The McGraw-Hill Companies

BUSINESS STATISTICS

THIRD EDITION

The McGraw-Hill Companies

BUSINESS STATISTICS

THIRD EDITION

G C BERI
Formerly Professor, Head and Dean
Faculty of Management Studies
M S University, Baroda

Tata McGraw Hill Education Private Limited
NEW DELHI

McGraw-Hill Offices
New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto

The McGraw-Hill Companies

Tata McGraw Hill

Published by the Tata McGraw Hill Education Private Limited,
7 West Patel Nagar, New Delhi 110 008.

Business Statistics, 3/e

ISBN-13: 978-0-07-008323-3
ISBN-10: 0-07-008323-1

Managing Director: *Ajay Shukla*
General Manager—Publishing (B&E/HSSL and School): *V Biju Kumar*
Publishing Manager—B&E: *Tapas K Maji*
Associate Sponsoring Editor: *Piyali Ganguly*
Editorial Executive: *Hemant K Jha*
Development Editor: *Shalini Negi*
Assistant Manager (Editorial Services): *Anubha Srivastava*
Senior Production Manager: *Manohar Lal*
General Manager—Marketing (Higher Ed & School): *Michael J Cruz*
Product Manager: *Vijay S Jagannathan*
General Manager—Production: *Rajender P Ghansela*
Assistant General Manager—Production: *B L Dogra*

Typeset at The Composers, 260, C.A. Apt., Paschim Vihar, New Delhi 110 063 and printed at
Avon Printers, Plot No. 16, Main Loni Road, Jawahar Nagar, Industrial Area, Shahdara, Delhi 110 094

Cover Design: K Anoop

Cover Printer: SDR Printers

RALLCRQFRCRRY

The McGraw-Hill Companies

The McGraw-Hill Companies

Dedicated to
The Memory of
My Revered Parents

PREFACE

At the outset, I would like to say that the favourable response to the second edition of this book was beyond my expectation. All the same, I believe that there is always some scope for improvement in any venture, and textbooks are not an exception to it. Accordingly, I decided to bring out the third edition with certain improvements. However, the basic approach of this text continues to be the same as in the previous editions, namely, the presentation of a modern introduction to statistical methods and data analysis for students pursuing courses in business administration, commerce and economics, both at the under-graduate and post-graduate levels.

Some improvements made in this edition are:

1. A new chapter on *Collection of Data* has been added just after the Introduction. It is the collection of data (and analysis) on which the entire subject of Statistics is based.
2. In Chapter 6 on *Measurements of Central Tendency*, relative merits and limitations of mean, median and mode have been explained, giving suitable examples.
3. The concepts of dispersion has been further explained emphasising upon its importance in knowing the characteristics of data. Some examples have been given where the study of dispersion would be most appropriate.
4. Chapter 9 on *Probability* has been enlarged especially on the Central Limit Theorem.
5. The components of a time series have been explained along with a chart. This apart, guidelines for time series analysis have been provided.
6. Barring a few chapters, new illustrative examples have been added in several chapters. These are mostly the questions set in university examinations.
7. Questions set in examinations of some universities in recent years have also been included at the end of several chapters.
8. Alongwith the above mentioned changes introduced in the third edition, I have retained the existing features such as:
   (i) learning objectives at the beginning of each chapter
   (ii) varying types of questions: 'true or false', multiple choice, and selected chapter-end questions along with their answer
   (iii) Statistical package for Social Sciences (SPSS)

(iv) Bibliography, and
(v) Index

It is hoped that this new edition would prove to be even more helpful to the students than the earlier edition.

I am extremely thankful to Mr Girish P and Ashok Kumar of SPSS South Asia, Bangalore for contributing a section comprising solutions to some problems using SPSS. The inclusion of SPSS solutions has become an important feature of this text.

I also gratefully acknowledge the suggestions concerning some topics, received from several reviewers. I may add that these suggestions were quite helpful and, accordingly, incorporated into this edition wherever feasible and appropriate.

I would like to put on record my appreciation of the staff of Tata McGraw Hill Education Private Limited, in particular to Mr Biju Kumar, Mr Tapas Kumar Maji, Mr Hemant Jha, Ms Anubha Srivastava, Mr Manohar Lal and Mr B L Dogra for their earnest efforts and courteous cooperation in the production of this edition. Last but not the least, I am grateful to my wife Sushila, who managed our domestic life single handed during the period when the manuscript for the third edition was under proportion.

At the end, I would like to urge the readers to inform me of any shortcomings that might have remained in the book so that these can be taken care of in subsequent edition.

G C Beri

The McGraw-Hill Companies

PREFACE TO THE FIRST EDITION

The purpose of this text is to present an introduction to statistical methods for students of both undergraduate and postgraduate management courses. The text has been written with the idea that this is their initial exposure to the subject of Statistics.

Although this is a modest attempt, it can be said safely that the text covers almost all topics in Statistics that a management student ought to learn. Besides, it is written in simple language and it does not require any advanced knowledge of mathematics on the part of the student.

To facilitate the students in their learning process, the text contains some features that may be mentioned here. Each chapter begins with chapter contents and chapter objectives and ends with a glossary of important terms, a list of formulae, and questions. Further, answers to all the concept questions and the selected numerical questions are given at the end of the book. These are followed by a short bibliography and an index. It is hoped that the students will find these features helpful not only in developing a genuine interest in the subject but also in acquiring the desired level of proficiency in it.

A textbook such as this one owes its existence to many authors whose writings form the basis of this book. It is almost impossible for me to acknowledge my indebtedness to all of them by name. All the same, I must acknowledge some of them whose help was of considerable importance.

I would like to express my sincere thanks to the anonymous reviewer, who offered helpful comments on the first ten chapters as also some other comments relevant for the book as a whole. I am grateful to Professor Y P Gupta, who pointed out certain discrepancies in the use of notations in some chapters. Thanks are also due to Professor Y K Bhushan, Director General, Narsee Monjee Institute of Management Studies, Mumbai; Professor Satendra Kumar, Head, Department of Business and Industrial Management, South Gujarat University, Surat; and Professor R Khasnabis, Head, Department of Business Management, University of Calcutta; for the favour they have done me by sending recent question-papers of their respective institutions.

During the course of writing the book, I had the benefit of using some libraries. Mention must be made of the library of Faculty of Management Studies, University of Delhi, for which I am thankful to the then Dean, Professor S P Gupta. My thanks are also due to the authorities of the British Council Library, New Delhi, as well as of the library of the Indian Statistical Institute, Delhi Centre.

The task of typing out the manuscript having 23 chapters, with a good deal of statistics and notations was not an easy one. Moreover, it turned out to be far more prolonged than what was envisaged at the time of its commencement. Mr Ishwinder Khanna performed this task with keen interest and competence for which I am extremely grateful to him.

I would like to express my deep sense of appreciation to the staff of the editorial and production departments of the Tata McGraw-Hill Publishing Company Limited, in particular to Mr V Biju Kumar, Mr Tapas Kumar Maji and Mr Raza Khan for their earnest efforts and courteous cooperation. It was a pleasant experience to have worked with such a harmonious team.

I am grateful to my wife, Sushila, who had to undergo a lot of tribulations during my prolonged involvement in writing the book.

At the end, I would like to say that despite my best efforts and the help received from various quarters, there might have remained some errors or deficiencies for which only I assume responsibility. I shall be extremely thankful to the learned professors in universities and management institutes if they do me a favour by pointing out these so that, in a subsequent edition, these can be taken care of.

G C Beri

CONTENTS

The McGraw-Hill Companies

CHAPTER 1 INTRODUCTION

Learning Objectives

After reading this chapter you would

- understand the meaning and definition of Statistics
- know the nature of Statistical study
- recognise the importance of Statistics in business and also its limitations
- know as to how Statistics can be misused
- differentiate descriptive Statistics from inferential Statistics.

1.1 MEANING AND DEFINITIONS OF STATISTICS

At the outset, it may be noted that the word 'Statistics' is used rather curiously in two senses—plural and singular. In the plural sense, it refers to a set of figures. Thus, we speak of production and sale of textiles, television sets, and so on. In the singular sense, Statistics refers to the whole body of analytical tools that are used to collect the figures, organise and interpret them and, finally, to draw conclusions from them.

It should be noted that both the aspects of Statistics are important if the quantitative data are to serve their purpose. If Statistics, as a subject, is inadequate and consists of poor methodology, we would not know the right procedure to extract from the data the information they contain. On the other hand, if our figures are defective in the sense that they are inadequate or inaccurate, we would not reach the right conclusions even though our subject is well developed. With this brief introduction, let us first see how Statistics has been defined.

Statistics has been defined by various authors differently. In the initial period the role of Statistics was confined to a few activities. As such, most of the experts gave a narrow definition of it. However, over a long period of time as its role gradually expanded, Statistics came to be considered as much wider in its scope and, accordingly, the experts gave a wider definition of it.

Spiegal, for instance, defines Statistics, highlighting its role in decision-making particularly under uncertainty, as follows:

> *Statistics is concerned with scientific method for collecting, organising, summarising, presenting and analysing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.*

This definition covers all the aspects and then tries to link them up with decision-making. After all, Statistics as a subject must help one to reach a reasonable and appropriate decision on the basis of the analysis of numerical data collected earlier.

Using the term 'Statistics' in the plural sense, **Secrist** defines Statistics as

> *"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 predetermined purpose, and placed in relation to each other".*

This definition of Secrist highlights a few major characteristics of statistics as given below:

1. Statistics are aggregates of facts. This means that a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is.
2. Statistics are affected by a number of factors. For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.
3. Statistics must be reasonably accurate. Wrong figures, if analysed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.
4. Statistics must be collected in a systematic manner. If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.
5. Finally, statistics should be placed in relation to each other. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

1.2 THE NATURE OF A STATISTICAL STUDY

Having briefly looked into the definition of Statistics, we should know at this stage as to what the nature of a Statistical study is. Whether a given problem pertains to business or to some other field, there are some well-defined steps that need to be followed in order to reach meaningful conclusions.

1. Formulation of the Problem To begin with, we have to formulate a problem on which a study is to be done. We should understand the problem as clearly as possible. We should know its scope so that we do not go beyond it or exclude some relevant aspect.

2. Objectives of the Study We should know what the objectives of the proposed study are. We should ensure that the objectives are not extremely ambitious or else the study may fail to achieve them because of limitations of time, finance or even competence of those conducting the study.

3. Determining Sources of Data The problem and the objectives, thus properly understood, will enable us to know as to what data are required to conduct the study. We have to decide whether we should collect primary data or depend exclusively on secondary data. Sometimes the study is based on both the secondary and the primary data. When study is to be based on secondary data, whether partly or fully, it is necessary to ensure that the data are quite suitable and adequate for the objectives of the study.

4. Designing Data Collection Forms Once the decision in favour of collection of primary data is taken, one has to decide the mode of their collection. The two methods available are: **(i)** observational method, and **(ii)** survey method. Suitable questionnaire is to be designed to collect data from respondents in a field survey.

5. Conducting the Field Survey Side by side when the data collection forms are being designed, one has to decide whether a census survey or a sample survey is to be conducted. For the latter, a suitable sample design and the sample size are to be chosen. The field survey is then conducted by interviewing sample respondents. Sometimes, the survey is done by mailing questionnaires to the respondents instead of contacting them personally.

6. Organising the Data The field survey provides raw data from the respondents. It is now necessary to organise these data in the form of suitable tables and charts so that we may be aware of their salient features.

7. Analysing the Data On the basis of the preliminary examination of the data collected as well as the nature and scope of our problem, we have to analyse data. As several statistical techniques are available, we should take special care to ensure that the most appropriate technique is selected for this purpose.

8. Reaching Statistical Findings The analysis in the preceding step will bring out some statistical findings of the study. Now we have to interpret these findings in terms of the concrete problem with which we started our investigation.

9. Presentation of Findings Finally, we have to present the findings of the study, properly interpreted, in a suitable form. Here, the choice is between an oral presentation and a written one. In the case of an oral presentation, one has to be extremely selective in choosing the material as in a limited time one has to provide a broad idea of the study as well as its major findings to be understood by the audience in proper perspective. In case of a written presentation, a report has to be prepared. It should be reasonably comprehensive and should have graphs and diagrams to facilitate the reader in understanding it in all its ramifications.

1.3 IMPORTANCE OF STATISTICS IN BUSINESS

There is an increasing realisation of the importance of Statistics in various quarters. This is reflected in the increasing use of Statistics in the government, industry, business, agriculture, mining, transport, education, medicine, and so on. As we are concerned with the use of Statistics in business and industry here, the description given below is confined to these areas only.

Three major functions where statistics can be found useful in a business enterprise.

1. ***The planning of operations*** *This may relate to either special projects or to the recurring activities of a firm over a specified period.*
2. ***The setting up of standards*** *This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output, and so forth.*
3. ***The function of control*** *This involves comparison of actual production achieved against the norm or target set earlier. In case the production has fallen short of the target, it gives remedial measures so that such a deficiency does not occur again.*

A point worth noting here is that although these three functions—planning of operations, setting standards, and control—are separate, but in practice they are very much interrelated.

Various authors have highlighted the importance of Statistics in business. For instance, **Croxton** and **Cowden** give numerous uses of Statistics in business such as project planning, budgetary planning and control, inventory planning and control, quality control, marketing, production and personnel administration. Within these also they have specified certain areas where Statistics is very relevant. **Irwing W. Burr**, dealing with the place of Statistics in an industrial organisation, specifies a number of areas where Statistics is extremely useful. These are: customer wants and market research, development design and specification, purchasing, production, inspection, packaging and shipping, sales and complaints, inventory and maintenance costs, management control, industrial engineering and research.

It can be seen that both the lists are extremely comprehensive. This clearly points out that specific statistical problems arising in the course of business operations are multitudinous. As such, one may do no more than highlight some of the more important ones to emphasise the relevance of Statistics to the business world.

Statistical Quality Control Methods

In the sphere of production, for example, Statistics can be useful in various ways to ensure the production of quality goods. This is achieved by identifying and rejecting defective or substandard goods. The sale targets can be fixed on the basis of sale forecasts, which are done by using varying methods of forecasting. Analysis of sales effected against the targets set earlier would indicate the deficiency in achievement, which may be on account of several causes: **(i)** targets were too high and unrealistic **(ii)** salesmen's performance has been poor **(iii)** emergence of increase in competition **(iv)** poor quality of company's product, and so on. These factors can be further investigated.

Personnel Management

This is another sphere in business where statistical methods can be used. Here, one is concerned with the fixation of wage rates, incentive norms and performance appraisal of individual employee. The concept of productivity is very relevant here. On the basis of measurement of productivity, the productivity bonus is awarded to the workers. Comparisons of wages and productivity are undertaken in order to ensure increase in industrial productivity.

Seasonal Behaviour

A business firm engaged in the sale of certain product has to decide how much stock of that product should be kept. If the product is subject to seasonal fluctuations then it must know the nature of seasonal fluctuations in demand. For this purpose, seasonal index of consumption may be required. If the firm can obtain such data or construct a seasonal index on its own then it can keep a limited stock of the product in lean months and large stocks in the remaining months. In this way, it will avoid the blocking of funds in maintaining large stocks in the lean months. It will also not miss any opportunity to sell the product in the busy season by maintaining adequate stock of the product during such a period.

Export Marketing

Developing countries have started giving considerable importance to their exports. Here, too, quality is an important factor on which exports depend. This apart, the concerned firm must know the probable countries where its product can be exported. Before that, it must select the right product, which has considerable demand in the overseas markets. This is possible by carefully analysing the statistics of imports and exports. It may also be necessary to undertake a detailed survey of overseas markets to know more precisely the export potential of a given product.

Maintenance of Cost Records

Cost is an important consideration for a business enterprise. It has to ensure that cost of production, which includes cost of raw materials, wages, and so forth, does not mount up or else this would jeopardise its competitiveness in the market. This implies that it has to maintain proper cost records and undertake an analysis of cost data from time to time.

Management of Inventory

Closely related to the cost factor is the problem of *inventory management*. In order to ensure that the production process continues uninterrupted, the business firm has to maintain an adequate inventory. At the same time, excessive inventory means blocking of funds that could have been utilised elsewhere. Thus, the firm has to determine a magnitude of inventory that is neither excessive nor inadequate. While doing so, it has to bear in mind the probable demand for its product. All these aspects can be well looked after if proper statistics are maintained and analysed.

Expenditure on Advertising and Sales

A number of times business firms are interested to know whether there is an association between two or more variables such as *advertising expenditure and sales*. In view of increasing competitiveness, business and industry spend a large amount on advertising. It is in their interest to find out whether such advertising expenditure promotes the sales. Here, by using correlation and regression techniques it can be ascertained that the advertising expenditure is worthwhile or not.

Mutual Funds

Mutual funds which have come into existence in recent years, provide an avenue to a person to invest his savings so that he may get a reasonably good return. Different mutual funds have different objectives as they have varying degrees of risk involved in the companies they invest in. Here, Statistics provides certain tools or techniques to a consultant or financial adviser through which he can provide sound advice to a prospective investor.

Relevance in Banking and Insurance Institutions

Banks and insurance companies frequently use varying statistical techniques in their respective areas of opertaion. They have to maintain their accounts and analyse these to examine their performance over a specified period.

The above discussion is only illustrative and there are numerous other areas where the use of Statistics is so common that without its use they may have to close down their operations.

1.4 LIMITATIONS OF STATISTICS

The preceding discussion highlighting the importance of Statistics in business should not lead anyone to conclude that Statistics is free from any limitation. As we shall see here, Statistics has a number of limitations.

1. There are certain phenomena or concepts where Statistics cannot be used. This is because these phenomena or concepts are not amenable to measurement. For example, beauty, intelligence, courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.
2. Statistics reveal the average behaviour, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous. For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet, when there may be some points in between where its depth is far more than four feet. On this understanding, one may enter those points having greater depth, which may be hazardous.
3. Since Statistics are collected for a particular purpose, such data may not be relevant or useful in other situations or cases. For example, secondary data (i.e., data originally collected by someone else) may not be useful for the other person.
4. Statistics is not 100 per cent precise as is Mathematics or Accountancy. Those who use Statistics should be aware of this limitation.
5. In Statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The results may not be appropriate as far as the universe is concerned. Moreover, different surveys based on the same size of sample but different sample units may yield different results.
6. At times, association or relationship between two or more variables is studied in Statistics, but such a relationship does not indicate 'cause and effect' relationship. It simply shows the similarity or dissimilarity in the movement of the two variables. In such cases, it is the user who has to interpret the results carefully, pointing out the type of relationship obtained.
7. A major limitation of Statistics is that it does not reveal all pertaining to a certain phenomenon. There is some background information that Statistics does not cover. Similarly, there are some other aspects related to the problem on hand, which are also not covered. The user of Statistics has to be well informed and should interpret Statistics keeping in mind all other aspects having relevance on the given problem.

1.5 MISUSE OF STATISTICS

Apart from the limitations of Statistics mentioned above, there are misuses of it. Many people, knowingly or unknowingly, use statistical data in wrong manner. Let us see what are the main misuses of Statistics so that the same could be avoided when one has to use statistical data. The misuse of Statistics may take several forms some of which are explained below.

1. Sources of Data not Given At times, the source of data is not given. In the absence of the source, the reader does not know how far the data are reliable. Further, if he wants to refer to the original source, he is unable to do so.

2. Defective Data Another misuse is that sometimes one gives inaccurate data. This may be done knowingly in order to defend one's position or to prove a particular point. This apart, the definition used to denote a certain phenomenon may be defective. For example, in case of data relating to unemployed persons, the definition may include even those who are employed, though partially. The question here is how far it is justified to include partially employed persons amongst unemployed ones.

3. Unrepresentative Sample In Statistics, several times one has to conduct a survey, which necessitates to choose a sample from the given population or universe. The sample may turn out to be unrepresentative of the universe. One may choose a sample just on the basis of convenience. He may collect the desired information from either his friends or nearby respondents in his neighbourhood even though such respondents do not constitute a representative sample.

4. Inadequate Sample Earlier, we have seen that a sample that is unrepresentative of the universe is a major misuse of Statistics. This apart, at times one may conduct a survey based on an extremely inadequate sample. For example, in a city we may find that there are 1,00,000 households. When we have to conduct a household survey, we may take a sample of merely 100 households comprising only 0.1 per cent of the universe. A survey based on such a small sample may not yield right information.

5. Unfair Comparisons An important misuse of Statistics is making unfair comparisons from the data collected. For instance, one may construct an index of production choosing the base year where the production was much less. Then he may compare the subsequent year's production from this low base. Such a comparison will undoubtedly give a rosy picture of the production though in reality it is not so.

Another source of unfair comparisons could be when one makes absolute comparisons instead of relative ones. An absolute comparison of two figures, say, of production or export, may show a good increase, but in relative terms it may turn out to be very negligible. Another example of unfair comparison is when the population in two cities is different, but a comparison of overall death rates and deaths by a particular disease is attempted. Such a comparison is wrong. Likewise, when data are not properly classified or when changes in the composition of population in the two years are not taken into consideration, comparisons of such data would be unfair as they would lead to misleading conclusions.

6. Unwarranted Conclusions Another misuse of Statistics may be on account of unwarranted conclusions. This may be as a result of making false assumptions. For example, while making projections of population in the next five years, one may assume a lower rate of growth though the past two years indicate otherwise. Sometimes one may not be sure about the changes in business environment in the near future. In such a case, one may use an assumption that may turn out to be wrong. Another source of unwarranted conclusion may be the use of wrong average. Suppose in a series there are extreme values, one is too high while the other is too low, such as 800 and 50. The use of an arithmetic average in such a case may give a wrong idea. Instead, harmonic mean would be proper in such a case.

7. Confusion of Correlation and Causation In Statistics, several times one has to examine the relationship between two variables. A close relationship between the two variables may not establish a cause-and-effect-relationship in the sense that one variable is the cause and the other is the effect. It should be taken as something that measures degree of association rather than try to find out causal relationship.

8. Suppression of Unfavourable Results Another wrong use of Statistics may be on account of suppressing results that are unfavourable to the organisation or an individual. Revealing such results may expose the concerned organisation or person in bad light. In order to avoid such a situation, one may be tempted to hide unfavourable, though true, facts emerging from a statistical study.

9. Mistakes in Arithmetic Finally, one may come across certain mistakes in calculations or in the application of a wrong formula. This human error may result in grossly wrong figures, leading to wrong conclusions.

The foregoing discussion on misuses of Statistics clearly indicates the pitfalls in which one is likely to be trapped if one does not exercise sufficient care in the collection, analysis and interpretation of data.

1.6 SUBDIVISIONS WITHIN STATISTICS

Having discussed the limitations and misuses of Statistics, we now turn to its subdivisions. There are many diverse techniques (which we will learn in the subsequent chapters of this book) that are used in Statistics. The statisticians commonly classify this subject into two broad categories: the *Descriptive statistics* and *Inferential statistics*.

Descriptive Statistics

As the name suggests *descriptive statistics* includes any treatment designed to describe or summarise the given data, bringing out their important features. These statistics do not go beyond this. This means that no attempt is made to infer anything that pertains to more than the data themselves. Thus, if someone compiles the necessary data and reports that during the financial year 2000–01, there were 1500 public limited companies in India of which 1215 earned profits and the remaining 285 companies sustained losses, his study belongs to the domain of descriptive Statistics. He may further calculate the average profit earned per company as also average loss sustained per company. This set of calculations too is a part of descriptive Statistics.

Methods used in descriptive Statistics may be called as descriptive methods. Under descriptive methods, we learn frequency distribution, measures of central tendency, that is, averages, measures of dispersion and skewness.

Inferential Statistics

Although descriptive Statistics is an important branch of Statistics and it continues to be so, its recent growth indicates a shift in emphasis towards the methods of Statistical inference. Inferential Statistics uses a number of quantitative techniques that enable us to make appropriate generalisations from limited observations. The events and phenomena in which the economists and business analysts are interested are too numerous and at times too inaccessible as a result a complete observation or census is not possible. As such sampling theory is used to assist in the rational decision-making process.

Let us take an example. Suppose a business firm is shortly going to introduce a new product in the market. For planning proper marketing of this product, the marketing manager of the firm wants to identify a niche market. He conducts a survey, collects data and estimates the income of households based on the sample. He calculates an interval estimate, i.e., the lower and upper limits within which the population income is likely to fall. On the basis of such an exercise, he is able to identify areas where the marketing efforts should be focused.

It may be noted that the whole area of statistical inference, whether we are concerned with interval estimation problems, tests of hypotheses, or correlation, leans heavily on the theory of probability.

In this text, we shall first discuss various aspects of descriptive Statistics. This will be followed by the discussion on different topics in inferential Statistics. The latter will understandably be far more comprehensive than the former.

GLOSSARY

Data or Data set	A collection of observations or measurements on one or more variables.
Descriptive statistics	A collection of methods that enable us to organise, display and describe data using such devices as tables, graphs and summary measures.
Inferential statistics	A collection of methods that enable us in making decisions about a population based on sample results.
Statistics (plural)	Numerical facts.
Statistics (singular)	A group of methods that are used to collect, analyse, present and interpret data and to make decisions.

QUESTIONS

1.1 Given below are ten statements. Indicate in each case whether it is true or false:

(a) The word "Statistics" has only one meaning, which implies numerical data.
(b) "Statistics is a science of counting." It is one of the most comprehensive definitions.
(c) Statistics are affected by a number of factors.
(d) Statistics has hardly any limitations.
(e) Statistics is as precise as mathematics or accountancy.
(f) Sometimes Statistics is used to suppress unfavourable results.
(g) Descriptive Statistics deals with 'how' and 'why' of a phenomenon.
(h) A single figure is not statistics.
(i) Recent growth of Statistics is more towards inferential studies as compared to descriptive studies.
(j) Statistics is only marginally useful so far as business problems are concerned.

Multiple Choice Questions (1.2 to 1.4)

1.2 Which of the following statements about Statistics is not correct?
(a) Statistics is sometimes misused by statisticians.
(b) Statistics has hardly any limitations.
(c) Statistics is indispensable in the management of business enterprises.
(d) Statistics is of two types: descriptive and inferential

1.3 Which of the following statements describes 'Statistics' most appropriately?
(a) Statistics is the science of counting.
(b) Statistics is the science of averages.

(c) Statistics is concerned with the collection, presentation and analysis of data leading to valid conclusions.
(d) Statistics is the aggregate of facts.

1.4 Which of the following statements is not true?
(a) Statistics deals with quantitative data.
(b) Statistics deals with qualitative data.
(c) Statistics deals with both quantitative and qualitative data.
(d) Statistics does not cover graphical devices.

1.5 Why is it necessary to learn Statistics?

1.6 Define Statistics. Discuss its use in business decision making.

1.7 "Statistics is all-pervading." Elucidate this statement.

1.8 Write a note highlighting the importance of Statistics to business.

1.9 "Statistics only furnishes a tool, necessary, though imperfect, which is dangerous in the hands of those who do not know its use and deficiencies."—Bowley. Discuss.

1.10 "Statistics are the straws out of which I, like every other economist, have to make the bricks."—Marshall. Elucidate the statement.

1.11 What are the major limitations of Statistics? Explain with suitable examples.

1.12 "Statisticians at times misuse Statistics." Elucidate this statement.

1.13 What are the functions of Statistics? Describe briefly.

1.14 Distinguish between descriptive Statistics and inferential Statistics.

1.15 "Statistical methods are most dangerous tools in the hands of an inexpert. Statistics is one of those sciences whose adept must exercise the self restraint of an artist." Elucidate.

1.16 "Science without Statistics bear no fruit, Statistics without science have no roots." Explain the significance of this statement.

1.17 What role does Business Statistics play in the management of business enterprise? Illustrate your answer with some typical business problems and the statistical techniques to be used there.

1.18 How statistical methods are likely to be of any use to a marketing firm? Give some suitable examples.

1.19 Discuss important applications of Statistics with special reference to decision-making in modern business and industry. Also, briefly describe its limitations, if any.

1.20 "Today Statistics is an indispensable tool in any sphere of human activity." Elucidate the statement.

1.21 For each of the following, indicate whether the study is descriptive or analytic (inferential), giving briefly the reason for your answer:
(a) A management institute would like to determine the reasons for a decline in the number of applicants to its MBA programme in the last three years.
(b) The proprietors of a magazine would like to know whether an increase in the subscription rate will adversely affect the number of subscribers.
(c) A company is interested to know whether its advertisements are followed by favourable results in the form of increased sales.
(d) A company would like to know as to how many applicants who have sent in their applications for a job are graduates of a foreign university.

The McGraw-Hill Companies

CHAPTER 2 COLLECTION OF DATA

Learning Objectives

After reading this chapter you would understand,

- the types of data that are used in Statistics
- advantages, disadvantages and evaluation of secondary data
- two main methods of collecting primary data
- designing of a questionnaire.

While discussing the nature of a statistical study in Chapter 1, it was mentioned that several steps are involved in a statistical study. After the problem has been formulated and the objectives of the study spelt out, the next step is to decide as to what type of data are required and how to collect such data. It should be noted that data collection is a major step in any statistical investigation. If the data collected are inaccurate or inadequate or defective in some other respect, the findings based on such data will undoubtedly be misleading. In order to avoid such a situation, it is necessary to devote enough time to plan proper collection of the requisite data. Accordingly, this chapter is devoted to the collection of data, an activity that gives rise to several other functions of Statistics.

When we talk of collection of data, we should be clear as to what does the word 'data' connote. The word *datum* is a Latin word, which literally means 'something given'. It means a piece of information, which can be either quantitative or qualitative. The term 'data' is the plural of 'datum' and means 'facts and statistics collected together for reference or analysis'. Thus, any information collected is data.

2.1 TYPES OF DATA

There are two types of data that are collected and analysed in Statistics. These are *Secondery data* and *Primary data.*

Secondary Data

Any data that have been gathered earlier for some other purpose are secondery data in the hands of an individual who is using them.

Primary Data

The data that are collected first hand by someone specifically for the purpose of facilitating the study are known as primary data. Thus, primary data collected by one person may become the secondary data for another.

Regardless of whether the data are primary or secondary, they can be classified in a different way viz. *qualitative* and *quantitative data*.

Qualitative Data are expressed by a non-numerical property such as satisfaction of a customer, rich, poor and superior.

Quantitative Data are numerically expressed. For example, weight, height, income, expenditure and price are quantitative data.

It will be seen that the latter use of data invariably involves numbers. It is in this sense that the word 'data' is used in Statistics most of the times as will be evident as we go on to study different topics covered in Statistics. At the same time, very occasionally, we shall concern ourselves with the qualitative data, too.

Even the quantitative data are ordinarily of two types—*measurements* or *scores* and *frequencies*. For example, height in centimeters is a measurement of linear body size. The performance of a worker is in the form of output produced by him in a given time period. Such a performance is the measurement. But, there are certain situations or areas where phenomena cannot be measured. In such cases, our observations are *qualitative* or *categorical*. For example, we can classify or categorise population by gender—male and female; students by their performance—intelligent and dull students; persons into rich and poor categories; and so on. Such classification shows the number of times that event has occurred. Data of this type are called *frequencies*.

A variable is a symbol such as x and y, which can assume any of a prescribed set of values, called the *domain* of the variable. In case the variable can assume only one value it is called a *constant*. A variable, that can theoretically assume any value between two given values is called a *continuous* variable; otherwise it is called *discrete* variable.

Based on the above classification, data too can be either discrete or continuous. For example, the number of students passed in an examination is the *discrete* data. Likewise, the number of children born in a city during a specified period is also discrete data. As against this, temperatures recorded by a meteorological bureau and litres of water used in washing clothes are the examples of *continuous* data.

We will first discuss secondary data in detail.

2.2 SECONDARY DATA

Any data that have been gathered earlier for some other purpose are secondary data in the hands of an individual who is using them. In contrast, the data that are collected first hand by someone specifically

for the purpose of facilitating the study are known as primary data. Thus, primary data collected by one person may become the secondary data for another. For example, the demographic statistics collected every ten years are the primary data with the Registrar General of India, but the same statistics used by anyone else would be secondary data with that individual. There are certain distinct advantages, as also the limitations, of using secondary data. One should, therefore, be fully aware of both the advantages and the limitations.

Advantages of Secondary Data

1. A major advantage in the use of secondary data is that it is far more economical as the cost of collecting original data is saved. In the collection of primary data, a good deal of effort is required—data collection forms are to be designed and printed, field staff is to be appointed and maintained until all the data have been collected, the travelling expenses are to be incurred, the sample design is to be selected, data are to be collected and verified for their accuracy, and finally, all such data are to be tabulated. All these activities would need large funds, which can be utilised elsewhere if secondary data alone can serve the purpose.
2. Another advantage is that the use of secondary data saves much of our time. This leads to prompt completion of the report for which, otherwise, primary data would have been required to be collected.
3. Search for secondary data is helpful, not only because secondary data may be useful but familiarity with such data indicates the deficiencies and gaps. As a result, one can make the primary data collection more specific and more relevant to one's study.
4. As one explores the availability of secondary data required for one's project, one finds, in the process, that one's understanding of the problem has improved. One may even have to change some of one's earlier ideas in the light of the secondary data.
5. Finally, secondary data can be used as a basis for comparison with the primary data that have been just collected.

Disadvantages of Secondary Data

In practice, secondary data seldom fit perfectly into the framework of a proposed study. This is on account of a number of factors.

1. The unit in which secondary data are expressed may not be the same as is required in the proposed study. For example, the size of firm can be expressed as (i) number of employees, (ii) paid-up capital employed, (iii) gross sales, (iv) gross or net profit and so on. It is just possible that the unit of measurement used in secondary data is different from the one needed in the proposed study. In such a case, secondary data cannot be used.
2. Even if the units are the same as those required by the research project, it may be that class boundaries are different from those desired. For instance, the weekly income of households may have a break-up of (i) less than Rs 500, (ii) Rs 501–1000, (iii) Rs 1001–1500, (iv) Rs 1501–2000, (v) Rs 2001+, so far as secondary data are concerned. If we want to find, for example, the number of households with a weekly income of Rs1800 or some similar figure, we will be at a loss with such secondary data.
3. One does not always know how accurate the secondary data are. In case the degree of inaccuracy is high, the use of such dubious data would undermine the utility of our study. In most cases, it is difficult to know with what care secondary data have been collected and tabulated. All the same,

in the case of well-established and reputed organisations, both official and non-official, secondary data would be far more accurate and reliable and they can be used without much reservation.

4. A severe limitation in the use of secondary data is that they may be somewhat out of date. A good deal of time is spent in the collection, processing, tabulation and publishing of such data and by the time the data are available, they are already two to three years old. As a result, the data are no longer up-to-date. It is a moot question as to how such data are relevant at the time of their use. Obviously, the utility of secondary data declines progressively as the time goes by, and they are finally useful only for historical purposes.

Evaluating Secondary Data

Since the use of secondary data is substantially cheaper than that of primary data, it is advisable to explore the possibility of using secondary data. In this connection, there are four requirements that must be met. These are: (i) Availability of secondary data, (ii) Relevance, (iii) Accuracy and (iv) Sufficiency. These requirements are briefly discussed here.

- The first and foremost requirement is that secondary data must be available for use. At times, one may find that secondary data are just not available on a problem at hand. In such cases, there is no alternative but to take recourse to the collection of primary data.
- Another precondition for the use of secondary data is their relevance to the marketing problem. Relevance means that the data available must fit the requirements of that problem. This would cover several aspects. First, the unit of measurement should be the same as that in the marketing problem. Second, the concepts used should be the same as are envisaged in the problem. Another pertinent issue is that the data should not be obsolete.
- The third requirement is that the data should be accurate. In this connection, one should consult the original source. This would not only enable one to get more comprehensive information but would also indicate the context in which data have been collected, the procedure followed and the extent of care exercised in their collection.
- Finally, the data should be sufficient. If the data are inadequate, then compliance with the preceding requirements will be in vain.

The foregoing requirements must be met to avoid an improper use of secondary data. One has to be extremely careful when deciding to use any secondary data. To help take a decision, one has to seek answers to such questions as: What sample design was used for collecting data? What questionnaire was used? What was the quality of the field staff that collected the data? What was the extent of non-responses and how was the problem handled by the organisation? These are some of the questions that are pertinent while deciding the reliability of secondary data. In the final analysis, it is the reputation of the organisation collecting and publishing such data, and its regularity in their publication that would carry more weight than anything else.

Sources of Secondary Data

Secondary data can be obtained internally, that is within the firm, or externally that is from one or more outside agencies. Internal secondary data are those that are generated within the firm. These include the financial accounts and sales and other records maintained by the firm.

The external secondary data do not originate in the firm and are obtained from outside sources.

It may be noted that secondary data can be collected from the originating sources or from secondary sources. For example, the Office of the Economic Adviser, Government of India, is the originating

source for the data on the wholesale price index. In contrast, a publication such as the *Reserve Bank of India Bulletin* containing some parts of the series of wholesale prices, is a secondary source.

Generally, the originating source of external secondary data should be preferred on account of several reasons. *First*, the originating source is more likely to explain the object and procedure of data collection. *Second*, the originating source is more likely to present all the data, whereas a secondary source may present a part of such data, depending on its requirement or convenience. *Finally*, the originating source would be more accurate as each additional repeating source of secondary data presents another possible source of error.

Despite these advantages of using the originating source of secondary data, many a time secondary sources of secondary data are used. There may be good reasons for this. *First*, the secondary source may be readily available and, as such, it is convenient to use it if the data are sufficiently reliable. At times, one may have to refer to different originating publications and search through numerous pages. The likely improvement in the quality of secondary data may not be commensurate with the time and effort required for using the originating source. *Second*, sometimes, secondary sources provide secondary data on magnetic tape for computer input. As a result of this facility, one may prefer the secondary source.

2.3 PRIMARY DATA

In everyday life, if we want to have first-hand information on any happening or event, we either ask someone who knows about it or we observe it ourselves or we do both. The same is applicable to any statistical study. Thus, the two main methods by which primary data can be collected are *observation* and *communication*. This section is devoted to these methods.

Observation

Observation is one of the methods of collecting data. It is used to get both past and current information. For example, instead of asking respondents about their current behaviour, we may observe it and record our observations. Although it is not possible to observe past behaviour, we may observe the results of such behaviour. In a way, secondary data reflect the results of the past behaviour of people as also of past occurrences.

There are some *advantages* of observation as a method of collecting information. To begin with, the direct observational technique enables the investigator to record behaviour as it occurs. In contrast, other techniques record the data mostly retrospectively, on the basis of the respondent's report after the event. Another merit of direct observation is that it can be used regardless of whether the respondent is willing to report or not. In a field survey, if an enumerator comes across an unwilling and hostile respondent, he cannot collect the desired information. But, this problem does not arise at all in the case of direct observation. Yet another advantage of observation is that it can be used even when it pertains to those who are unable to respond, such as infants and animals.

There are, however, some *limitations* of this method. *Firstly*, only the current behaviour of a person or a group of persons can be observed. One is unable to observe the past behaviour nor can one observe a person's future behaviour because the act of observation takes place in the present. *Secondly*, observation does not help us in gauging a person's attitude or opinion or knowledge on a certain subject. *Thirdly*, the observational method is very slow and, therefore, when a large number of persons are to be contacted, it becomes unsuitable because of the long time required for this purpose.

Questionnaire

The communication method, in effect, is the method of designing questionnaires with a view to collect the requisite information. The questionnaires can be classified into four main types:

(i) Structured-non-disguised
(ii) Structured-disguised
(iii) Non-structured-non-disguised
(iv) Non-structured-disguised

It may be mentioned here that some authors prefer to call the 'non-disguised' as direct and 'disguised' as indirect questionnaires.

A structured questionnaire is a formal list of questions framed so as to get the facts. The interviewer asks the questions strictly in accordance with a prearranged order. For example, if he is interested in knowing the amount of expenditure incurred on different types of clothing, viz., cotton, woollen or synthetic, by different households classified according to their income, he may frame a set of questions seeking this factual information.

A structured questionnaire can be of two types, namely, *disguised* and *non-disguised*. This classification is based on whether the object or purpose of the survey is revealed or undisclosed to the respondent. Thus, a structured non-disguised questionnaire is one where the listing of questions is in a pre-arranged order and where the object of enquiry is revealed to the respondent. In the case of a structured-disguised questionnaire, the investigator does not disclose the object of the survey. He feels that if the respondent comes to know the object of the survey, he may not be objective in giving the necessary information and, as such, its purpose may be defeated. He is, therefore, very particular not to divulge the purpose of the investigation.

A non-structured questionnaire is one in which the questions are not structured and the order in which they are to be asked from the respondent is left entirely to the investigator. He asks the questions in the manner in which he deems fit in a particular situation. In fact, he may only have certain main points on which he may develop the questions at the time of the actual interview.

Designing a Questionnaire

Designing a questionnaire is not a simple job. A statistician intending to collect primary data has to be extremely careful in deciding what information is to be collected, how many questions are to be formulated, what should be their sequence, what should be the wording of each question, and what should be the layout of the questionnaire. All these aspects need considerable time and effort of the statistician. If he is able to develop a questionnaire suitable for his field investigation, he will find that his task of collecting the data has become much easier than otherwise.

1. Type of Information to be Collected While designing a questionnaire, the statistician has to first ask himself what type of information he needs from the survey. He should seriously consider this question as it will have considerable repercussion on the usefulness of the survey. For, if he omits to collect information on some relevant and vital aspects of his survey, his study is unlikely to be useful. At the same time, if he collects information on some issues not directly relevant to his study, he not only raises the total cost of the survey but also increases the time factor. This being the case, the survey will take much more time than is actually necessary. It will also lead to greater inaccuracy as the respondent will have to answer many more questions than are necessary and he will, therefore, not be

sufficiently careful in giving the exact answer. In either case, the statistician will be the loser. To avoid this situation, he should give serious thought to the specific information to be sought.

2. Types of Questions The second important aspect in the designing of a questionnaire is to decide on the types of questions to be used. Questions can be classified in various ways. One way of classification is as follows:

(i) Open-ended questions
(ii) Dichotomous questions
(iii) Multiple-choice questions

An *open-ended* or simple 'open' or 'free answer' question gives the respondent complete freedom to decide the form, length and detail of the answer. Open questions are preferred when the investigator is interested in knowing what is uppermost in the mind of the respondent. However, open questions pose certain problems. At the time of the actual interview, it becomes difficult for the interviewer to note down the respondent's answer verbatim. Further, if several interviewers are conducting interviews and each one is recording the answers to open-ended questions according to his understanding, and in his own way, then there is likely to be an element of bias in the recorded answers.

The *dichotomous question* has only two answers in the form 'yes' or 'no', 'true' or 'false', 'use' or 'do not use', and so on. It may be pointed out that dichotomous questions are most convenient or least bothersome to respondents, who have simply to indicate their choice from the two possible answers. As such, these questions require the minimum possible time of the respondents. Also, answers to such questions are easy to edit, tabulate and interpret.

In the case of *multiple-choice questions*, the respondent is offered more than two choices. The investigator exhausts all the possible choices and the respondent has to indicate which one is the most appropriate. Obviously, the respondent is likely to take more time to answer a multiple-choice question as compared to a dichotomous one. Also, more time is required in the editing, tabulation and interpretation of data.

3. Phrasing of the Questions The next issue in the preparation of the questionnaire is how to phrase the questions. The way in which a question is drafted is very important as even a slightly suggestive wording would elicit a very different answer from the respondent. For example: Don't you think that this is a sub-standard product?

A question of this type would prompt respondents to answer in the affirmative. Many of them, who do not have a definite opinion about the product are likely to agree that it is of sub-standard quality. However, if the above question is worded a little differently, the answer is likely to be different. Suppose this question is put as follows:

Do you think that this is a sub-standard product? As it is a straight forward question, respondents are not likely to be prompted to say 'yes' as was the case in the earlier question, which was suggestive.

4. Order of Questions Another aspect that should receive the attention is the sequence or order of questions to be contained in a questionnaire. Since, in the beginning, the interviewer has to establish some rapport with the respondent, it is necessary that questions asked at the beginning are simple and thereby helpful in establishing the rapport. Difficult questions or those on sensitive issues should be relegated to the end of the questionnaire. Further, questions of a general type should be asked in the beginning, while those which are specialised, needing some in-depth information from the respondents, should be left to the end.

5. How many Questions to be Asked The interviewer has also to decide how many questions are to be asked. We may add that the number of questions is not so important as the actual length of the questionnaire. We have just mentioned above that the interest of the respondent should be sustained until the last moment so that the interview can be completed successfully and the requisite information obtained. Too lengthy a questionnaire would obviously be a disadvantage and the response to it may be quite poor.

6. Layout of the Questionnaire Finally, the layout of the questionnaire has to be decided. This implies that the document should be set in such a way that it leaves a favourable impression in the mind of the respondent. It should be neatly printed and the individual pages should not have too many questions so as to appear crowded. Proper spacing between the questions and within a question should be provided for. The more important wordings to which the attention of the respondent is to be drawn should be set in bold type or underlined. If it is a lengthy questionnaire, special care should be taken to reduce its size by providing two columns in a page and by using finer type. But, this can be done up to a certain point for too fine, a print may cause inconvenience to the respondent. The questionnaire should have 'easy looks', which means that it should be short and printed on superior quality paper so that writing with pen or pencil is smooth.

A Specimen Questionnaire

A specimen questionnaire is given at the end of this chapter. It will be seen that the questionnaire relates to the introduction of a product in the market. The first eight questions (or items) seek information about the respondent such as his age, educational level, occupation, income, and so on. Questions 9 to 18 relate to the new product. The questions are simple so that the respondent can understand. These questions are close-ended questions except the last one, which is an open-ended question wherein, the respondent is free to provide any answer or information, which he thinks, has not been covered earlier and which is relevant to the survey. A point worth noting is that most of the questions are pre-coded as different responses are given numerical codes as (1), (2), (3) and so on. Such coding facilitates categorisation of responses so that tabulation work can be expeditiously completed.

Pre-testing the Questionnaire

Once the questionnaire is ready, it should be pre-tested. Pre-testing of the questionnaire implies that it is tried out on a few respondents and the reaction to the questionnaire is observed. It helps us in deciding whether any changes in the question-content or the wording of questions are called for. If so, specific changes that are desirable can also be ascertained and incorporated in the questionnaire. This would improve it, and if it is a mail questionnaire, it would perhaps increase the response rate as well.

The other benefit of pre-testing the questionnaire is that we can know the suitability of the instructions given to the interviewers as also their capability. In case certain changes are required, the same can be introduced. Interviewers will also have an opportunity to familiarise themselves with the problems they might face in the collection of data. This apart, pre-testing may indicate whether a particular sample design is feasible or some other sample design, which may be more appropriate, should be selected. Sometimes pre-testing of a questionnaire is undertaken to find out the suitability of data for particular needs. For this purpose, we may have to tabulate the data collected in the pilot survey (or pre-testing) and prepare dummy tables. With the help of these tables, we can examine whether such data would be appropriate and adequate for the objectives of the survey. In the light of this investigation, the questionnaire can be revised to elicit additional information.

2.4 EDITING

Once the data have been collected, the first task is the editing. It is the process by which data are prepared for subsequent coding. As it is a very subjective process, it is necessary that persons who are well qualified and trained in the job of editing, should alone be given this responsibility.

Editing is the process of examining errors and omissions in the collected data and making necessary corrections in the same. This is desirable when there is some inconsistency in the response or responses as entered in the questionnaire or when it contains only a partial or a vague answer.

In all cases where editorial corrections are to be made, it is necessary that these should be kept distinct from the changes made either by the respondent or by the interviewer. The editor can ensure this by using a different coloured pencil for editing the raw data.

2.5 CODING

Coding is the procedure of classifying the answers to a question into meaningful categories. The symbols used to indicate these categories are codes. Coding is necessary to carry out the subsequent operations of tabulating and analysing data. If coding is not done, it will not be possible to reduce a large number of heterogeneous responses into meaningful categories with the result that the analysis of the data would be weak, ineffective and without proper focus.

Coding involves two steps. The first step is to specify the different categories or classes into which the responses are to be classified. The second step is to allocate individual answers to different categories.

There is no definite rule for the number of categories or classes that can be used. This will depend on the nature of the problem as also the extent of analysis one would like to carry out.

A practice, which is frequently followed, is to edit and code the data simultaneously. These two operations are regarded as one operation, which is looked after by one person. Although this may perhaps be the quickest and most efficient method, it may lead to the neglect of editing, as the editor who is expected to code becomes just a coder. In view of this, it may be advisable to get these jobs done by two persons.

QUESTIONS

2.1 Differentiate between primary data and secondary data.

2.2 Which of the following is primary data?

- **(a)** Census of population data
- **(b)** Wholesale Price Index Numbers
- **(c)** Statistics contained in an official publication such as the Reserve Bank of India Bulletin
- **(d)** Data collected through your own field survey

2.3 Collection of secondary data is

- **(a)** Cheaper
- **(b)** Faster
- **(c)** More accurate
- **(d)** (a) and (b)
- **(e)** (a), (b) and (c)

2.4 What are the advantages and disadvantages of secondary data?

2.5 How would you evaluate secondary data?

2.6 What are the different types of questionnaires?

2.7 What factors you would keep in mind while desiginning a questionnnaire?

2.8 Differentiate among the following types of questions:

— Open-ended questions
— Dichotomous questions
— Multiple-choice questions.

2.9 What do you understand by pre-testing a questionnaire? Why is it necessary?

2.10 Explain the term 'coding'. How does it differ from 'editing'?

A Specimen Questionnaire

Questionnaire on the Introduction of a Product

1. Respondent's name _____
2. Address _____
   _____
3. Telephone no. (if any) _____
4. Age in years
   21 to 30 _____ (1) 31 to 40 _____ (2)
   41 to 50 _____ (3) 51 to 60 _____ (4)
   60 + _____ (5)
5. Level of Education
   Primary _____ (1) Middle _____ (2)
   Higher Secondary _____ (3) Graduation _____ (4)
   Post Graduation _____ (5)
6. Marital status
   Single _____ (1) Married _____ (2) Widowed/Divorced _____ (3)
7. Occupation
   Govt. service _____ (1) Pvt. service _____ (2)
   Industry _____ (3) Self-employed _____ (4)
   Any other (please specify)_____ (5)
8. Family income per month (Rupees)
   Below 3,000 _____ (1) 3,000–6,000 _____ (2)
   6,000–9,000 _____ (3) 9,000–12,000 _____ (4)
   12,000–15,000 _____ (5) 15,000 + _____ (6)
9. Have you heard of this product before?
   Yes _____ (1)
   No _____ (2)
10. If yes, from whom?
    Neighbours _____ (1) Friends _____ (2)
    Relatives _____ (3) Newspapers/ magazines _____ (4)
    Radio _____ (5) Television _____ (6)
    Seen in the store _____ (7) Other (please specify) _____ (8)
11. Have you already used this product?
    Yes _____ (1) No _____ (2)
12. Do you currently use this product?
    Yes _____ (1) No _____ (2)

13. If you have already used this product, how would you rate it?
    Extremely good _____ (1) Good _____ (2)
    Satisfactory _____ (3) Bad _____ (4)

    Extremely bad _____ (5)
14. In your opinion, the claims of this product are:
    Highly exaggerated _____ (1) Exaggerated _____ (2)
    Justified _____ (3) Found better than the claims _____ (4)
15. If you have not already used the product, you are:
    Most likely to use it in the near future _____ (1)
    Likely to use it _____ (2) Not likely to use it _____ (3)
    Most unlikely to use it _____ (4) Can't say _____ (5)
16. How long do you think this product has been in the market?
    Less than 6 months _____ (1) 6 months–1 year _____ (2)
    1–2 years _____ (3) 2–3 years _____ (4)
    Over 3 years _____ (5)
17. To what extent do you think that this product has been accepted or rejected by consumers?
    Widely accepted _____ (1) Slightly rejected _____ (4)
    Moderately accepted _____ (2) Moderately rejected _____ (5)
    Slightly accepted _____ (3) Widely rejected _____ (6)
18. In case you would like to provide any additional information about this product, please do so in the space given below.

    _____
    _____
    _____

The McGraw-Hill Companies

CHAPTER 3 TABULATION OF COLLECTED DATA

Learning Objectives

After reading this chapter you would understand,

- how to arrange raw data in an array, and then classify data to construct frequency table and cumulative frequency table
- how to transform frequency tables into relative frequency and percentage distribution
- the main parts of a statistical table
- the general rules for the construction of tables
- for a given set of data, the type of table that would be most appropriate.

3.1 INTRODUCTION

In Chapter 1 we explained the nature of statistical study, while in Chapter 2 we looked into the collection of primary and secondary data, emphasizing the importance of collection of data. Here, we shall see how the collected data can be tabulated.

3.2 THE DATA ARRAY

When we collect data, there are not one or two observations but a large number of them. Even the data array, particularly when the number of observations is very large, is not very helpful for any analysis. Thus, it becomes necessary to organise the mass of data so that they are reduced to meaningful proportions. This brings us to tabular representation.

Suppose there is a class of 30 management students and each student in the class is asked to toss a coin five times and record each time whether he gets a head. As a result of this experiment, the following figures emerge for the 30 students:

3, 2, 0, 4, 1, 2, 3, 2, 5, 3, 3, 1, 1, 3, 5, 4, 2, 2, 3, 1, 0, 4, 3, 2, 2, 4, 2, 3, 3, 1

It is seen that even with only 30 observations the data need some better display. One way of doing this is to show the data in a certain order. For instance, we may show the same data in an ascending order as follows:

0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5

The same data could have been arranged in descending order in which case the figure 5 will appear first and 0 the last. The arrangement of data in ascending or descending order is called an *array*.

Advantages of Array One advantage of an array is that one can immediately find the range, that is, the difference between the highest and, the lowest value. Another advantage is that it can reveal at a glance whether there is a tendency of some items to concentrate near a certain value. In our example, we find that there is greater concentration of number three than any other number. Despite these advantages, array is not frequently used because arranging observations in accordance with their respective values is not only a monotonous and cumbersome job but it needs a good deal of time as well.

3.3 THE FREQUENCY TABLE: DISCRETE SERIES

It will be noticed that the data relating to the toss of a coin five times by 30 students show that each figure 0 to 5 has occurred a certain number of times. The frequency of their occurrence varies in each case. The same data can be shown differently. We can condense these data by pairing each of that value with its *frequency*. This is shown in Table 3.1.

Table 3.1 Frequency Distribution for Number of Heads Obtained by 30 Students

Observation (x)	*Frequency (y)*
0	2
1	5
2	8
3	9
4	4
5	2

The arrangement and display of data in this form where the observed value is paired with its frequency is called a *frequency distribution*. It may be noted that the frequency distribution does not show the details as are shown in an array as it presents the data in a summarised form. It enables the statistician to understand the data in a more meaningful manner. A major advantage of the frequency distribution is that it enables the statistician to obtain further characteristics of the variables with the help of various statistical techniques that will be discussed in Chapters 6, 7 and 8.

It can be seen from Table 3.1 that the largest number of frequencies are against observation 3 while the lowest number of frequencies are against observation 0 and 5. Sometimes the frequency with which a particular observation is observed is less revealing as compared to its *relative* frequency.

The Concept of Relative Frequency This is very simple; the concerned frequency is divided by the number of observations or the total frequencies. In our example in Table 3.1 the total number of frequencies is 30. By dividing each observation from 0 to 5 by 30, we can get the relative frequency in each case. This is shown in Table 3.2.

Table 3.2 Relative Frequency Distribution for Number of Heads Obtained by 30 Students

Observation	*Frequency (y)*	
(x)	*(f/N)*	*(f/N) × 100*
0	0.07	7
1	0.17	17
2	0.27	27
3	0.30	30
4	0.12	12
5	0.07	7
Total	**1.00**	**100**

We may convert the relative frequencies given in fractions to whole numbers or integers by multiplying them by 100. This is shown in column 3 of Table 3.2. The total of column 2 is obviously 1 while that of column 3 is 100. A frequency distribution in which every observed value is paired with its relative frequency is known as a *relative frequency distribution*.

As a frequency distribution facilitates us in condensing a large mass of data, the process of data analysis and interpretation also becomes far more manageable than it would have been otherwise.

In the section on Types of Data, we distinguished between the discrete variable and the continuous variable. It may be reiterated that discrete variable can have only specific values in a given class interval. In contrast, a continuous variable can take any value in a given class interval. Although there is thus a distinct difference between a discrete variable and a continuous variable, in practice, data collected on a continuous variable look like the data pertaining to a discrete variable.

The classification schemes can be either qualitative or quantitative and either discrete or continuous. We now turn to the formation of a frequency table when the data are continuous. However, before attempting this, we should have a clear idea of certain terms, which we shall come across frequently.

Class Limits These are the lowest and the highest values of a class. For example, take the class 30–50. Here, we find that the lowest limit is 30 and the highest limit is 50. When we categorise individual observations within this class, it is clear that none of the included observations is below 30 or above 50. Take another example; a class 60–79 indicates that no value below 60 can be included here and, likewise, no value above 79 can be included.

Class Interval The difference between the upper limit and the lower limit of a class is known as a class interval. It is the width of the class. Thus, the class interval of class 30–50 is 20. The class 60–79 has a class interval of 19.

Class Frequency The number of observations belonging to a particular class is known as the frequency of that class or the class frequency. Suppose there are 20 students who have obtained marks ranging from 30–40 and 44 students have obtained marks ranging from 50–60. In the first case, the class-interval 30–40, the class frequency is 20, while in the second case, in the class interval 50–60, the class frequency is 44. If we add up all the class frequencies, we get the total frequency of the entire series. For instance, if there are 500 students who have been classified in varying class intervals in accordance with their marks in an examination, then the total frequency is 500.

Class mid-point When we add up the lower and the upper class limits of a class interval, we get a certain value. This value is divided by two, which gives us the class mid-point. Thus, the mid-point of class interval 40–60 is (40 + 60) ÷ 2 = 50. The formula for obtaining class mid-point is as follows:

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

As we shall see subsequently, the mid-point of each class-interval is taken to represent it for the purpose of statistical calculations. One should also know that there are two methods of classifying data into different class-intervals. These are **(a)** *exclusive method* and **(b)** *inclusive method*.

Exclusive Method In order to explain this method, let us take some hypothetical data, which are given below:

Profits Earned by Companies

Profits (Rs in lakh)	*Number of Companies*
10–20	12
20–30	17
30–40	30
40–50	25
50–60	16
Total	**100**

This is an exclusive method where the upper limit of one class is shown as the lower limit of the next class. This ensures continuity of data. For example, a company earning a profit of Rs 30 lakh will not be classified in the class-interval 20–30. It will be included in the next class-interval: 30–40. In a field survey, proper instructions should be given to the enumerators where to put a certain response otherwise some enumerators will include a company earning Rs 30 lakh in the 20–30 class while others in the 30–40 class. Obviously, this will lead to erroneous results. It should be noted that if the class intervals are given as shown in the above table, it is presumed that the upper limit is exclusive and that the item of that value is included in the next class interval.

Inclusive Method In the case of inclusive method, the upper limit of one class is included in that class itself. Suppose we have the following frequency distribution:

Profits (Rs in lakh)	*Number of Companies*
10–19	12
20–29	17
30–39	30
40–49	25
50–59	16
Total	**100**

In this case, the class intervals are formed on the basis of inclusive method, which shows that the upper limit of the class interval in reality is taken in that class. Thus, a company earning Rs 30 lakh will be in the class 30–39 and another one earning Rs 39 lakh will also be in the same class interval. When a company earns a profit of say, Rs 30.6 lakh, where it should be included? Here we have to use the mid-points of the class interval. We may use the exclusive method and then calculate the mid-points. Thus, we may have classes 10.5–20.5, 20.5–30.5 and so on. The company earning Rs 30.6 lakh will be included in 30.5–40.5 class interval. Another way of classifying it would be 9.5–19.5, 19.5–-29.5 and so on.

At this point one may ask: How to decide whether inclusive or exclusive method is to be used? One has to first find out whether the variable under observation is a continuous or discrete. If the variable is a continuous one, then the exclusive method should be used. If, on the other hand, there is a discrete variable, then the inclusive method should be used.

Having made certain terms clear, we now turn to the formation of a grouped frequency table.

3.4 FORMATION OF A GROUPED FREQUENCY TABLE

The formation of a frequency distribution table comprises the following steps:

1. Deciding the appropriate *number* of class groupings
2. Choosing a suitable size or *width* of a class interval
3. Establishing the *boundaries* of each class interval
4. Classifying the data into the appropriate classes
5. Counting the number of items (i.e. frequency) in each class.

Steps 4 and 5 are purely mechanical. The first three steps assume considerable importance and are discussed below.

Deciding the Appropriate Number of Class Groupings The first and the foremost question one confronts is: How many classes should be formed?

The number of class intervals depends mainly on the number of observations as well as their range. As a general rule, the number of classes should not be less than six nor should be more than 15. If the number of observations is small, obviously the classes will be few as we cannot classify small data into 12 or 15 classes. If the classes are too few, then the original data will be so compressed that only limited information will be available. There is, however, Struges' formula available for guidance. The number of classes can be determined by applying Struges' formula, which is as follows:

$$n = 1 + 3.322 \log_{10} N$$

where n = *number of classes (rounded to the next whole number)*
N = *the total number of observations.*

For example, if the total number of observations is 100, then the number of classes would be 1 + 3.322 (2) = 7.644 or 8. In practice, the number of classes is determined keeping in mind the requirement in a given problem. It would, therefore, vary from problem to problem and the statistician has to decide as to how many classes should be formed in a particular problem.

Classes should be chosen in such a way that they cover all the data to be categorised.

Choosing a Suitable Size or Width of a Class Interval Another major consideration while forming a frequency table is the size of the class width. It is desirable to have each class grouping

of equal width. In order to ascertain the width of each class, the difference between the highest value and the lowest value, which is known as the *range,* should be divided by the number of class groupings desired:

$$\text{Width of class interval} = \frac{\text{Highest value} - \text{Lowest value}}{\text{Number of class groupings}}$$

Suppose we have 200 observations pertaining to income of workers where the highest income is Rs 6,200 p.m. and lowest is Rs 1,400 p.m. We have decided that the number of class groupings should be 6.

$$\text{Width of class interval} = \frac{6,200 - 1,400}{6} = 800$$

At times this value may be in fraction. For the sake of convenience, the width of each class grouping is rounded up to 5.

As far as possible, the class intervals should have equal width. For example, instead of having the class intervals as 1–7 and 8–10, we should have 1–5 and 6–10. Normally, the range should be a multiple of 5, 10 or 100 as it facilitates in computation. Occasionally, it becomes necessary to have unequal width in class intervals, particularly when there is considerable variation in the individual observations. For example, while dealing with data on income in a certain territory, we may find some extremely high values occurring infrequently while most values are within low or moderate range. In such a case, it is preferable to have larger width in the higher range of class intervals as compared to the width in the lower range values. One has to be careful in dealing with such tables having varying class intervals. In particular, one should ensure that graphs and charts based on such data do not give a distorted picture to the reader.

Establishing the Boundaries of the Classes The next step in the formation of a frequency table is to decide *class boundaries* for each class-interval so that observations can be placed into one class only. The point to note is that classes should not be overlapping as it would cause confusion and an observation could be included sometimes in one class and at other times in another class. Suppose we have class boundaries as follows:

0 –5
5 –10
10 –15

In such a case, an observation with value 5 can be included either in 0–5 or 5–10 class. This should be avoided. Instead, we may put the class boundaries as:

Below 5
5 to under 10
10 to under 15

The class boundaries, thus formed, are clear and there will not be any cause of confusion by placing individual observations into different classes where they should belong.

Example 3.1 The following data pertain to weights (in kg) of 33 students of a class:
42, 74, 40, 60, 82, 115, 41, 61, 75, 83, 63, 53, 110, 76, 84, 50, 67, 65, 78, 77, 56, 95, 68, 69, 104, 80, 79, 79, 54, 73, 59, 81 and 110.
Prepare a suitable frequency table.

Solution Here, we find that the highest value is 115 and the lowest is 40. We can decide the width of the class-interval. Since the range is 115 – 40 = 75, we may decide there should be 8 classes starting from 40. Thus, class intervals would be as shown in Table 3.3.

Table 3.3 Classification of Students by Weight

Class Intervals	*Tally Marks*	*Frequencies*
40 to under 50	|||	3
50 to under 60	𝍸	5
60 to under 70	𝍸 ||	7
70 to under 80	𝍸 |||	8
80 to under 90	𝍸	5
90 to under 100	|	1
100 to under 110	|	1
110 to under 120	|||	3
Total		**33**

We have thus allocated different observations to their respective classes with the help of tallies. The figures on the extreme right are the frequencies.

The main advantage of this frequency distribution table is that it highlights some important characteristics of the data. We find that the range of the data is 75 (115 – 40). Another point to note is that there is a concentration in the class-interval 70 to under 80 as it contains the highest frequency.

A major limitation of grouped frequency distributions is that some information is lost in this process. We cannot know how the values are within a particular class interval without going to the original source.

A point worth noting is that the selection of class boundaries for frequency distribution tables is very subjective. If the same set of data is given to two persons who are asked to form a frequency table, they may use different width and number of class intervals. Thus, the two frequency tables based on the same set of original data may be very different looking. This may be due to shift in the data concentration in the two frequency tables. However, in case the original data relate to a large number of observations, the effect of changes in the class boundaries on the concentration of frequencies would be far less as compared to the case when the data are extremely limited.

3.5 RELATIVE FREQUENCY AND PERCENTAGE DISTRIBUTIONS

Our discussion so far was confined to absolute frequencies. We can transform the frequency distribution into a *relative frequency distribution.* The relative frequency may be obtained from

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Total number of observations}}$$

Thus, for the data presented in Table 3.3, the relative frequency for the class interval 70 to under 80 is obtained from

$$\text{Relative frequency} = \frac{8}{33} = 0.242$$

This relative frequency is the proportion to the total number of observations. By multiplying it by 100, we can change it as a percentage to the total number of observations. This comes to 24.2 per cent. In other words, the absolute frequency of 8 in the total number of observations of 33 is 24.2 per cent of it. Let us take the following example.

Example 3.2

Table 3.4 A Relative Frequency Table

Class Interval	*Frequency*	*Relative Frequency*
10 but less than 20	3	0.055
20 but less than 30	5	0.091
30 but less than 40	9	0.164
40 but less than 50	18	0.327
50 but less than 60	10	0.182
60 but less than 70	8	0.145
70 but less than 80	2	0.036
Total	**55**	**1.000**

The foregoing table can be converted into a percentage frequency distribution by dividing frequency in each class by the total number of observations and multiplying by 100. This percentage distribution is shown in Table 3.5.

Table 3.5 Percentage Frequency Distribution

Size of Item	*Percentage Frequency*
10 but less than 20	5.5
20 but less than 30	9.1
30 but less than 40	16.4
40 but less than 50	32.7
50 but less than 60	18.2
60 but less than 70	14.5
70 but less than 80	3.6
Total	**100.0**

It may be noted that at times the use of relative frequencies is more appropriate than absolute frequencies. Whenever two or more sets of data contain different number of observations, a comparison with absolute frequencies will be erroneous. In such cases, it is necessary to use the relative frequency. Also, the class boundaries should be the same (or at least be multiples of each other) if proper comparisons are to be made.

Let us take another example.

Example 3.3 The data given below relate to the number of years that 50 workers of a small factory have worked for:

1.4	2.4	0.6	5.1	4.1	4.8	10.9	3.9	11.6	0.9
11.0	8.6	4.4	0.8	5.7	2.3	1.3	7.6	9.3	14.4

5.4	6.9	8.6	3.2	10.6	6.8	7.1	8.4	2.1	11.3
0.4	4.9	8.2	10.8	15.0	9.3	2.3	0.7	3.9	6.2
2.2	5.7	13.8	10.1	0.7	3.2	4.6	9.8	3.9	2.7

Construct a frequency distribution starting from 0 and under 2 and having equal width in each class.

Solution

Table 3.6 A Frequency Distribution with Equal Width Classes

Class-intervals	*Tally Marks*	*Frequencies*
0 and under 2	𝍸 |||	8
2 and under 4	𝍸 𝍸 |	11
4 and under 6	𝍸 ||||	9
6 and under 8	𝍸	5
8 and under 10	𝍸 ||	7
10 and under 12	𝍸 ||	7
12 and under 14	|	1
14 and under 16	||	2
Total		**50**

3.6 CUMULATIVE FREQUENCY DISTRIBUTION

At this stage, we may introduce another concept relating to frequency distribution. This is known as *cumulative frequency distribution* or simply *cumulative distribution*. Table 3.7 gives such a distribution along with original distribution, which is non-cumulative.

Table 3.7 Cumulative Frequency Distribution

Marks Obtained	*Number of Students*	*Marks Obtained*	*Cumulative Frequency*
1	10	Not more than 1	10
2	30	Not more than 2	40
3	35	Not more than 3	75
4	28	Not more than 4	103
5	39	Not more than 5	142
6	20	Not more than 6	162
7	20	Not more than 7	182
8	12	Not more than 8	194
9	6	Not more than 9	200
10	0	Not more than 10	200
Total	**200**		

It may be noted that instead of 'not more than' we could have used the words 'less than'. Such a cumulative frequency table is helpful for knowing the number of students who have scored a certain level of marks. Suppose we want to know the number of students who have scored up to 4 marks, the

answer is obvious—103, as shown in the above table. Suppose we want to know the number of students who have scored 60 percent and above marks. Here, from the total number of students, 200, we have to subtract 142, that is, the number of students scoring up to 50 percent marks. This gives us the number of students scoring 60 percent and above marks as 58.

Example 3.4 Let us take another example. Suppose we have a grouped frequency distribution as follows:

Class-interval	0–10	10–20	20–30	30–40	40–50
Frequency	7	10	18	9	6

We can transform this distribution into a cumulative frequency distribution as shown in the following table:

Table 3.8 Cumulative Frequency Distribution

Class-interval	*Cumulative Frequency*
Less than 10	7
Less than 20	17
Less than 30	35
Less than 40	44
Less than 50	50

The cumulative frequency can be converted into percentages as is given in Table 3.9.

Table 3.9 Cumulative Percentage Frequency Distribution

Class-interval	*Percentage Frequency*
Less than 10	14
Less than 20	34
Less than 30	70
Less than 40	88
Less than 50	100

The cumulative percentage distribution can be used in finding the percent of observations falling below or above a specified value of interest. Thus from Table 3.9, we know that 70 percent of the observations are below 30; only 12 percent of the observations lie between sizes 40 and 50; and so on.

3.7 TWO-WAY AND THREE-WAY FREQUENCY DISTRIBUTION

Sometimes two separate frequency distributions but related with each other are combined in the form of a two-way frequency distribution. Such a distribution is also called a *cross-tabulation*, which is helpful in analysing the relationship between two variables. In constructing cross-classification tables, one has to decide which data should be given primary emphasis and which should be given secondary emphasis. The rule to follow is that data with primary emphasis are given in columns while those with secondary emphasis are shown in rows. An example of cross-classified table is given as follows.

Table 3.10 Break-up of Demand by Price Variations

(Number of units)

Price per Unit (in Rs)	*0–10*	*10–20*	*20–30*	*30–40*	*Total*
Less than Rs 5	—	—	32	48	80
5–10	—	28	35	—	63
10–15	35	22	—	—	57
Total	**35**	**50**	**67**	**48**	**200**

A close look into the figures given in the above table indicates that when price is low, the demand is high and *vice versa*. In other words, there is an inverse relationship between price and demand. However, such a table is unable to measure the statistical relationship between the two variables.

Table 3.10 can be transformed into percentages and then these percentages can be shown either as a separate table or side-by-side with the original data. The data can be percentaged in either direction; one has to decide which base should be used as 100 percent. Percentages should be based on totals of rows or columns, whichever is relatively more important. Another example of a two-way classification is given in Table 3.11, which is a blank table.

Table 3.11 Two-way Classification

Population (in lakh)	*Years*				
	1961	*1971*	*1981*	*1991*	*2001*
Rural					
Urban					
Total					

This table can be extended to a three-way classification as follows:

Table 3.12 Three-way Classification

Population (in lakh)	*Years*				
	1961	*1971*	*1981*	*1991*	*2001*
Rural Total					
Males					
Females					
Urban Total					
Males					
Females					
Total					

As was mentioned earlier, the data with primary emphasis are put in columns and those with secondary emphasis are put in rows; for higher-order tables this process is repeated. In the above two tables, it will be seen that years are given in columns and classification of population into rural and urban is

shown in rows. This indicates that the table has assigned primary emphasis to the decennial change in the population and the secondary emphasis to the rural and urban break-up of population as also its break-up by sex.

A far more elaborate classification of workers in an industrial enterprise is shown in Table 3.13, which is a blank table. This table shows a number of characteristics of the workers.

(i) Classification of workers as skilled, semi-skilled and unskilled
(ii) Classification of workers by sex
(iii) Classification of workers by length of service put in by them
(iv) Classification of workers by age groups

It should be clear from the table that these four characteristics of workers are not shown in isolation. They are shown in an integrated manner such that one can derive any information as needed from the table. It should be obvious that the formation of such a table is not easy as it can be formed in more than one way. One should be clear as to which feature or features are to be given prominence in the format.

Table 3.13

Workers	*Skilled*			*Semi-skilled*			*Unskilled*			*Total*		
	M	*F*	*T*	*M*	*F*	*T*	*M*	*F*	*T*	*M*	*F*	*T*
Working for less than 2 years												
Ages												
20–30												
30–40												
40–50												
50 and above												
Working for a period of 2–5 years												
Ages												
20–30												
30–40												
40–50												
50 and above												
Working for a period over 5 years												
Ages												
20–30												
30–40												
40–50												
50 and above												
Total												

M = Males F = Females T = Total

3.8 MAIN PARTS OF A STATISTICAL TABLE

In the preceding pages we have seen how raw data are transformed into suitable statistical table. In order to have proper tabulation, it is necessary to know the main parts of a table. These are explained below:

Guidelines Regarding Structure of a Table

1. Table Number Every table should be numbered so that it can be identified. The number is normally indicated at the top of the table.

2. Title Each table must bear a title indicating the type of data contained. The title should not be very lengthy so as to run in several lines. It should be clear and unambiguous.

3. Captions and Stubs A table consists of rows and columns. The headings or sub-headings given in columns are known as captions while those given in rows are stubs. It is necessary that a table should have captions and stubs to indicate what columns and rows stand for. It is also desirable to provide for an extra column and row in the table for the column and row totals.

4. Main Body of the Table As this part of the table contains data, it is the most important part. Its size and shape should be suitable to accommodate the data. The data are entered from the top to the bottom in columns and from left to right in rows.

5. Ruling and Spacing A good table should invariably have proper ruling and spacing. Vertical lines are drawn to separate columns but horizontal lines are normally not drawn except in case of totals which must be separated from the main body of horizontal lines. In addition, the horizontal lines are drawn both at the top just below the title of the table as also at the bottom. Due care must be exercised in providing proper space in accordance with the dimensions or magnitude of figures.

6. Head Note A head note is invariably given just below the title of a table indicating the unit of measurement applicable to the data displayed. This is generally shown within brackets.

7. Footnote Sometimes it becomes necessary to give some explanation for the data used or to explain the meaning of the abbreviation used. This should be given in the form of a footnote at the bottom of the table just below the last horizontal line.

8. Sourcenote Whenever secondary data are used, it is necessary to show the source from which such data are taken. This is also in the form of a footnote. The purpose of providing the sourcenote is to facilitate the reader to refer to that source if he so wants.

A blank table, given below, shows the main parts of a table as explained above.

Table No. _____
Title_____

Stub Heading	*Captions*		*Captions*		*Total*
	Caption	*Caption*	*Caption*	*Caption*	
Stub Entries		Main Body			
Total					

Footnote:
Source:

Having looked into the main parts of a statistical table, we should now be familiar with certain general rules for constructing a suitable table. These are briefly discussed below.

3.9 RULES FOR TABULATION

1. The purpose of tabulation is to condense the mass of data and to make it understandable. As such, the table should not be overloaded with data and it should be as simple as possible. If too much data are given in one table alone, it will defeat the very purpose of its construction. It will be too difficult to understand and interpret the data.
2. The table should serve the purpose or objective of the investigation. Sometimes one finds that a table has been constructed but it has not been put to any use. Such tables should be avoided.
3. A table should be complete in all respects including the unit of measurement, the time period for which the data relate, abbreviations used, footnote and source note, if necessary.
4. The data presented in the table should be free from any inconsistency or inaccuracy. It is desirable to go over the data again to ensure that they are accurate and properly recorded from the raw data.
5. The table should contain, as far as possible, totals, ratios and percentages, which will provide a better understanding of the data shown in it.
6. Finally, the table should not be made in a hurried and haphazard manner. Care should be taken to ensure that the format of the table is appropriate with reference to the nature of the data to be presented therein. Further, it should be made attractive so that the reader feels interested to go through it.

GLOSSARY

Class boundary	The upper and lower limits of each class interval.
Class frequency	The number of values in a data set that belongs to a certain class.
Class-interval	An interval that includes all the values in a data set that fall within two numbers, the lower and the upper limits of the class.
Classification	The process of arranging data into sequences and groups according to their common characteristics.
Continuous data	Quantitative data concerning a parameter in which all measured values are possible.
Cross-section data	Data collected on different elements at the same point in time or for the same period of time.
Cumulative frequency distribution	A frequency distribution showing how many observations lie below or above certain values.
Cumulative frequency	The frequency of a class that includes all values in a data set that fall below the upper boundary of that class.
Data array	Raw data arranged in either an ascending or a descending order.
Data point	A single observation from a data set.
Data set	A collection of data.

Discrete data	Data not available on a continuous scale.
Discrete series	A series of data where the data are distinct numbers with gaps between them.
Frequency distribution	The arrangement and display of data wherein the observed value is paired with its frequency.
Frequency	Any item that occurs a specific number of times.
Primary data	Data collected specifically for the study currently undertaken.
Raw data	Any data before they are processed or analysed by statistical methods.
Relative frequency	Frequency in a category divided by total number of observations.
Relative frequency distribution	A frequency distribution in which every observed value is paired with its relative frequency.
Secondary data	Data collected by a previous study.
Tabulation	Presentation of data collected in an orderly and summarised form.
Time-series data	Data collected on the same element for the same variable at different points of time or for different periods of time.
Width of class-interval	Difference between the highest value and the lowest value divided by the number of class intervals.

QUESTIONS

3.1 Given below are seven statements. Indicate in each case whether it is true or false.

(a) A discrete series always consists of whole numbers or integers.
(b) A class width of a frequency distribution should always be of equal size.
(c) A relative frequency distribution does not have all-inclusive and mutually-exclusive classes.
(d) A relative frequency can be obtained by dividing the frequency by the total number of observations.
(e) A two-way classification hardly reveals any more information than a one-way table.
(f) A table based on secondary data must always indicate the source of data.
(g) Too much of data given in one table may reduce its utility.

Multiple Choice Questions (3.2 to 3.10)

3.2 Which of the following is the most accurate method of classifying data?

(a) Qualitative method
(b) Quantitative method
(c) Method depending on the nature of information required for a particular situation/problem
(d) (a) and (b)

3.3 While tabulating the grouped data:

(a) Each group must have frequencies.
(b) Frequency can be negative.
(c) It is necessary to have frequencies.
(d) None of the above.

3.4 The first step in the formation of a frequency table is
- **(a)** Choosing a suitable size or width of class intervals
- **(b)** Establishing the boundaries of each class intervals
- **(c)** Deciding an appropriate number of class groupings
- **(d)** Classifying the data into the appropriate classes

3.5 In a relative frequency distribution, frequencies are in
- **(a)** Whole numbers
- **(b)** percentages
- **(c)** fractions
- **(d)** both (b) and (c)

3.6 The number of classes in a frequency distribution depends on
- **(a)** range of observations in the data set
- **(b)** size of the population
- **(c)** number of data points
- **(d)** (a) and (c)
- **(e)** (a), (b) and (c)

3.7 The upper limit of class intervals is required when we consider the formation of
- **(a)** relative frequency
- **(b)** 'more than' cumulative frequency
- **(c)** 'less than' cumulative frequency
- **(d)** none of the above

3.8 Which of the following is *not* an example of compressed data?
- **(a)** A frequency curve
- **(b)** Frequency distribution
- **(c)** Data array
- **(d)** Histogram

3.9 In a frequency distribution, classes are all inclusive because
- **(a)** the given data must fit into one class or another
- **(b)** none of the data points falls into more than one class
- **(c)** there are always more classes than data points
- **(d)** all of these

3.10 While preparing a frequency table, which of the following number of classes is generally used?
- **(a)** less than five
- **(b)** between 5 and 10
- **(c)** between 10 and 20
- **(d)** none of these

3.11 What is meant by 'classification'? State its important objectives.

3.12 Describe different kinds of classification, giving one example in each case.

3.13 "Tabulation of data is a very important method of presenting data." Elaborate the statement and describe different parts of a statistical table.

3.14 What do you understand by 'tabulation'? What considerations should be kept in mind while tabulating the data?

3.15 What are the uses of tabulation? Describe main parts of a statistical table.

3.16 Which of the following do you think are discrete and which are continuous variables?

(a) The weight of bags of fruit-drops (normally 100 gm and 200 gm bags) that are filled by an automatic machine.

(b) The number of fruit-drops in these bags.

(c) The stock-level of a retail shoe dealer.

(d) The stocks of grain held by a wholesale animal feed merchant.

3.17 You have been given some raw data and asked to present them in a suitable frequency distribution with some class-intervals. Explain what factors would you consider in deciding the number of classes and the lower and upper limits of class-intervals for the proposed frequency distribution.

3.18 The following numbers are given:
20, 25, 27, 11, 9, 30, 32, 29, 17 and 23
Arrange them in an array and determine the range.

3.19 In a survey it was found that 64 families bought milk in the following quantities (litres) in a particular month:

19	22	09	22	12	39	19	14	23	06	24	16	18
07	17	20	25	28	18	10	24	20	21	10	07	18
28	24	20	14	24	25	34	22	05	33	23	26	29
13	36	11	26	11	37	30	13	08	15	22	21	32
21	31	17	16	23	12	09	15	27	17	21	16	

Using Struges' rule, convert the above data into a frequency distribution by 'inclusive method'.

3.20 A market survey collected response from 50 persons regarding acceptability of a new product. The scores of the respondents on appropriate scale are as follows:

40	45	41	45	30	39	08	48	25	45
26	11	23	24	13	29	08	40	41	42
39	35	18	25	35	40	42	43	44	36
37	32	28	27	25	26	38	37	40	35
32	28	40	41	43	44	45	40	39	41

Prepare a frequency table.

3.21 Prepare a blank table in which the following types of information can be shown properly:

(i) Total number of workers by sex.

(ii) Classification of male and female workers by skill, that is, skilled, semi-skilled and unskilled.

(iii) Classification of workers according to broad age groups, that is, below forty years and above forty years.

3.22 Suppose you have to show in a proper tabular form the following information relating to a certain city:

(i) Number of males and females.

(ii) Number of children, working adults and old persons.

(iii) Classification of population by five areas of residence, viz., north, south, east, west and central.

Prepare a blank table to show the information.

3.23 The marks obtained by 30 students in Statistics test are given below:

42	48	57	31	40	30	18	52	59	45
65	67	29	22	72	62	60	58	55	40
18	25	44	61	75	77	48	32	26	50

Prepare a frequency table and a cumulative frequency table. Having prepared the tables, answer the following:

(i) The highest marks, the lowest marks and the range.
(ii) How many students received 75 marks and above?
(iii) How many students received marks below 40?
(iv) What percentage of students passed this test, taking 40 marks as the minimum marks required to pass the test?
(v) What percentage of students secured first class, taking 60 and above marks for getting the first class?

3.24 The following table shows a frequency distribution of the monthly wages in rupees of 50 employees of a company:

Wages in rupees	1000–1500	1500–2000	2000–2500	2500–3000	3000 +
No. of employees	6	11	15	10	8

With reference to this table determine:

(i) The lower limit of the third class
(ii) The upper limit of the fourth class
(iii) Percentage of workers earning between Rs 2000 and Rs 2500
(iv) Percentage of workers earning below Rs 2500.

3.25 From the table given in question 3.24, prepare a cumulative frequency table.

3.26 From the table given in question 3.20, construct an 'or more' cumulative frequency distribution.

3.27 A survey of 370 students from Commerce Faculty and 130 students from Science Faculty revealed that 180 students were studying for only CA examination, 140 for only Costing examination and 80 for both CA and Costing examinations. The rest had opted for part-time Management courses. Of those studying Costing only, 13 were girls, and 90 boys belonged to Commerce Faculty. Out of 80 studying for both CA and Costing, 72 were from Commerce Faculty amongst whom 70 were boys. Amongst those who opted for part-time Management courses, 50 boys were from Science Faculty and 30 boys and 10 girls from Commerce Faculty. In all there were 110 boys in Science Faculty.

Present the above information in a tabular form, showing the break-up of students (girls and boys taken together) for commerce and science faculties by different courses.

3.28 Tabulate the following:

A super market divided into five main sections—grocery, vegetables, medicines, textiles and novelties, recorded the following sales in 1991, 1992 and 1993.

In 1991, sales in grocery, vegetables, medicines and novelties were Rs 6,25,000, Rs 2,20,000, Rs 1,88,000 and Rs 24,000 respectively. Textiles accounted for 30 percent of the total sales during the year. In 1992, the total sales showed 10 percent increase over the previous year. While grocery and vegetables registered 8 percent and 10 percent increase over

their corresponding figures in 1991, medicines dropped by Rs 13,000, textiles stood at Rs 5,36,000. In 1993, though the total sales remained the same as in 1992, grocery declined by Rs 22,000, vegetables by Rs 32,000, medicines by Rs 10,000 and novelties by Rs 12,000.

3.29 The regional transport authority is concerned about the speed of motor bikes college students are driving on a section of the road. Following are the speeds of 45 drivers (speed in km/hr).

15	31	44	56	38	32	48	42	58	29	45	
49	38	48	62	46	56	52	47	49	42	52	
55	52	69	39	39	58	57	18	68	48	62	
64	61	47	69	58	29	55	18	61	48	55	49

(i) Classify the data considering classes 15–24, 25–34, 35–44, …

(ii) The state report says that not more than 10% college students exceed speed 55 km/hr. State whether the collected data supports the given report or not.

3.30 The time required in minutes for each of the 50 students to read 20 pages of a book is recorded below:

42	52	47	49	36	48	41	47	50	32	45	48
40	43	48	36	51	44	49	53	37	34	42	47
45	47	44	50	31	48	43	45	44	36	49	51
43	53	46	39	50	42	42	47	38	51	46	40
38	45										

Classify the data by considering classes 30–34, 35–39, ….

3.31 Present the following information in a suitable table supplying the missing figures. “In the Lok Sabha there were 542 members. When a certain bill was put to vote, 306 voted as Ayes of whom 30 belonged to oposition benches. In all, 54 members abstained from voting, of whom 30 belonged to treasury benches. Out of the total of 236 members of opposition benches, 182 voted as Noes. The bill was passed as 306 Ayes against 182 Noes.”

The McGraw-Hill Companies

CHAPTER 4

GRAPHIC PRESENTATION OF DATA

Learning Objectives

After reading this chapter you would

- understand the guidelines for the use of graphs and diagrams
- know the various types of graphs
- make the most suitable choice of method of graphic presentations for a given set of data
- plot a set of data on varying types of graphs
- display a set of data by means of a graph that is most suitable in the given case
- understand and critically interpret graphic presentations of data attempted by others.

Chapter Prerequisites

Before starting work on this chapter, you should ensure that you are conversant with

1. the calculation of percentages
2. the plotting of data on graphs

4.1 INTRODUCTION

In the preceding chapter, we have seen how huge and scattered data can be condensed and presented in the form of suitable tables. It should be obvious to everyone that without proper tabulation of data, we cannot proceed further with the data, and they would remain unutilised. *Tabulation is one way of presenting data. Another way is in the form of diagrams and graphs*. In this and the next chapter, our focus is on these forms of presentation of statistical data. Before we look into the various types of diagrams and graphs that can be used in the presentation of data, we shall look into the importance and limitations of such presentation. In addition, we shall look into the qualities of a good presentation of data.

4.2 IMPORTANCE OF GRAPHIC AND DIAGRAMMATIC PRESENTATION

1. On account of their visual impact, the data presented through graphic and diagrammatic presentation are better grasped and remembered than the tabulated data.

2. These forms of presentation transform data in simple, clear and effective manner.
3. They are able to attract the attention of the reader particularly when several colours and pictures are used in presentation.
4. A major advantage of these presentations is that they have better appeal even to a layman. For the layman, simple charts, maps and pictures facilitate a much better undestanding of the data on which these are based.
5. Since they lead to a better understanding, they save considerable time.
6. Even when data show highly complex relations among variables, these devices make them much clear. They thus greatly facilitate in the interpretation and analysis of data.
7. As we shall see subsequently, these devices are extremely helpful in depicting mode, median, skewness, correlation and regression, normal distribution, time series analysis, and so on.

4.3 LIMITATIONS OF GRAPHS AND DIAGRAMS

Despite the foregoing advantages of graphs and diagrams in presenting statistical data, they are subject to certain limitations.

1. In presenting data by these devices, it is not possible to maintain 100 percent precision. As such these devices are not suitable where precision is needed.
2. These cannot be a complete substitute for tabulation. They can serve the purpose better when they are accompanied by suitable tables.
3. When too many details are to be presented, these devices fail to present them without loss of clarity.
4. In those cases, where mathematical treatment is required, these devices turn out to be extremely unsuitable.
5. Small differences in large measurements cannot be properly brought out by means of graphs and diagrams.
6. While diagrams and graphs are generally simple to understand, one should know that all graphic devices are not simple. Particularly when ratio graphs and multidimensional figures are used, these may be beyond the comprehension of the common man. A proper understanding of these figures needs some expertise on the part of the reader.

4.4 GUIDELINES FOR THE USE OF GRAPHS AND DIAGRAMS IN PRESENTING DATA

1. Title To begin with, a graph or diagram should always have a suitable title and, if necessary, a subtitle as well. The title shows the subject-matter while the subtitle shows the time period covered such as '1995 to 2000' or '1995–96 to 1999–2000', or unit of measurement such as 'rupees in millions', and so forth. It is desirable to provide a brief title as too long a title would be cumbersome and would not go well with the graph or the diagram. However, the brief title should not be at the cost of clarity. *The title should be distinctive from the rest of the chart. It should be given prominence by using* **bold letters** *either at the top or below the graph or the diagram.*

2. Scale The scale used for presentation of data should be appropriate keeping in mind the availability of space and the need for proper presentation. If the scale is very arbitrary and unsuitable, the

graph or diagram would be very odd in appearance and is unlikely to be understood properly. Both horizontal and vertical scales should be specified if they are different from each other. If, however, both the scales are on the same basis, then this should be specified at an appropriate place or on the graph itself.

3. Size and Proportion Another aspect that needs to be looked into while using these devices is the size and proportion. Both would depend mainly on the size of the paper and availability of space in the report. *A major guideline is that the size should neither be too large nor too short.* As regards the proportion of length and breadth, it is suggested that these should be in the ratio of $\sqrt{2} : 1$, that is, 1.414:1. Any side, horizontal or vertical, can be a longer side, which should broadly be a little less than 1.5 times the shorter side.

4. Footnotes At times, the graph or diagram needs some clarification, which obviously cannot be shown in the body of the graph or the diagram. In such cases, a footnote should be given below the chart. For example, if a chart shows the national income for different years, the figures for the last one or two years may be merely estimates unlike the figures for the earlier years, which are actual figures. This fact has to be specified by a footnote.

5. Sourcenote *When the graph or diagram is based on data taken from some external source, it is necessary to specify this source below the chart.* For example, if a chart has used some data on the total value of production and the number of workers employed from the Census of Indian Manufactures, then it must specify this source along with the concerned year. This apart, if any modification or adjustment has been made while using such data, this should also be adequately mentioned along with the source. This is very necessary so that any reader who would like to look into the original source can easily have an access to it.

6. Attractive Presentation Finally, a graph or diagram should have an attractive presentation. The purpose of drawing such a chart will be defeated if it is unattractive and the reader is disinclined to see it. *A chart can be made attractive if the statistician gives a due consideration to several aspects such as colour shades, size and proportion, proper lettering, and so on.* These considerations become all the more relevant in case the diagram happens to be a three-dimensional diagram. This is because, in general, three-dimensional diagrams are difficult to understand. As such, extra care has to be exercised in order to make them appealing and attractive.

4.5 WHICH TYPE OF GRAPH OR DIAGRAM TO BE USED

There are several ways by which statistical data can be presented. These techniques will be discussed shortly. But before we do so, we should know-how to choose a particular technique from amongst several of them. In choosing a proper technique, we should be guided by the following considerations.

1. Purpose A major consideration while choosing a particular technique is the purpose of presenting the data. The purpose may be merely the presentation of results, or illustration or analysis. For example, if we are dealing with the time series, the graphic device will be most suitable where we may measure time period on the horizontal scale and the concerned variable on the vertical scale. If the presentation is made for the use of the general public, the choice should be in favour of a very simple graph or diagram. In case the presentation is made before professional people or technologists, the technique could be comprehensive and more refined.

2. Circumstances of Use Another consideration is the circumstances under which a diagram is to be used. There are several ways in which it can be put to use. For instance, it may be used during the course of a lecture using a computer and power point slides, or it may be included in a book or report. Other possible uses could be to hang it on a wall or to put it on a table or to show it on a TV programme or in the movie. In all such cases, the choice will depend on the specific use as well as on the level of uncertainty of the user.

3. Subject-matter to be Presented The choice is also guided by the nature of the data to be presented. If original data are to be presented, a simple graph can be a proper choice, but if the data are in the form of ratios of change, then an appropriate device would be a semi-logarithm graph. Again, for discrete and continuous series, the techniques of presentation should be different. This apart, the magnitudes of data should be taken into consideration. When the data consist of wide differences, the two- or three-dimensional charts would be more suitable than other techniques. Likewise, on the nature of the data will depend whether one has to use a simple bar, a multiple bar or a component bar diagram.

4. Time and Resources Available Some charts are such that need a lot of time before they can be used. They also need more resources. A map or a multi-coloured pictorial presentation would need far more resources and time than the ordinary graphs and simple bar diagrams. Sometimes multi-coloured pictorial devices may need the services of a well-qualified cartographer or an artist. Besides, the services of a good printing press may be needed if such charts are to be included in published reports.

5. Type of Audience Finally, a major consideration before the statistician should be the type of audience before whom the presentation is to be made. Obviously, for the general public, more refined or elaborate devices would be most unsuitable. In case of such an audience, simple time-series graphs or simple bar diagrams would be more suitable. In contrast, where the audience is professionally qualified or otherwise quite competent, the statistician should prefer a more elaborate device.

We now turn to specific devices that are used in the presentation of data. They can be divided into two categories viz. *graphical devices* and *diagrammatic devices*. In this chapter, we discuss graphic devices, while the next chapter will focus on diagrammatic devices.

4.6 GRAPHIC DEVICES

There are two major categories of graphs—the *natural scale graph* and the *ratio scale graph*; the former is more frequently used. Within the natural scale graph, again there are two types: **(a)** *time series graph*; **(b)** *frequency graph.*

Time series graph, as the name implies, shows the data against time, which could be any measure such as hours, days, weeks, months and years.

Thus, a graph showing a number of industrial workers employed in a company for each of the years 1991–2000 is time series.

Some other variables such as income of employees and the number of employees earning that income, if plotted on a graph, will be known as a frequency graph. Graphs such as histogram, frequency polygon and the ogive curve are popular frequency graphs.

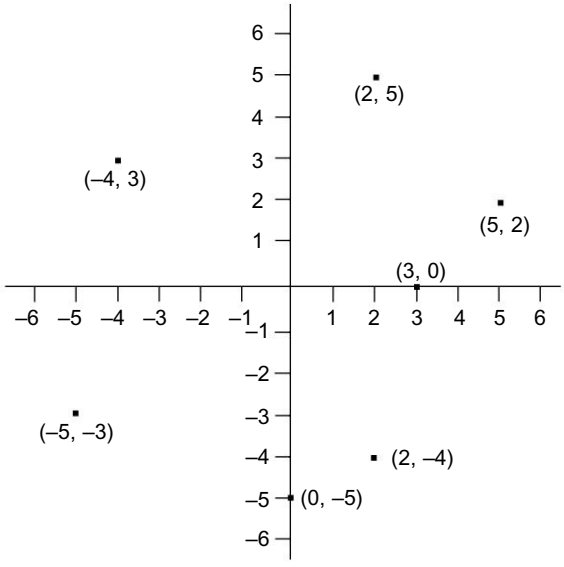


Fig. 4.1 **Four Quadrants of a Graph**

In this section, we shall discuss and illustrate with examples both types of graphs. However, we first start with the basic approach of dividing a graph sheet into four quadrants.

Let us first use the four quadrants of a graph. Suppose we have to locate on a rectangular coordinate system, the points having coordinates (a) 2, 5; (b) 5, 2; (c) –4, 3; (d) 2, –4; (e) –5, –3; (f) 0, –5; and (g) 3, 0. We assume that all the given numbers are exact. See Fig. 4.1.

In the same manner, we can draw graphs of equations. For example, we are given an equation $y = x^2 - 2x + 5$. Here, we assume certain values of x on the basis of which the corresponding values of y can be obtained. This is shown below.

Since all these values are positive, these points will be plotted on the first quadrant where both x and y have positive values only. This is shown in Fig. 4.2.

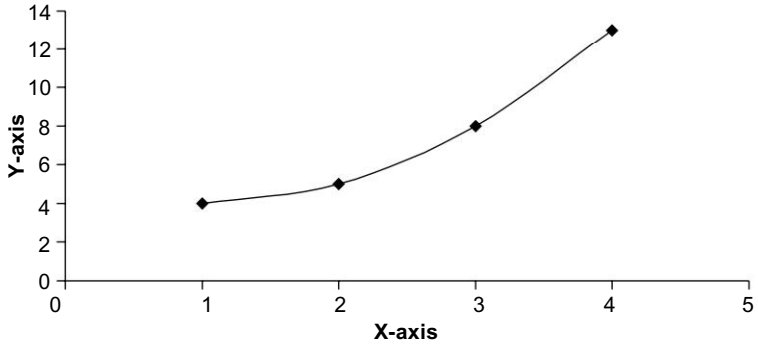


Fig. 4.2 **Graph of Equation $y = x^2 - 2x + 5$**

X	*Y*
1	4
2	5
3	8
4	13

Line Graph

Let us now move on to a time series graph. Here time period is measured along the X-axis and the corresponding values are on the Y-axis. Suppose we have the following time series:

Year	1990	1991	1992	1993	1994	1995	1996
Sales, (in '000 Rs)	5	8	13	10	15	17	20

The sales data are shown in Fig. 4.3 as given below:

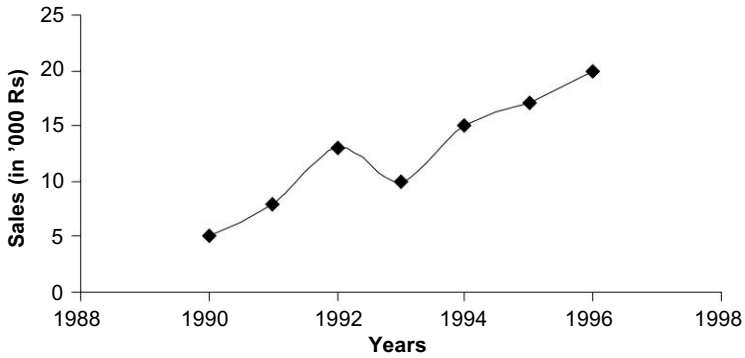


Fig. 4.3 **Graph Showing Sales from 1990–1996**

A graph of this type clearly shows that there was a decline in sales in 1993 while all other years, 1991 to 1996, showed an increase in the same.

It may be noted that the *time series graphs are very commonly used where one variable is the time factor*. The data on population, number of workers in an industrial unit or the entire industry, number of students enrolled in a college, profits earned by a company, sales of a given product are some of the examples where time series can be effectively used to present the given series. It may also be noted that in the same graph we can present two or more related series. For example, we have two series of data on exports and imports for the same years. Using a proper scale, which should be common for both the series, the export and import figures can be plotted on the same graph. Such a chart will give an idea of the gap or difference between the two series. A hypothetical example is given in Fig. 4.4.

Figures 4.3 and 4.4 both are the examples of line graph wherein the values of a given variable have been plotted against time (years). As mentioned earlier, the time could be a year, a month, a week, a day or even an hour. Such graphs are extremely simple though one has to be very careful in selecting a suitable scale depending on the range of data to be plotted against time. A multiple line graph showing two or more series on the same graph needs more care to accommodate varying values of the different series.

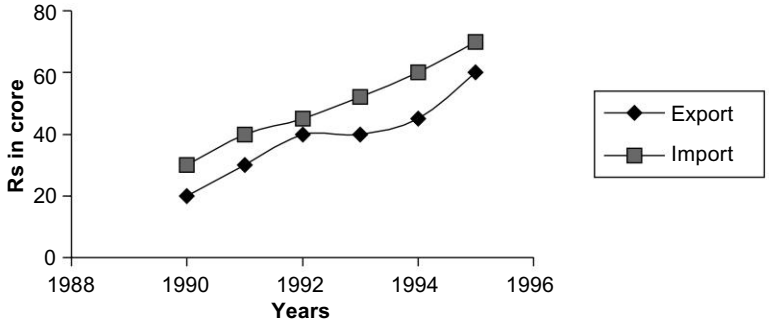


Fig. 4.4 **Graph Showing Exports and Imports from 1990–1995**

False Base Line

An important rule in the construction of graphs is that the scale of the Y-axis should begin from zero even when the lowest figure in the Y-series happens to be far above zero. However, sometimes we have to graph such data that it may be extremely difficult to adhere to this basic rule. This happens when the space available for the graph cannot accommodate the entire scale beginning from zero. An example will make this point clear.

Suppose the data given in Table 4.1 relating to the growth of population of a city during the decennial periods are to be shown by means of a graph.

Table 4.1 **Growth of Population of a Certain City**

Year	*Population (in lakh)*
1951	45
1961	53
1971	60
1981	73
1991	91
2001	115

If we take the starting point zero on the vertical scale (Y-axis), we find that we are unable to show all the data, particularly for 1991 and 2001, on account of limitations of space. As such, we use the false base-line to get over this difficulty. Figure 4.5 shows the data using the false base line.

The use of a false base line should be resorted to when it is absolutely necessary due to the non-availability of adequate space. Sometimes this device is used to magnify the minor fluctuations on the graph. Without using the false base line, such fluctuations will remain almost unnoticed to the reader. In such a case, the use of the false base line is legitimate and justified. However, sometimes one may use this device with the intention to magnify the fluctuations in the given data, say, pertaining to production, sales, profits, and so forth. Obviously, the use of false base line in such cases is unfair. *As such, the reader should be very vigilant while interpreting graphs based on false base line.*

Silhouette or Net Balance Graph

In such a graph the two related series are plotted in such a manner as to highlight the difference or gap between them. For example, when we plot export and import data using the same scale on a graph, the variation between the two showing the balance of trade gets highlighted. Figure 4.4, given earlier, is an

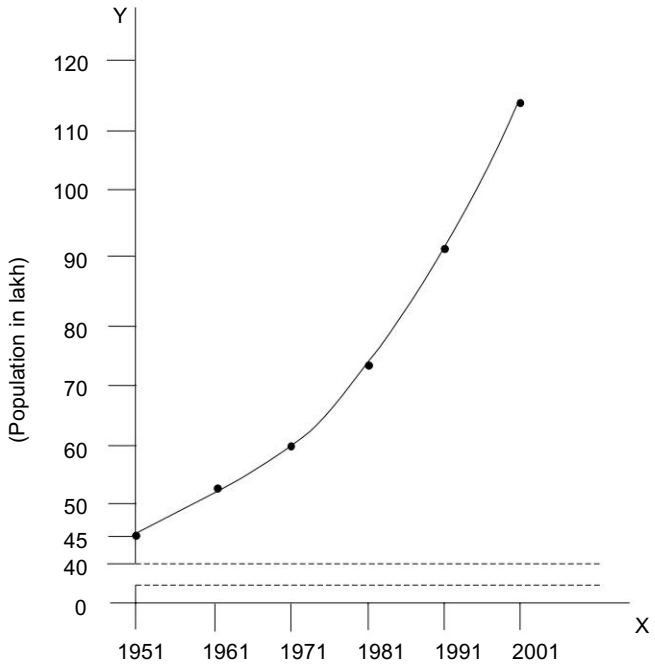


Fig. 4.5 **Use of the False Base Line**

example of this type. While it was based on a hypothetical example, we can certainly use the same type of graph showing the balance of trade by plotting real export and import data of the country. Let us show these actual data. Figure 4.6 plots these data.

Table 4.2 **Exports, Imports and Balance of Trade of India, 1993–94 to 1999–2000**

(Rs '000 crore)

Year	*Exports (including re-exports)*	*Imports*	*Balance of Trade*
1993–94	70	73	–3
1994–95	83	90	–7
1995–96	106	123	–17
1996–97	119	139	–20
1997–98	130	154	–24
1998–99	140	178	–38
1999–2000	163	205	–42

Source: Government of India: *Economic Survey* 2000–2001, p. S-81 (The figures given in Rs crore have been rounded off to Rs thousand crore.)

Component or Band Graph

Under this device, phenomena, which form part of the whole, are shown by successive bands or components to enable an overall picture along with the successive contributions of the components. The

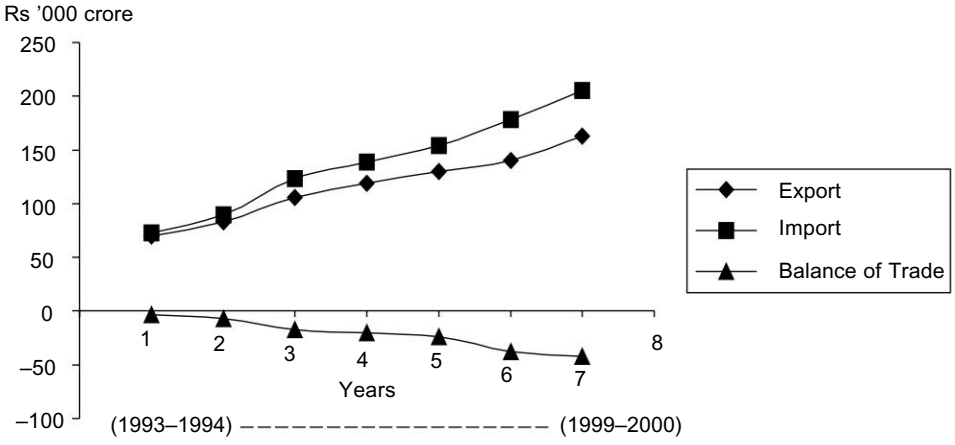


Fig. 4.6 **Exports, Imports and Balance of Trade of India (1993–94 to 1999–2000)**

data pertaining to tax revenue of the Central and State Governments and Union Territories for some recent years are given below.

Table 4.3 **Tax Revenue of Central and State Governments and Union Territories**

(Rs '000 crore)

Years	*Income & Corporate Tax*	*Customs*	*Union Excise Duties*	*Sales Tax*	*Others*	*Total*
1995–96	32	36	40	36	31	175
1996–97	37	43	45	42	33	200
1997–98	37	40	48	46	42	213
1998–99	47	48	56	56	48	255

Source: Government of India: *Economic Survey* 2000–2001, p. S-81 (The figures given in Rs crore have been rounded off to Rs thousand crore.)

These data can be plotted one after the other. However, the data need to be made cumulative. For example, the figure for 1995–96 for income and corporate tax is Rs 32 crore. While plotting the figures for customs for 1995–96, we have to first add up 32 and 36, the total being Rs 68 crore. Therefore, the point will be at Rs 68 crore on the vertical scale and not Rs 36 crore. This process of cumulation will go on till the last item, 'others'. Here the cumulative value will be the same as shown in the last column of the table, this being the total tax revenue. Figure 4.7, thus, emerges as a component or a band graph.

Range Graph

As the name implies, this graph shows the range, that is, the highest and the lowest of a certain product or item under reference. It may also show the average by taking an average of the two extreme values. For example, we are interested in the price behaviour of an equity share 'A' for a few months. In this connection, the following data are given to us.

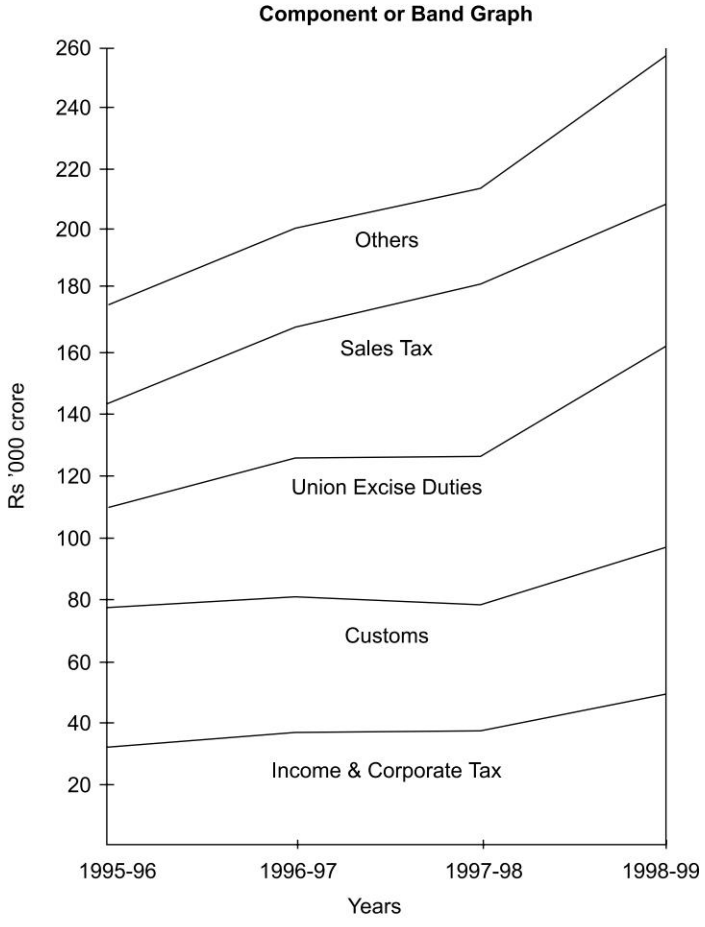


Fig. 4.7 **Tax Revenue of Central and State Governments and Union Territories**

Table 4.4 **Price Behaviour of an Equity Share 'A' (in Rupees)**

Months	*Highest Price*	*Lowest Price*	*Average Price*
January	210	176	193
February	218	180	199
March	190	160	175
April	230	150	190
May	200	160	180
June	260	190	225
July	275	205	240

The data are displayed in Fig. 4.8.

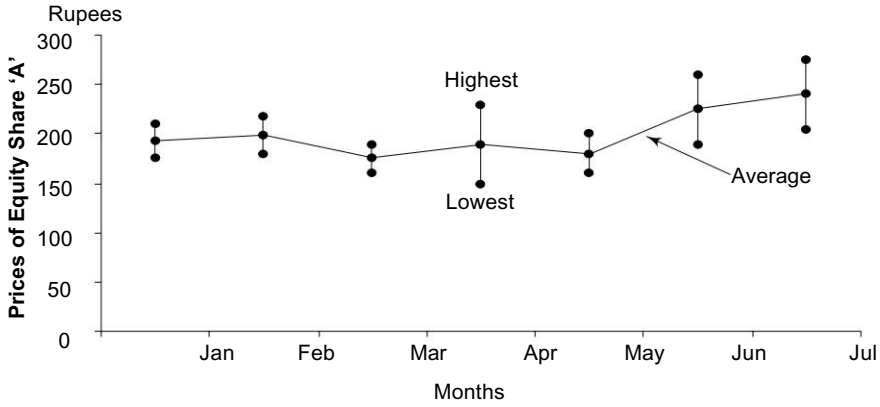


Fig. 4.8 **An Example of a Range Graph**

Frequency Graphs

We now come to another category of graphs known as frequency graphs. The following types of frequency graphs are discussed below along with a suitable illustration in each case.

(i) Histogram
(ii) Frequency Polygon
(iii) Frequency Curve
(iv) Ogive
(v) Z-Chart

Histogram In histogram, we measure the size of the item in question, given in terms of class intervals, on the axis of X while the corresponding frequencies are shown on the axis of Y.

Unlike the line graph, in histogram the frequencies are shown in the form of rectangles the base of which is the class interval. Another feature of this graph is that the rectangles are adjacent to each other without having any gap amongst them.

It may be recalled that this was not the case in the line graph where vertical frequency lines were separate and unconnected with each other. This will become clear with an example. Suppose we are given the following data to be displayed by means of a histogram.

Marks	*Number of Students*
0–20	10
20–40	22
40–60	35
60–80	28
80–100	5

It will be seen from Fig. 4.9 that the area of each rectangle is proportional to the frequency unlike the line graph shown earlier. This is because a histogram generally represents a continuous frequency distribution in contrast to line graph, which represents either a discrete frequency distribution or a time series. It should be obvious from Fig. 4.9 that the area of each rectangle is equal to the number of frequencies multiplied by the concerned class interval. Similarly, the total area of this figure comprising five rectangles of varying sizes is equal to the total number of frequencies multiplied by the corresponding class intervals.

Frequency Polygon A frequency polygon like any polygon consists of many angles. A histogram can be easily transformed into a frequency polygon by joining the mid-points of the rectangles by straight lines. The preceding figure (Fig. 4.9) has been thus transformed into a frequency polygon and shown as Fig. 4.10.

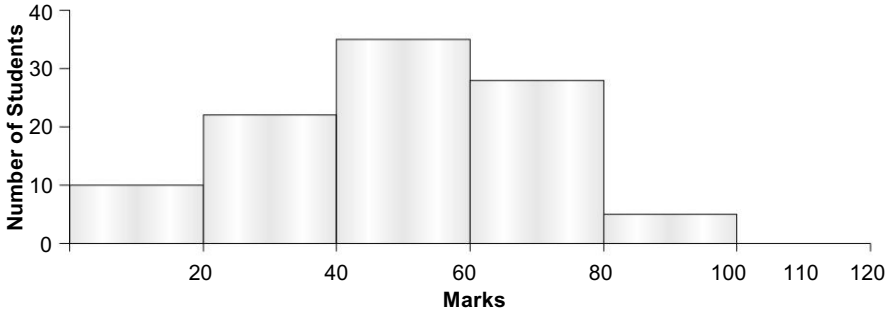


Fig. 4.9 **An Example of Histogram**

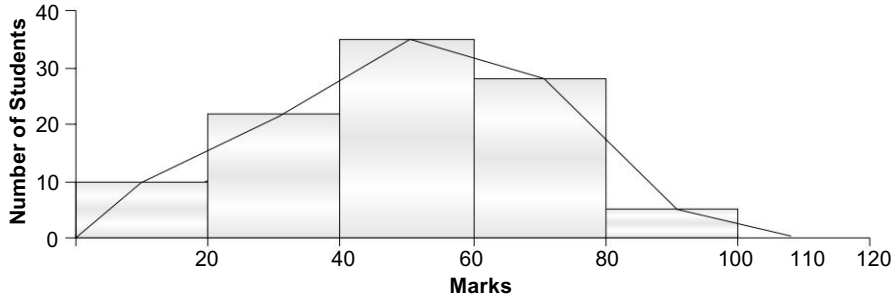


Fig. 4.10 **An Example of Frequency Polygon**

It may be noted that instead of transforming a histogram into a frequency polygon, one can draw straightaway a frequency polygon by taking the mid-point of each class-interval and by joining the mid-points by the straight lines. Another point to note is that this can be done only when we have a continuous series. In case of a discrete distribution, this is not possible.

Instead of having a frequency polygon, we can have a relative frequency polygon. While the relative frequency polygon has the same shape as the frequency polygon drawn from the same data, it has a different scale of values on the vertical axis. Instead of having absolute frequencies, the frequencies in the relative frequency polygon will be the number of frequencies in each class as a proportion of the total number of frequencies.

As the histogram and frequency polygon are very similar graphs, a question that frequently arises is: Why is it necessary to have two similar graphs? It would, therefore, be worthwhile to know their respective advantages.

The advantages of histogram are:

1. *Each rectangle shows distinctly separate class in the distribution.*
2. *The area of each rectangle in relation to all other rectangles shows the proportion of the total number of observations pertaining to that class.*

Frequency polygons, too, have certain advantages.

1. *The frequency polygon is simpler as compared to its histogram.*
2. *The frequency polygon shows more vividly an outline of the data pattern.*
3. *As the number of classes and the number of observations increase, so also the frequency polygon becomes increasingly smooth.*

Frequency Curve A frequency polygon is angular as the mid-points of class-intervals are joined by straight lines. When a frequency polygon is smoothened and rounded at the top, then it is known as a frequency curve. On account of smoothening angularities of a frequency polygon, the frequency curve becomes more suitable for the purpose of interpolation.

Figure 4.11, given below, shows the frequency curve. As is evident, this is based on the smoothening of the frequency polygon of Fig. 4.10.

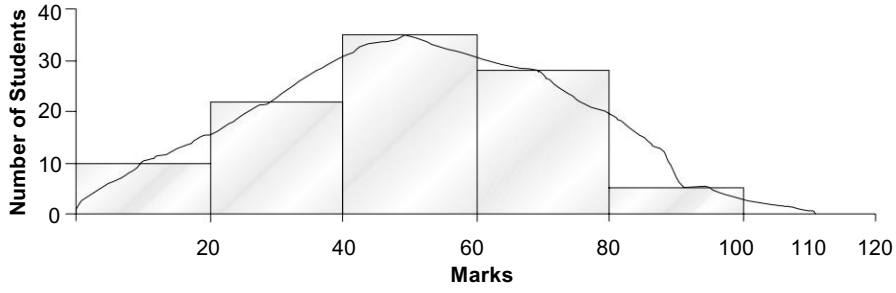


Fig. 4.11 A Frequency Curve

Cumulative Frequency Curve or Ogive So far we have discussed the graphic devices that showed frequencies as are given to us or we may say non-cumulative frequencies. We now take up another type of graph, which is based on cumulative frequencies. A cumulative frequency distribution enables us to know how many observations are above or below a certain value. A cumulative frequency curve is also known as an ogive curve.

Let us take an example to show the shape of this curve. Suppose we are given the following data:

Example 4.1

Weekly Earnings (Rs)	*Number of Employees*
Below 550	5
550–600	10
600–650	22
650–700	30
700–750	16
750–800	12
800–850	15

Solution First of all we have to transform the above series into a cumulative series. This can be done in two ways: 'less than' or 'more than' the varying amounts of income. We transform the above data as 'less than' series.

Weekly Earnings (Rs)	*Number of Employees*
Less than 550	5
Less than 600	15
Less than 650	37
Less than 700	67
Less than 750	83
Less than 800	95
Less than 850	110

We now plot these figures on the graph, which is shown in Fig. 4.12.

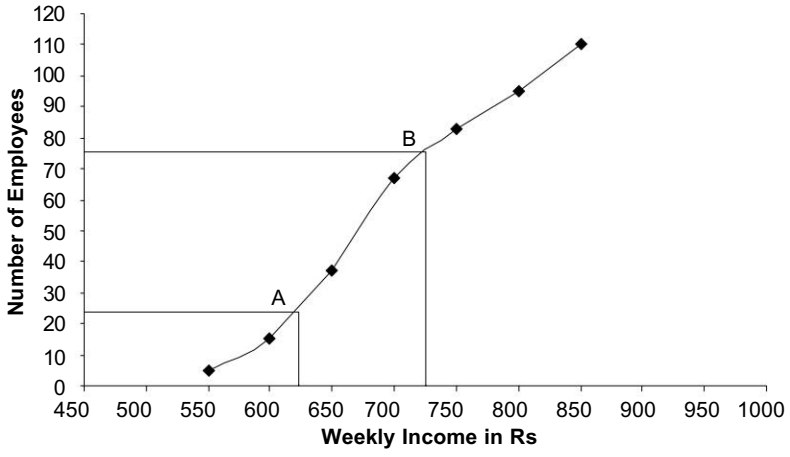


Fig. 4.12 **'Less than' Ogive of the Distribution of Weekly Income of 110 Employees**

It will be seen from Fig. 4.12 that 'less than' ogive is moving up and to the right. If we plot a 'more than' curve then it would show a declining slope and to the right, as we shall see shortly. From the ogive of Fig. 4.12, we can find the number of employees earning weekly wages, say, between Rs 625 and Rs 725. What we have to do is to draw a perpendicular from the horizontal line at Rs 625 meeting the ogive at point A. From this point, a straight line is to be drawn to meet the vertical line. This would give a certain number. Likewise, we have to do with the upper amount of Rs 725. Thus, we would get two figures of number of employees. By subtracting the smaller figure from the higher one, we can find the actual number of employees earning between Rs 625 and Rs 725. This has been shown in Fig. 4.12. The upper line which meets the vertical line shows the figure of 76 employees, while the lower line which meets the vertical line shows the figure of 24 employees. Accordingly, the number of employees whose weekly earnings are between Rs 625 and Rs 725 comes to 76 – 24 = 52.

Similarly, we can find the weekly earnings of the middle 50 percent of the employees. As we shall see later, we can ascertain graphically the values of median, quartiles, percentiles and so on with the help of a cumulative frequency graph.

Let us take another example where we use 'more than' ogive.

Example 4.2 We are given the following data and asked to draw an ogive and determine graphically the number of observations lying between 360 and 440.

Size of Item	*Number of Observations*
More than 200	400
More than 250	370
More than 300	315
More than 350	220
More than 400	115
More than 450	45
More than 500	15
More than 550	0

Solution We can find out the number of observations lying between 360 and 440 sizes. As described earlier in the case of Fig. 4.12, we have drawn the perpendiculars from the two value of 360 and 440 to meet the ogive at points A and B respectively. From these points, straight lines have been drawn to meet the vertical line at points C and D respectively. Now, C gives the value of 195 observations and D gives the value of 57 observations. Thus, the difference between C and D, that is, C – D gives us the number of observations lying between these values, which is 138 approximately.

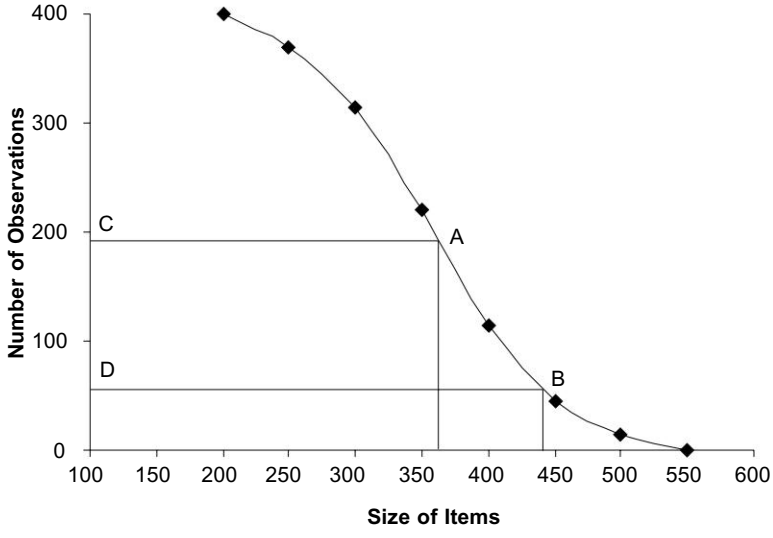


Fig. 4.13 **An Example of 'More than' Ogive**

It is worthwhile here to recall the steps needed in drawing an ogive after the data have been collected. The data are first arrayed and then a frequency distribution is obtained. Both these steps were discussed in the preceding chapter. From the frequency distribution, we obtain a cumulative frequency distribution on the basis of either 'less than' or 'more than' patterns. The cumulative data are then plotted on the graph, which gives us an ogive. Finally, the ogive curve thus obtained, can be used to find graphically the values of any given size or for any given value, the corresponding size. But, we cannot get the exact original data from any of the graphical devices.

Before we close our discussion on the ogive, we may point out that one major advantage of this curve is that it enables us to find the value of the median. This aspect will be taken up in Chapter 6.

'Z' Curve This graphic presentation is commonly used in business. The name of this device is derived because it takes the shape that resembles the letter 'Z'. In fact, it is a combination of three curves, namely, **(i)** the curve based on the original data, **(ii)** the curve based on the cumulative frequency, and **(iii)** the curve based on the moving totals. This will be clear from an example.

Example 4.3 Suppose, we are given the following data pertaining to the sales of an enterprise:

('000 Rs)

Months	*Monthly Sales*	*Cumulative Totals*	*Moving Totals*
October	20	20	120
November	30	50	110
December	30	80	130
January	40	120	140
February	25	145	160
March	35	180	180

Solution The moving total for a month has been obtained by adding the total sales of the last five months to the sales of the current month. We have to represent these data by the Z curve.

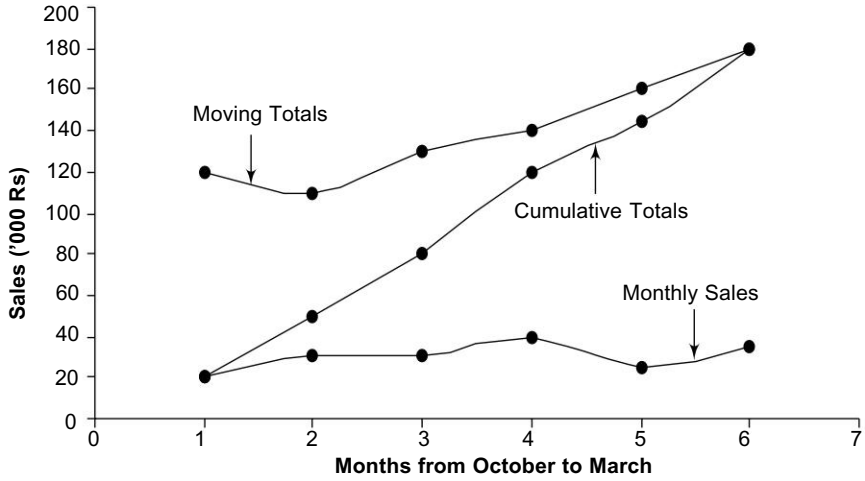


Fig. 4.14 The Z Curve

Ratio Scale Graphs

So far our discussion was confined to the natural scale graphs, where we have seen how data can be represented by varying types of graphic devices. We now come to ratio scale graphs, which are of two types viz. *semi-logarithmic* and *logarithmic*.

It may be noted that in all our previous graphic presentation of data, we have used the absolute values on both the horizontal and the vertical scales. Sometimes we are interested in the rate of change rather than the absolute magnitude of change. In such cases, the ratio scale or the logarithmic scale is used.

There are two types of logarithmic scales.

In an exclusively logarithmic scale, both the horizontal and the vertical scales show the logarithmic values instead of absolute values.

In the other case, only the vertical scale shows the logarithmic values and the horizontal scale, as usual, is on the absolute values. It is, therefore, known as the semi-logarithmic scale, which is more commonly used than the exclusive logarithmic scale.

In order to use a semi-logarithmic scale, one should have the basic understanding of logarithms. Suppose we have the following equation:

$$Y = ab^X$$

which is an exponential trend line and is similar to the compound interest formula. It may also be used to find the trend line of, for example, population, national income, or other phenomena that grow geometrically.

By taking logarithms, the above equation becomes a linear function:

$$\log Y = A + BX$$

where $\log a = A$, $\log b = B$. When applying the method of least squares, we find a and b that make $\Sigma (\log Y - \log \hat{Y})^2$ a minimum. When this is plotted on a semi-log scale, we shall obtain a straight line. The slope of the line is $\log b$, where b may be considered as the rate of increase of Y.

Using this characteristic, we may plot data that show the behaviour of a geometric sequence on a semi-log scale, fit a straight line graphically and estimate $\log b$ from the graph. From this, we can find b, which will be the rate of increase of Y.

The preceding discussion vividly shows that there are a large number of graphic devices that are used to depict the statistical data. In order to ensure that the right mode of display is adopted, the statistician should be well conversant with all these devices. It should also be clear that the pattern of data should be closely examined to finalise the most appropriate choice of graph.

GLOSSARY

Band graph	A graph having successive bends displaying an overall picture as well as contributions of components.
Bar chart	A chart in which the length of the bar represents the amount or the frequency of the item associated with the bar.
False base line	A device used to display data on the Y-axis, which otherwise cannot be shown on account of limited space.
Frequency curve	A frequency polygon smoothed by adding classes and data points to a data set.
Frequency polygon	A line graph connecting the mid-points of each class in a data set, plotted at a height corresponding to the frequency of the class.
Histogram	A form of bar chart in which the height of the bar represents the absolute or relative frequency of occurrence of the variable of interest.
Line graph	A graph displaying time period on the X-axis and the corresponding values on the Y-axis.
Ogive	A curve drawn for cumulative frequency distribution.
Pie chart	A circle divided into portions that represent the relative frequency or percentages of different categories or classes.
Polygon	A graph formed by joining the mid-points of the tops of successive bars in a histogram by straight lines.

Range graph	A graph showing the highest and the lowest values of a certain item under reference.
Ratio scale graph	A graph which uses logarithmic values instead of absolute values.
Silhoutte or net balance graph	A graph displaying two related series in such a manner as to highlight the difference between them.
Time series graph	A graph wherein time period is shown on X-axis and the corresponding values of an item on the Y-axis.
'Z' curve	A graph displaying (a) the original data, (b) cumulative frequencies, and (c) the moving totals. When these three types of data are plotted, the resultant chart resembles the letter 'Z'.

QUESTIONS

4.1 State the advantages and limitations of graphical representation of data.

4.2 Explain briefly the various methods that are used for graphical representation of frequency distribution.

4.3 What is a histogram? How do you construct it?

4.4 What do you mean by a frequency curve? State its important characteristics.

4.5 Illustrate graphically the distinction among a frequency polygon, a histogram and an ogive curve. Comment on their uses.

4.6 Graph $Y = x^3 - 2x^2 + 20x - 9$.

4.7 The following table shows the value of exports and imports from 1993–94 to 1998–99. You are required to show these data by means of a graph.

(Rs '000 crore)

Year	*Exports*	*Imports*
1993–94	70	73
1994–95	83	90
1995–96	106	123
1996–97	119	139
1997–98	130	154
1998–99	142	176

4.8 Graph the following data:

Year	*R&D Expenditure as Percent of GNP*
1980–81	0.62
1985–86	0.89
1990–91	0.85
1992–93	0.81
1994–95	0.71
1995–96	0.69
1996–97	0.66

4.9 Given below are the data on birth rate and death rate (per thousand) in India for a number of years. Graph these data.

Year	*Birth Rate*	*Death Rate*
1950–51	39.9	27.4
1960–61	41.7	22.8
1970–71	36.9	14.9
1980–81	33.9	12.5
1990–91	29.5	9.8
1997–98	26.4	8.8

4.10 The following data relate to the general index of wholesale prices (Base 1981–82 = 100) for the period 1990–91 to 1999–2000. Show these data by a suitable graph.

Year	*Wholesale Price Index*
1990–91	183
1991–92	208
1992–93	229
1993–94	248
1994–95	275
1995–96	296
1996–97	315
1997–98	330
1998–99	353
1999–2000	362

4.11 In Question 3.20, it was asked to prepare a frequency table. You have now to show the information contained in that frequency table by a histogram.

4.12 Given the following distribution, calculate and plot 'or more' and 'less than' frequency curves.

Class	*Frequency*
40–49	4
50–59	8
60–69	4
70–79	24
80–89	6
90–99	4

4.13 Plot the frequency polygon from the data given in the preceding question.

4.14 Using the distribution given in Question 4.12, calculate and plot the percentage frequency distribution.

4.15 The following table gives the distribution of weekly wages of 600 workers of a factory:

Weekly Wages (in Rs)	*Frequency*
Below 375	69
375–450	167
450–525	207
525–600	65
600–675	58
675–750	24
750–825	10

Draw an ogive for the above data and find out graphically the limits of weekly wages of the central 50% of the workers.

4.16 Draw an ogive for the following distribution and find out graphically the number of workers who earned monthly wages between Rs 980 and Rs 1080.

Monthly Wages (in Rs)	*Number of Workers*
900–950	6
950–1000	10
1000–1050	22
1050–1100	30
1100–1150	16
1150–1200	12
1200–1250	15

4.17 Draw an ogive for the following frequency distribution:

Marks	*Number of Students*
0–10	8
10–20	10
20–30	12
30–40	25
40–50	40
50–60	35
60–70	30
70–80	20
80–90	15
90–100	5

(i) Find graphically the number of students who obtained at least 75% marks.
(ii) Find the number of students scoring between 55% and 75% marks.
(iii) If a student has to obtain at least 35 marks to pass the examination, find out how many students failed in the examination.

4.18 Draw a histogram, a frequency polygon, and an ogive curve for the following distribution:

Marks less than	*Number of Students*
10	4
20	6
30	24
40	46
50	67
60	86
70	96
80	99
90	100

4.19 The following data relate to the foreign tourist arrivals and the foreign exchange earnings. Represent the data graphically.

Year	*Number of Foreign Tourists (in Lakhs)*	*Foreign Exchange (US $ in Million)*
1997–98	24	2914
1998–99	24	2993
1999–2000	25	3036
2000–01	27	3168
2001–02	24	2910
2002–03	25	3029
2003–04	29	3979
2004–05	36	5029

The McGraw-Hill Companies

CHAPTER 5

DIAGRAMMATIC PRESENTATION OF DATA

Learning Objectives

After reading this chapter you would

- understand different types of diagrams and draw them without any difficulty
- understand the relative merits and limitations of each type of diagram
- make the most suitable choice of a diagram keeping in mind the given data set
- critically examine and interpret any diagrammatic presentation attempted by others.

Chapter Prerequisites

Before starting work on this chapter, you should ensure that you are conversant with

1. the calculation of percentages
2. the plotting of data on graphs

5.1 INTRODUCTION*

Having discussed the graphic presentation of data, we now turn to the diagrammatic presentation of quantitative information. There are broadly four types of diagrams:

1. One-dimensional diagrams such as bar diagrams
2. Two-dimensional diagrams such as rectangles and squares
3. Three-dimensional diagrams such as cubes, cylinders and spheres
4. Pictograms and cartograms such as maps

We shall discuss all these types of diagrams, giving suitable illustrations, in this chapter. Let us first begin with the one-dimensional diagrams.

5.2 ONE-DIMENSIONAL DIAGRAMS

Line Diagram

The first one under this head is the *line diagram*. An example of line diagram is given in Fig. 5.1.

* The discussion given in Sections 4.2 to 4.5 in Chapter 4 is also relevant so far as diagrammatic presentation is concerned.

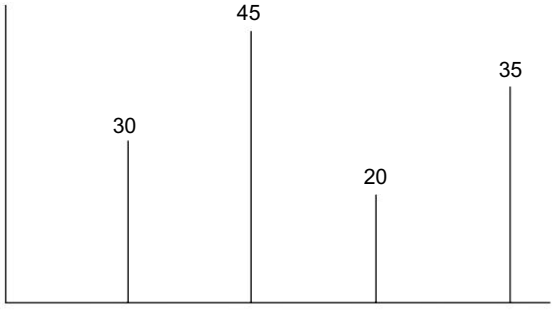


Fig. 5.1 **Line Diagram**

This is the simplest form of diagram. The height of each line indicates the value of an item that is being measured. The line diagram is drawn taking a suitable scale.

Simple Bar Diagrams

Unlike the line diagram, a simple bar diagram shows a width or column. It is used to represent only one variable. Suppose we have the production of an item X for three years. This information can be shown as in Fig. 5.2.

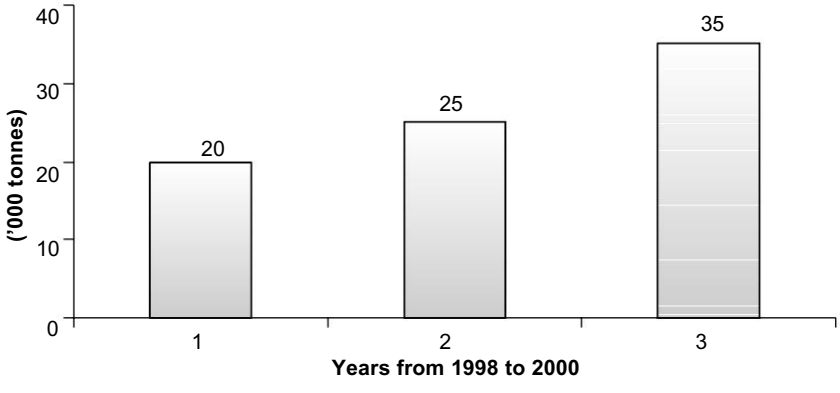


Fig. 5.2 **Production of Item X ('000 tonnes)**

It will be seen that each bar has an equal width but unequal length. The length indicates the magnitude of production. From such a diagram, it becomes obvious that there is an increase in trend in the production of X commodity. In view of its simplicity and ease of drawing it, a bar diagram is very popular in practice. All the same, it suffers from a major limitation. Such a diagram can display only one classification or one category of data.

It may be noted that the simple bars shown in Fig. 5.2 are drawn vertically. They are, therefore, known as vertical bars. But the same bars can be drawn horizontally as shown in Fig. 5.3.

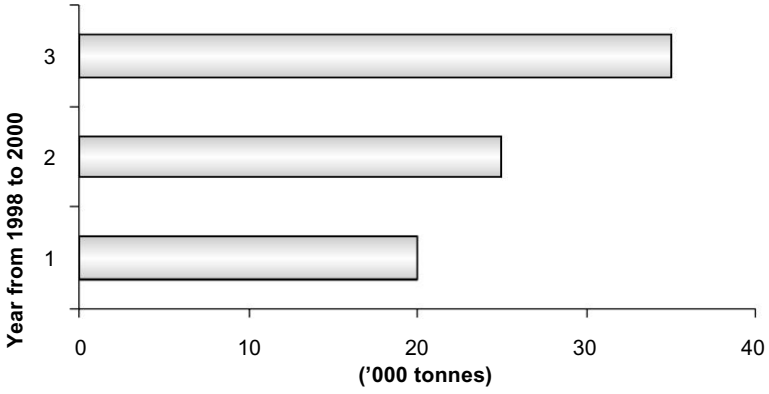


Fig. 5.3 **Horizontal Bars**

Multiple Bars

When two or more interrelated series of data are depicted by a bar diagram, then such a diagram is known as a multiple-bar diagram. Suppose we have export and import figures for a few years. We can display by two bars close to each other, one representing exports while the other representing imports. Figure 5.4 shows such a diagram based on hypothetical data.

It should be noted that multiple bar diagrams are particularly suitable where some comparison is involved.

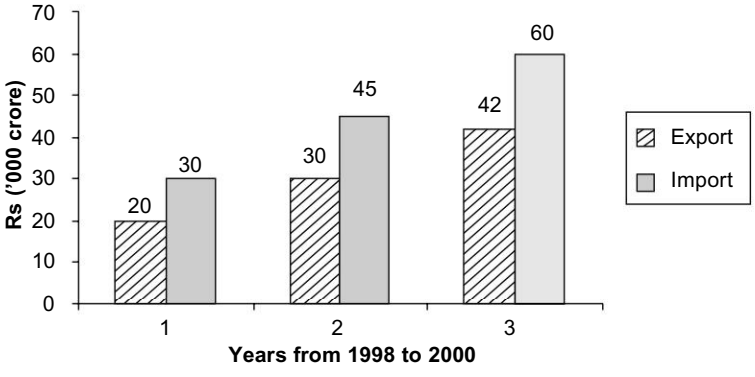


Fig. 5.4 **Multiple Bars**

For example, we may represent admission data pertaining to, say, an MBA programme for boys and girls or performance of boys and girls in a particular examination. Similarly, the number of skilled, semi-skilled and unskilled workers in a factory for a couple of years can be represented well by means of a multiple-bar diagram.

Subdivided or Component Bar Diagram

As the name implies, this diagram shows subdivisions of components in a single bar. For example, a bar diagram may show the composition of revenue expenditure of the Government of India. The components of this bar could be defence expenditure, interest payments, major subsidies, grants to states and union territories and others. Such bar diagrams are shown in Fig. 5.5 for two years.

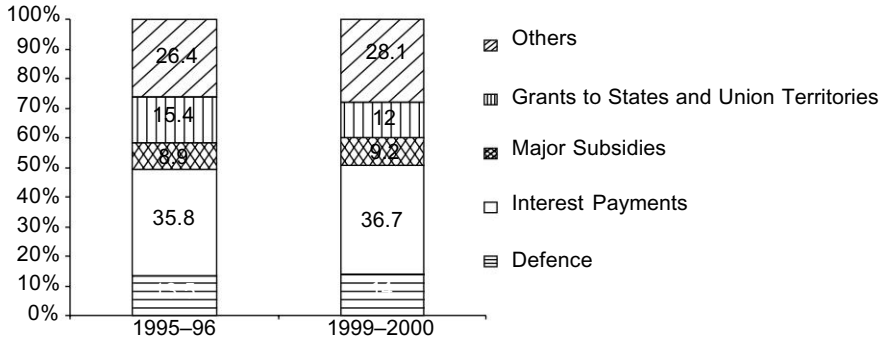


Fig. 5.5 **Composition of Revenue Expenditure (percent)**

It will be seen from these bars that the order of components in each bar is the same. This is necessary for proper comparison of two or more bars. It is also to be noted that Fig. 5.5 shows the components in percentages. These could have been shown in absolute figures as well, but in such a case the length of each bar would have been different in accordance with the total revenue expenditure of the Government in the two years. Since the bars are in terms of percentages, the length of each bar is the same indicating the total as 100 percent.

In order to distinguish the different components, different colours or shades should be used. This should be accompanied by a legend or key for identifying these components.

Another point worth noting is that such a diagram should not be used when there are too many components or subdivisions. This is because of two reasons. *First*, it may be difficult to provide so many subdivisions particularly when their magnitudes are small. *Second*, it may be difficult to understand and interpret such a diagram. Component bars can be very suitable mode of presenting data on the enrolment of students by the type of courses such as MA, M.Com., M.Sc., MBA, and so on. Likewise, data on the employment of the workers in the public sector, private sector and the total can be suitably displayed by component bars.

Percentage Subdivided Bar Diagrams

These bar diagrams are useful when the relative changes in data are to be displayed. Our previous example is a case in point. We may give another example.

Suppose we are given the following data pertaining to the monthly expenditure of two families under some heads.

Item of Expenditure	*Expenditure in Rupees*	
	Family A	*Family B*
Food	1000	1600
Clothing	500	800
Rent	800	1000
Education	400	800
Recreation	200	500
Miscellaneous	100	300

Since these data are in absolute terms, we have to convert them into percentages. This has been done in the following table.

Calculation of Percentage Expenditure under Different Heads

Item of Expenditure	*Family A*			*Family B*		
	Rs	*%*	*C %*	*Rs*	*%*	*C %*
Food	1000	33.3	33.3	1600	32.0	32.0
Clothing	500	16.7	50.0	800	16.0	48.0
Rent	800	26.7	76.7	1000	20.0	68.0
Education	400	13.3	90.0	800	16.0	84.0
Recreation	200	6.7	96.7	500	10.0	94.0
Miscellaneous	100	3.3	100.0	300	6.0	100.0
Total	**3000**		**Total**	**5000**		

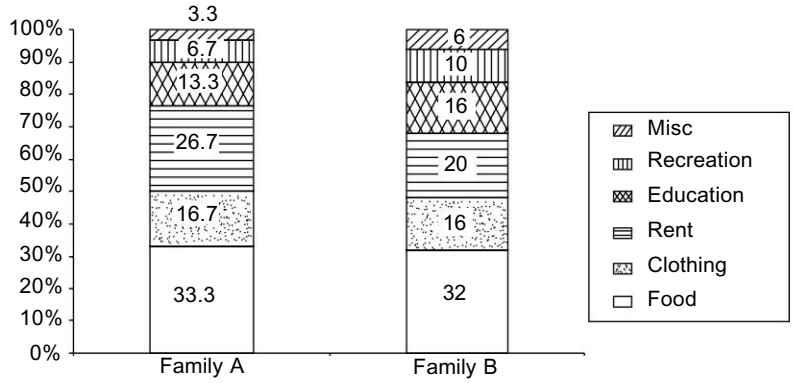


Fig. 5.6 **Percentage Subdivided Bars**

We may now show the percentage expenditure of the two families by percentage bars. It may be noted that we have calculated cumulative percentages, which have been plotted (Fig. 5.6).

Although the above diagram gives some idea of relative expenditure on various items of the two families, a rectangular diagram would be a much better means of displaying the expenditure, which will be taken up later.

Broken Bars

In some cases we find that the data may have very wide variations in values in the sense that some values may be very large while others extremely small. In such cases, larger bars may be broken to provide space for the smaller bars. An illustration will make this clear.

Suppose we are given data pertaining to the number of students enrolled in certain faculties in a university. The data regarding the enrolment are given below:

Faculty	*Number of Students*
Arts	1500
Science	2000
Commerce	6800
Management	500

These data are now shown in Fig. 5.7.

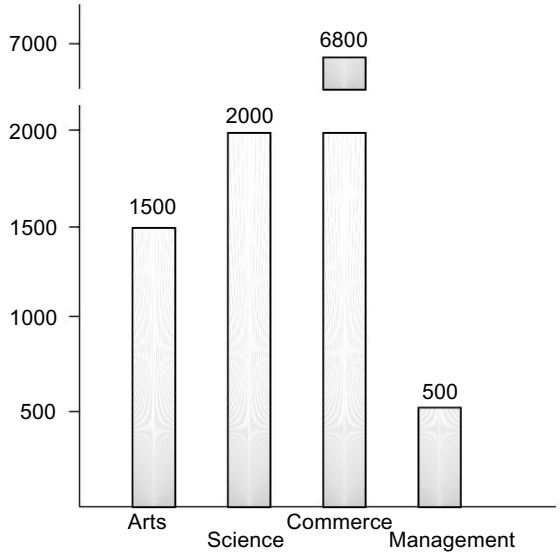


Fig. 5.7 **An Example of Broken Bar**

Deviation Bar Diagrams

Deviation bars are used to show both positive and negative values. For instance, we may show net profit data for a company for some years. It is possible that in one or two years, instead of earning net profit the company might have sustained net loss. In such a case, the data on net profit will be displayed above the base line while the data on net loss below it. The bars can show the absolute data or the percentages as may be thought proper. Let us take an example.

Suppose we have the following data relating to rates of change (percent) in agricultural production (all crops) for a few years. These data are shown in Fig. 5.8.

Year	*Percent Change Over Previous Year*
1996–97	9.3
1997–98	–6.1
1998–99	7.4
1999–2000	–2.2

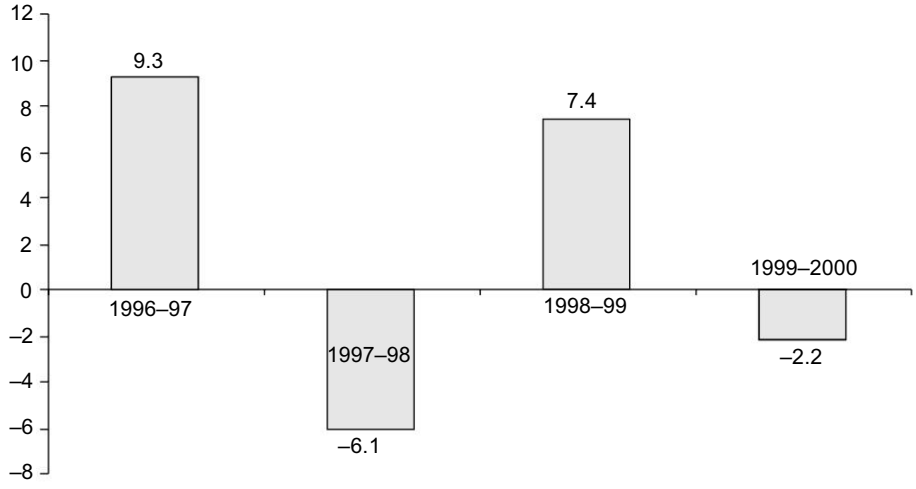


Fig. 5.8 **Agriculture Production—Rate of Change (percent)**

Duo-directional Bar Diagram

A duo-directional bar diagram as its name indicates is a diagram on both sides of the axis of X. One component is shown above the horizontal line while the other is shown below it. The two components taken together give the total value of the item displayed. This will be clear from the example and Fig. 5.9 on the next page.

Suppose we have the following data for two years—1998–99 and 1999–2000.

Items	*1998–99 ('000 Rs)*	*1999–2000 ('000 Rs)*
Total earnings from production	100	80
Cost of production (all expenses)	60	55
Net Income	40	25

Sliding Bar Diagram

A sliding bar diagram is similar to the duo-directional bar diagram. The only difference between the two is that the latter is based on the absolute values, as we have seen above, while the former is based on percentages. Sliding bar diagrams can be shown in either way—*horizontally* or *vertically*. Let us take an example.

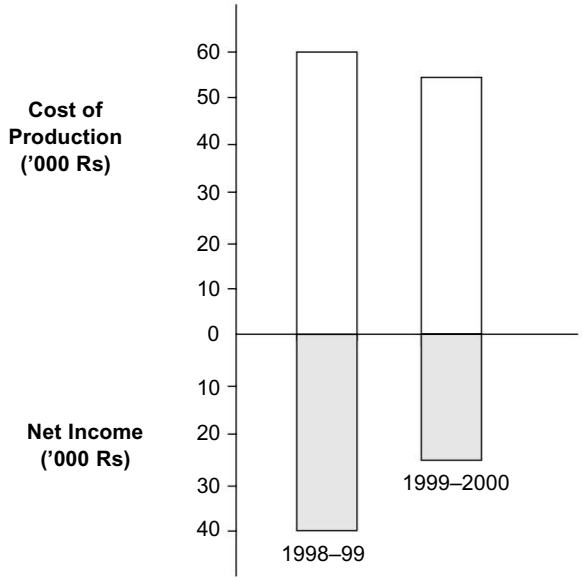


Fig. 5.9 **Duo-directional Bar Diagram**

Suppose we are given the following data to be presented by a sliding bar diagram.

Results of the B.Com. Examination of Three Colleges Affiliated to a Certain University

College	*Pass*	*Fail*
A	60	40
B	75	25
C	80	20

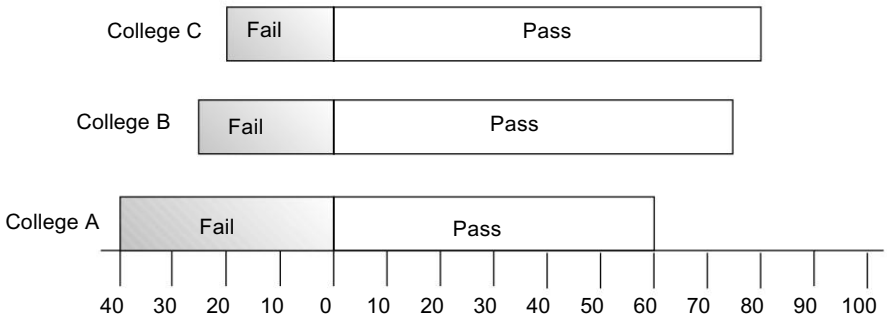


Fig. 5.10 **Sliding Bar Diagram**

Pyramid Diagram

A pyramid diagram shows a number of horizontal bars, which are arranged in such a manner as to give an appearance of a pyramid. Such diagrams are suitable to present data on population, occupation, education, and so forth.

Suppose we have the following data pertaining to the inhabitants of a locality.

(Figures in '000)

Age Group	*Male*	*Female*	*Total*
0–20	25	23	48
20–40	20	18	38
40–60	15	15	30
60–80	10	8	18
80+	5	3	8

Figure 5.11 displays these data in the form of a pyramid diagram.

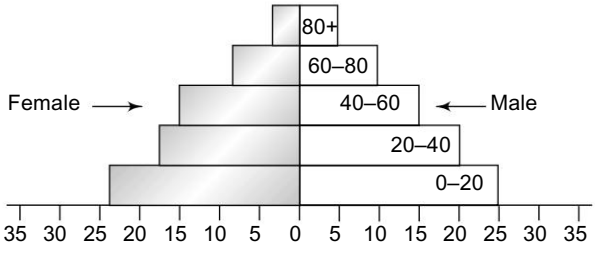


Fig. 5.11 **Pyramid Diagram**

5.3 TWO-DIMENSIONAL DIAGRAMS

In case of one-dimensional diagrams, only one side, namely, length of the bar is relevant to measure the magnitude of a phenomenon as we have seen earlier. In contrast to one-dimensional diagrams, the two-dimensional diagrams consider both length and breadth, that is, the area in rectangles and squares and π, that is, radius in circles. Thus, the three forms of such diagrams are rectangles, squares and circles, which are discussed below.

Rectangular Diagrams

Rectangular diagrams are used to compare values of different items in two or more situations. The area of rectangular diagram is in proportion to the absolute value of a product or item. This would become clear by the following example.

A travel agency has a fleet of taxis and luxury buses. It has 15 taxis and 6 luxury buses. During the past year, it has earned the revenue of Rs 90 lakh from taxis and Rs 120 lakh from luxury buses. We have to show these data by rectangular diagrams. Figure 5.12 shows the two rectangles displaying this information.

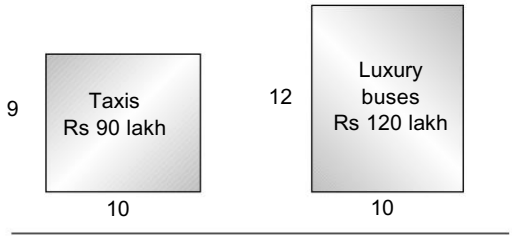


Fig. 5.12 Rectangular Diagram

It may be noted that in both the cases, the width has been taken equal so that the diagrams can be easily understood and interpreted. Here we have shown the revenue from the two sources, namely, 15 taxis and 6 luxury coaches. We can, in fact, show several items in the same diagram, which implies subdivisions of rectangles. We take another example to explain this aspect.

Example 5.1 The data given below relate to firms A and B for a particular month. These are to be shown by rectangular diagrams so that cost and profit per unit in the two firms can be compared.

('000 Rs)

Items	*Firm A*	*Firm B*
Raw material cost	10,000	7,000
Labour cost	7,000	3,000
Other overhead expenses	4,000	1,500
Miscellaneous expenses	3,000	500
Total cost	24,000	12,000
Total revenue	30,000	18,300
Profit	6,000	6,300
No. of units produced and sold	1,200	900

Solution Here, first we have to calculate per unit cost of different items as well as revenue per unit for both the firms. This is shown as follows.

Cost Per Unit

('000 Rs)

Items	*Firm A*	*Firm B*
Raw material cost	8.34	7.78
Labour cost	5.83	3.33
Other overhead expenses	3.33	1.67
Miscellaneous expenses	2.50	0.55
Total cost	20.00	13.33
Revenue	25.00	20.33
Profit	5.00	7.00

Total length of rectangle for firm A = Total revenue /Output
= Rs 30,000/1,200 = 25 units

Length of the corresponding total cost = Rs 24,000/1,200 = 20 units

Therefore, length for per unit profit = 25 – 20 = 5 units

Similarly, total length of rectangle for firm B = Total revenue/Output
= Rs 18,300 / 900 = 20.33 units

Length of the corresponding total cost = Rs 12,000 / 900 = 13.33 units

Therefore, length for per unit profit = 20.33 – 13.33 = 7 units

On the basis of the above calculations, the length of the rectangle for firm A would be 25 and for firm B 20.33, while width would be their respective output in units. Further, we shall have subdivisions of the rectangle in accordance with the cost per unit for the different items. Figure 5.13 displays this information.

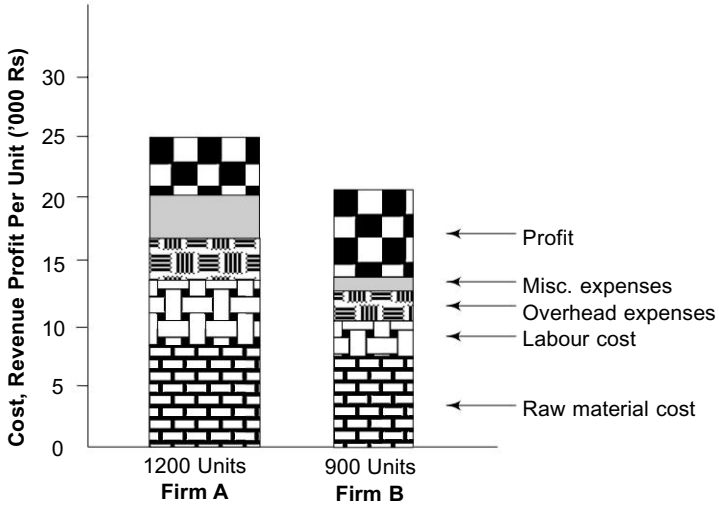


Fig. 5.13 **Rectangular Diagrams for Firms A and B**

Percentage Rectangular Diagram

The foregoing rectangular diagrams were based on the absolute values of different items of expenditure, total cost, total revenue and profit. The same data on these items can be presented in the form of percentage rectangular diagrams. In such a case, the area pertaining to each item represents its percentage to the total, which is taken as 100. For this purpose, we have to first convert absolute values in terms of percentages and then use them in the rectangular diagram. These percentages for different items for the two firms A and B are shown on the next page.

Percentages to the total

Items	*Firm* A		*Firm* B	
	Percentage	*C. Percentage*	*Percentage*	*C. Percentage*
Raw material cost	33.3	33.3	38.3	38.3
Labour cost	23.3	56.6	16.4	54.7
Other overhead expenses	13.4	70.0	8.2	62.9
Miscellaneous expenses	10.0	80.0	2.7	65.6
Total cost		80.0		65.6
Profit	20.0	100.0	34.4	100.0
Total revenue	100.0		100.0	

Figure 5.14 displays these data. It may be noted that the width of the two rectangles is in the ratio of the output of the two firms viz. 1200 units and 900 units. The length of the rectangle is, of course, the same, being 100 percentage in each case.

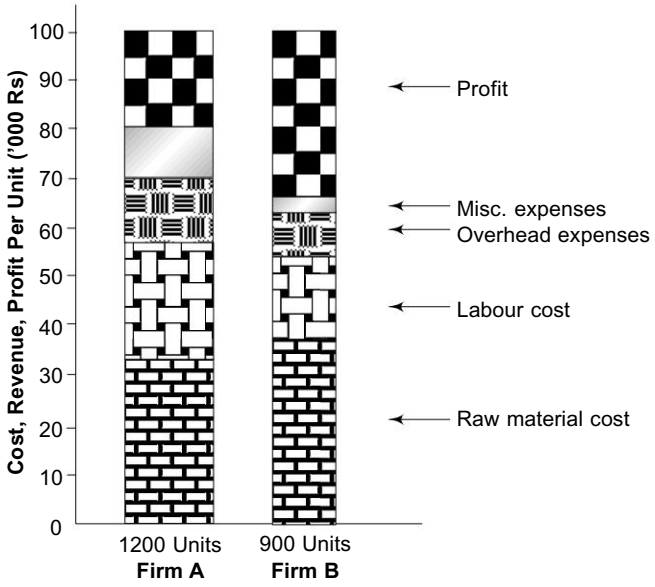


Fig. 5.14 **Percentage Rectangular Diagrams**

Square Diagrams

In a square diagram the length and width are of the same dimension. Sometimes a square diagram is preferable to rectangle. This is because when a comparison is involved between any two phenomena but their magnitudes are widely different from each other, it is difficult to draw rectangles on the same page. One rectangle would have very wide width while the other a very small width. This apart, a meaningful comparison between the two diagrams would be quite difficult. To overcome this difficulty, square diagrams are preferred. Let us take an example.

Example 5.2 We are given the following data:

Year	*Production of Food Grains (million tonnes)*
1950–51	50.8
1960–61	82.0
1970–71	108.4
1980–81	129.6
1990–91	176.4

Source: GOI, *Economic Survey*, 1999–2000.

Solution As mentioned earlier, first we have to calculate the square roots of these figures. The square roots are as follows:

Year	*Production of Food Grains (Million Tonnes)*	*Square Roots (Million Tonnes)*
1950–51	50.8	7.12
1960–61	82.0	9.06
1970–71	108.4	10.41
1980–81	129.6	11.38
1990–91	176.4	13.28

The diagrams are shown below.

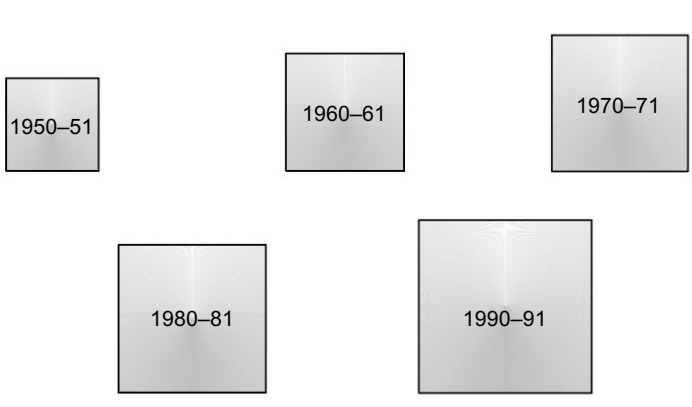


Fig. 5.15 Square Diagrams

5.4 CIRCULAR OR PIE DIAGRAMS

Another type of diagram, which is more commonly used than the square diagrams, is the circular or pie diagram. In fact, circles can conveniently display the data on production of foodgrains presented in the preceding square diagrams.

Let us take an example to show how a circle or pie diagram can be drawn. In this connection, the following data are given to us.

Items	*Expenditure as Percent of Total*
Food	50
Clothing	15
Housing	10
Fuel and lighting	5
Education	10
Recreation	5
Miscellaneous	5
Total	**100**

We have first to calculate the degrees for each of the above-mentioned items. These calculations are shown below:

Food = $50/100 \times 360 = 180°$

Clothing = $15/100 \times 360 = 54°$

Housing = $10/100 \times 360 = 36°$

Fuel and lighting = $5/100 \times 360 = 18°$

Education = $10/100 \times 360 = 36°$

Recreation = $5/100 \times 360 = 18°$

Miscellaneous = $5/100 \times 360 = 18°$

The total of all these angles will be 360°.

The pie diagram is also known as an angular sector diagram though in common usage the term pie diagram is used. It is advisable to adopt some logical arrangement, pattern or sequence while laying out the sectors of a pie chart. Usually, the largest sector is given at the top and others in a clockwise sequence. The pie chart should also provide identification for each sector with some kind of explanatory or descriptive label. If the space within the chart is sufficient, the lables can be placed inside the sectors, otherwise these should be shown outside the circle, using an arrow or pointing out to the concerned sector.

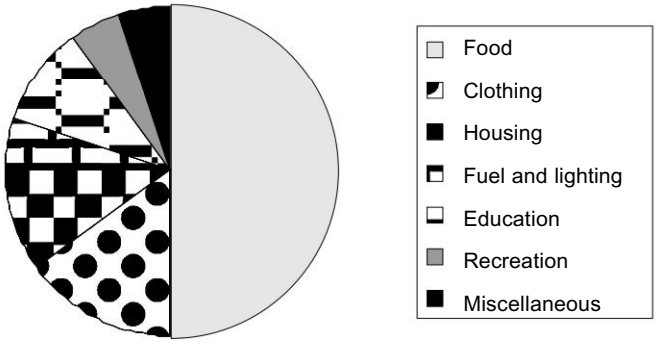


Fig. 5.16 **An Example of Pie Diagram**

Limitations of Pie Diagrams There are certain *limitations* of pie diagrams.

They are not as effective as bar diagrams for accurate reading and interpretation. This limitation becomes all the more obvious when the series are divided into a large number of components or the differences among the components are too small. When a series comprises more than five or six categories, pie chart would not be a proper choice since it would be confusing to differentiate the relative values of several small sectors having more or less the same size. Although pie diagram is frequently used, it turns out to be inferior to a bar diagram whether it is simple bar or a divided bar or a multiple bar.

5.5 THREE-DIMENSIONAL DIAGRAMS

Three-dimensional diagrams are the volume diagrams. The common forms are *cubes*, *cylinders* and *spheres*. In the case of cubes, all the three dimensions, length, width and height are taken into consideration. In the case of a cylinder, the length and the diameter of the circle are taken into consideration. A sphere in the shape of a ball can be used in a three-dimensional form.

Advantage of Three-dimensional Diagrams The question one may ask here is: What is the main advantage of these diagrams over the other types that we have just studied? When the data to be displayed have very wide magnitudes, then such diagrams are preferred. For example, if we have two magnitudes, say, 125 units and 1331 units of output of two firms, it would be difficult to use a diagram other than cubes, as there is an extremely wide difference between the two values. But, if we take the cube root of these values, then it becomes possible to display the magnitudes. The cube root of 125 is 5 and of 1331 is 11. Based on these figures we can draw the diagram. This is shown in Fig. 5.17.

It may be noted that these two cubes are of different sizes. In the case of the smaller one we have taken 1 inch as its all sides. In the case of larger cube, we have taken 2.2 inch as all sides. This is because we have divided 11 by 5 which gives 2.2 as the measurement for the second cube.

Though in the beginning one may think it is difficult to construct a cube, but it is not so.

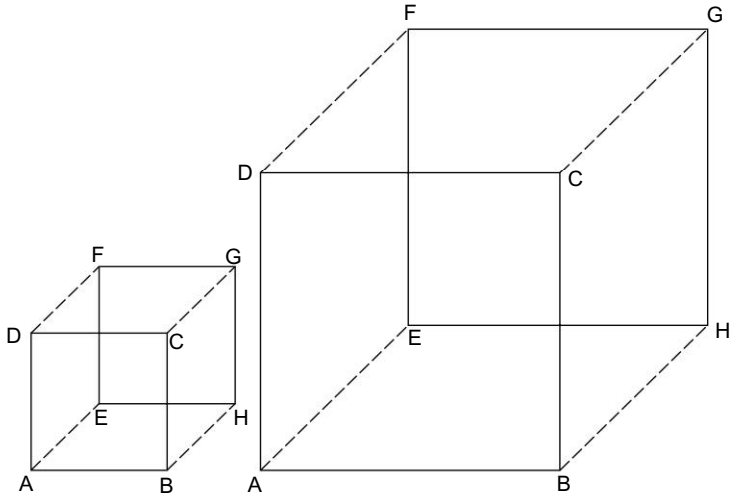


Fig. 5.17 **An Example of Cube Diagrams**

Procedure to Construct a Cube

1. Construct a square with sides of 1 inch. The square is represented by ABCD (smaller diagram).
2. Locate the mid-point of the line DC and from this mid-point draw a perpendicular of 1 inch in such a way that 0.5 inch should be on either side of this line. This perpendicular is represented by EF.
3. Join DF and from C draw a line CG which should be parallel as well as equal to DF.
4. Join FG. From the point G draw a line GH, which should be parallel as well as equal to FE.
5. Join B and H as also E and H, thus giving the required cube, ADFGHB.

The same procedure is to be followed for the second cube.

Instead of going through the above procedure, we now have the computer facility. We may now construct the required size cubes using the computer, which has been attempted here as well (Fig. 5.17).

5.6 PICTOGRAMS

Pictograms are the pictures that are frequently used in presenting data. As they are attractive and easy to understand, they are an appropriate mode for presentation of data particularly for the layman. In pictograms, a pictorial symbol is used to clearly indicate the item that is being displayed. For example, if we are dealing with the production of cars, then we may use the symbol of car. If the data relate to the enrolment of students, then we may use a symbol of a student to present those data. It may be noted that pictograms are not abstract presentations such as bar diagrams.

Figure 5.18 shows some hypothetical data pertaining to cars sold by a manufacturer. A symbol of one car represents 1000 cars sold by the manufacturer. From the figure it becomes obvious that in 1998 only 3000 cars were sold, which rose to 6000 cars in the following year and finally to 8000 cars in 2000.

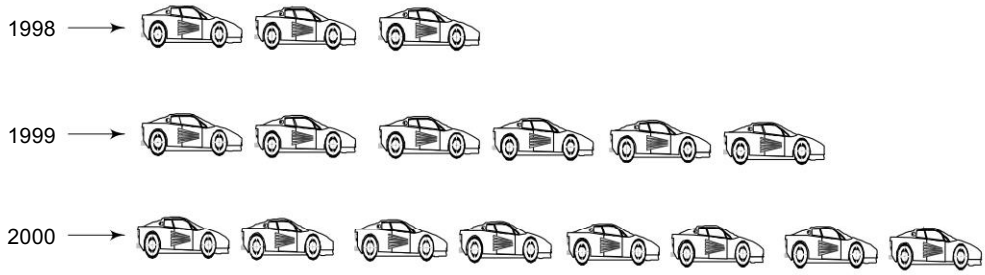


Fig. 5.18 **An Example of Pictogram**

Major Advantages of Pictograms *First*, they are far more attractive as compared to other diagrams. As such, they generate interest in the audience. *Second*, it has been observed that the facts presented by pictograms are remembered for a longer time than tables, bars and other diagrams.

Limitations of Pictograms *First*, they are difficult to draw. *Second*, sometimes we cannot show the actual data properly. For example, if we have to show the number of enrolment of students, in view of limited space, we may say that the picture of one student shows that 100 students have been enrolled. Now, if the actual enrolment is, say, 279, then how to represent it as we cannot divide or split

the symbol of one student. In case we round up this figure to 300, then it is clear that a higher figure is being shown, resulting in an inaccuracy.

5.7 CARTOGRAMS

Cartograms are the maps that are used to present statistical data on a geographical basis. Suppose we have to show the rate of literacy in different parts of India. What we can do is to group different literacy rates into three or four categories. Then we display these rates on India's map, using different-coloured lines to identify different literacy rates. Similarly, we can show the number of registered factories in different States of India for a particular year. The actual number can be put within the geographical boundaries of the States as shown in the map. Alternatively, we can show the States into three distinct categories—highly industrialised, moderately industrialised and poorly industrialised and use different colours for these categories to bring them into sharp focus.

Cartograms too are very attractive but they should be used especially where geographic comparisons are to be made and where approximate measures can serve the purpose. This is understandable as maps are unable to provide 100 percent accuracy.

5.8 CHOICE OF A SUITABLE DIAGRAM

The foregoing discussion shows that a variety of diagrams are available, which can be used to display statistical data. At times it may be difficult for one to select a particular type of diagram from amongst several types. However, there are some major considerations which should be kept in mind while deciding on the mode of data presentation. These have been given in (Section 4.5) of the preceding chapter.

As all these considerations are very relevant, one should consider the pros and cons of different modes of display in each case before making a final choice. Above all, one should be aware of a major limitation of diagrams. While they can be useful up to a certain point, they cannot take the place of tabular form of data presentation.

GLOSSARY

Bar diagram	A simple diagram, whose heights represent the frequencies of respective categories.
Cartograms	Maps that are used to present statistical data on a geographical basis.
Component bar	A bar diagram that shows subdivisions or components of a certain item.
Deviation bar	A bar diagram that shows both positive and negative values. Positive values are shown above the zero-base line, while negative values below it.
Duo-directional bar	A bar diagram that shows data on both sides of axis of X in such a way that the two bars, taken together, give the total value of the item displayed.

Horizontal bar	A diagram where a bar is shown horizontally instead of vertically, which is the usual practice.
Multiple bar	Two or more interrelated series of data are shown by a set of bars.
Pictogram	A diagram in the form of picture for displaying data.
Pie chart	A circular diagram showing different items or components by different angles.
Pyramid diagram	A diagram displaying a number of bars, arranged horizontally in such a manner as to give an appearance of a pyramid.
Rectangular diagram	A diagram that displays data with the help of one or more rectangles.
Sliding bar	A bar diagram that shows percentage data on both sides of axis of X.
Vertical bar	A usual form of bar diagram that depicts the value of a certain item on the vertical axis.

QUESTIONS

5.1 "Diagrams help us visualise the whole meaning of a numerical complex at a single glance." Comment.

5.2 State the advantages of a diagrammatic presentation of statistical data. Describe briefly, with suitable examples, a single bar diagram, a multiple bar diagram and a component bar diagram.

5.3 What are the different types of diagrams used in Statistics? Write a short note on the utility of diagrams in business.

5.4 What are the advantages and limitations of diagrammatic presentation of data?

5.5 What factors would you take into consideration while deciding the type of diagram to be used for a given data set?

5.6 Name the different types of diagrams commonly used and mention the situations where the use of each type of diagram would be appropriate.

5.7 "Charts are more effective in attracting attention than are any of the other methods of presenting data." Do you agree? Give reasons for your answer.

5.8 What are the different types of bar diagrams? Discuss their relative merits and demerits.

5.9 The following table shows the marital status of males and females (18 years and older) in a metropolitan city. Draw a pie chart separately for males and females to display these data.

Marital Status	*Male (Percent of Total)*	*Female (Percent of Total)*
Single	21	16
Married	65	73
Widowed	9	4
Divorced	5	7

5.10 The following data relate to monthly income and expenditure under different heads for two families. Show these data by a suitable diagram.

Item of Expenditure	*Family A (Monthly Income Rs 12,000)*	*Family B (Monthly Income Rs 16,000)*
Food	4,000	4,800
Clothing	2,500	3,000
Rent	3,000	4,000
Education	1,500	2,500
Fuel and lighting	400	800
Others	600	900

5.11 Draw a pie chart for comparing the various costs (Rs in lakh) of the house building activities in two periods, 1984 and 1989.

Items	*1984*	*1989*
Land cost	0.75	1.00
Material cost	1.00	1.75
Labour cost	0.60	1.00
Fixtures and furnitures	0.40	0.75
Miscellaneous	0.25	0.50

5.12 The following data pertain to agricultural production and industrial production, showing percentage change over previous year.

Year	*Agricultural Production*	*Industrial Production*
1996–97	9.3	5.6
1997–98	–6.1	6.6
1998–99	7.4	4.0
1999–2000	–2.2	6.2

Show these data by bar diagrams separately for agricultural and industrial production.

5.13 The data given below pertain to R&D expenditure by public, private and industrial sectors for three years. You are required to draw a suitable diagram to show these data. (You may avoid the fractions by rounding them off.)

Sector	*R&D Expenditure (Rs in crore)*		
	1994–95	1995–96	1996–97
Public	414.61	427.58	536.05
Private	1,318.87	1,627.07	1,796.96
Industrial	1,733.48	2,054.65	2,333.01

Source: GOI, *Department of Science & Technology*, New Delhi.

5.14 A company having a supermarket chain conducted a survey of its customer's opinion. Respondents were asked to express their opinion regarding the quality of the product sold by the company. The results showed that of the 400 car owners who responded, 220 rated the quality of products 'good', 120 rated it 'fair' and the remaining 'poor'. For non-car owners, the numbers were 147, 108 and 95 respectively. Tabulate this information adding suitable derived statistics, and draw a diagram or diagrams to show the results.

5.15 The following information shows the annual growth rate in major sectors of industry.

(Percent)

Year	*Mining*	*Manufacturing*	*Electricity*	*General*
1996–97	–2.0	6.7	4.0	5.6
1997–98	5.9	6.7	6.6	6.6
1998–99	–1.7	4.3	6.5	4.0

Source: Economic Survey, 1999–2000.

Depict the above information by means of a suitable diagram.

5.16 The following table shows the sectoral growth rates in GDP (at factor cost) for the two years 1998–99 and 1999–2000.

Sector(s)	*Years (Percentage Change over the Previous Year)*	
	1998–99	*1999–2000*
Agriculture and allied sectors	7.2	0.8
Industry	4.0	6.9
Services	8.3	8.2
Total GDP	6.8	5.9

Use the bar charts to show these data.

5.17 The data on literacy rate for India and a few selected states for the year 1991 are given below:

India and States	*Literacy Rate (Percent)*		
	Total	*Male*	*Female*
India	52.21	64.13	39.29
Delhi	75.29	82.01	66.99
Kerala	89.81	93.62	86.17
Punjab	58.51	65.66	50.41
Rajasthan	38.55	54.99	20.44

You are asked to show these data by a suitable diagram after rounding them off to their nearest integer.

5.18 Display the data given in Table 4.3 by means of a suitable diagram.

5.19 Draw a suitable diagram showing the information contained in Question 4.12.

5.20 Draw a suitable diagram showing the information given in Question 4.17.

CHAPTER 6

MEASURES OF CENTRAL TENDENCY

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the requirements of a good average
- calculate various types of averages—simple arithmetic mean, weighted mean, geometric and harmonic means
- calculate mode, median, quartiles, percentiles, and so on
- know the main properties of each measure of central tendency and select the most appropriate one for use with a given set of data.

Chapter Prerequisites

Before starting work on this chapter, you should ensure that you are conversant with

1. the construction of frequency tables and
2. quite comfortable in carrying out simple calculations.

6.1 INTRODUCTION

At the outset, it may be noted that the description of statistical data may be quite elaborate or quite brief, depending on two factors, namely, the nature of the data themselves and the purpose for which the data have been collected. While describing data statistically or verbally, one must ensure that the description is neither too brief nor too lengthy. In this chapter and the next one, we shall concentrate mainly on two types of descriptions—measures of central tendency and measures of variation.

The measures of central tendency (also called measures of location) that will be discussed here are the arithmetic mean, weighted mean, the median, the mode, the geometric mean and the harmonic mean. Before discussing arithmetic mean or any other mean, the question arises: Why should we use such a mean? The answer is that there are two main objects of using mean.

- *First*, to get a single value that indicates the characteristic of the entire data. For instance, when we talk of per capita income of a country, it gives a broad idea of the standard of living of the people in that country.

- *Second*, to facilitate comparisons. Measures of central tendency enable us to compare two or more distributions pertaining to the same time period or within the same distribution over time. For example, the average consumption of tea in two different territories for the same period or in a territory for two years, say, 1999 and 2000, can be attempted by means of an average.

6.2 THE ARITHMETIC MEAN

The arithmetic mean is obtained by adding all the observations and dividing the sum by the number of observations. Suppose we have the following observations:

10, 15, 30, 7, 42, 79 and 83

These are seven observations. Symbolically, the arithmetic mean, also called simply *mean* is

$\bar{x} = \Sigma x/n$, *where* $\bar{x}$ *is sample mean.*

This formula is the basic formula that forms the definition of arithmetic mean and is used in case of ungrouped data where weights are not involved.

$$= \frac{10+15+30+7+42+79+83}{7}$$

$$= \frac{266}{7} = 38$$

It may be noted that the Greek letter μ is used to denote the mean of the population and N to denote the total number of observations in a population. Thus, the population mean $\mu = \Sigma X/N$.

Ungrouped Data—Weighted Case

In case of ungrouped data where weights are involved, our approach for calculating arithmetic mean will be different from the one used earlier.

Example 6.1 Suppose a student has secured the following marks in three tests:

Mid-term test	30
Laboratory	25
Final	20

The simple arithmetic mean will be $\frac{30+25+20}{3} = 25$.

However, this will be wrong if the three tests carry different weights on the basis of their relative importance. Assuming that the weights assigned to the three tests are:

Mid-term test	2 points
Laboratory	3 points
Final	5 points

Solution On the basis of this information, we can now calculate a weighted mean as shown below:

Table 6.1 Calculation of a Weighted Mean

Type of Test	*Relative Weight (w)*	*Marks (x)*	*wx*
Mid-term	2	30	60
Laboratory	3	25	75
Final	5	20	100
Total	**$\Sigma w = 10$**	**—**	**235**

$$\bar{x} = \frac{\Sigma wx}{\Sigma w} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3}$$

$$= \frac{60 + 75 + 100}{2 + 3 + 5} = 23.5 \text{ marks}$$

It will be seen that weighted mean gives a more realistic picture than the simple or unweighted mean.

Let us take another example of weighted mean in the case of ungrouped data.

Example 6.2 An investor is fond of investing in equity shares. During a period of falling prices in the stock exchange, a stock is sold at Rs 120 per share on one day, Rs 105 on the next and Rs 90 on the third day. The investor has purchased 50 shares on the first day, 80 shares on the second day and 100 shares on the third day. What average price per share did the investor pay?

Solution

Table 6.2 Calculation of Weighted Average Price

Day	*Price Per Share (Rs) x*	*No. of Shares Purchased w*	*Amount Paid Rs wx*
1	120	50	6000
2	105	80	8400
3	90	100	9000
Total	**—**	**230**	**23,400**

$$\text{Weighted average} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3} = \frac{\Sigma wx}{\Sigma w}$$

$$= \frac{6000 + 8400 + 9000}{50 + 80 + 100} = \text{Rs } 101.7$$

Thus, the investor paid an average price of Rs 101.7 per share.

It will be seen that if merely prices of the shares for the three days (regardless of the number of shares purchased) are taken into consideration, then the average price would be

$$\frac{120 + 105 + 90}{3} = \text{Rs } 105$$

This is an unweighted or simple average and as it ignores the quantum of shares purchased, it fails to give a correct picture. A simple average, it may be noted, is also a weighted average where weight in each case is the same, that is, only 1. *When we use the term 'average' alone, we always mean that it is an unweighted or simple average.*

Arithmetic Mean in Case of Grouped Data

For grouped data, arithmetic mean may be calculated by applying any of the following methods:

(i) Direct method
(ii) Short-cut method
(iii) Step-deviation method

In the case of direct method, the formula $\bar{x} = \Sigma fm/n$ is used. Here 'm' is the mid-point of various classes, 'f' is the frequency of each class and 'n' is the total number of frequencies. The calculation of arithmetic mean by the direct method is shown below.

Example 6.3 The following table gives the marks of 58 students in Statistics. Calculate the average marks of this group.

Marks	*No. of Students*
0–10	4
10–20	8
20–30	11
30–40	15
40–50	12
50–60	6
60–70	2
Total	**58**

Solution

Table 6.3 Calculation of Arithmetic Mean by Direct Method

Marks	*Mid-point m*	*No. of Students f*	*fm*
0–10	5	4	20
10–20	15	8	120
20–30	25	11	275
30–40	35	15	525
40–50	45	12	540
50–60	55	6	330
60–70	65	2	130
			Σfm = 1940

$$\bar{x} = \frac{\Sigma fm}{n} = \frac{1940}{58} = 33.45 \text{ marks or 33 marks approximately.}$$

It may be noted that the mid-point of each class is taken as a good approximation of the true mean of the class. This is based on the assumption that the values are distributed fairly evenly throughout the interval. When large numbers of frequency occur, this assumption is usually accepted.

In the case of short-cut method, the concept of arbitrary mean is followed.

The formula for calculation of the arithmetic mean by the short-cut method is:

$$\bar{x} = A + \frac{\Sigma fd}{n}$$

where A = arbitrary or assumed mean
f = frequency
d = deviation from the arbitrary or assumed mean

When the values are extremely large and/or in fractions, the use of the direct method would be very cumbersome. In such cases, the short-cut method is preferable. This is because the calculation work in the short-cut method is considerably reduced particularly for calculation of the product of values and their respective frequencies. However, when calculations are not made manually but by a machine calculator, it may not be necessary to resort to the short-cut method, as the use of the direct method may not pose any problem.

As can be seen from the formula used in the short-cut method, an arbitrary or assumed mean is used. The second term in the formula—$\Sigma fd \div n$—is the correction factor for the difference between the actual mean and the assumed mean. If the assumed mean turns out to be equal to the actual mean, $\Sigma fd \div n$ will be zero.

The use of the short-cut method is based on the principle that the total of deviations taken from an actual mean is equal to zero. As such, the deviations taken from any other figure will depend on how the assumed mean is related to the actual mean. While one may choose any value as assumed mean, it would be proper to avoid extreme values, that is, too small or too high to simplify calculations. A value apparently close to the arithmetic mean should be chosen.

For the figures given earlier pertaining to marks obtained by 58 students, we calculate the average marks by using the short-cut method.

Example 6.4

Table 6.4 Calculation of Arithmetic Mean by Short-cut Method

Marks	*Mid-point m*	*f*	*d*	*fd*
0–10	5	4	–30	–120
10–20	15	8	–20	–160
20–30	25	11	–10	–110
30–40	35	15	0	0
40–50	45	12	10	120
50–60	55	6	20	120
60–70	65	2	30	60
				$\Sigma fd = -90$

It may be noted that we have taken arbitrary mean as 35 and deviations from mid-points. In other words, the arbitrary mean has been subtracted from each value of mid-point and the resultant figure is shown in column '*d*'.

$$\bar{x} = A + \frac{\Sigma fd}{n}$$
$$= 35 + \left(\frac{-90}{58}\right)$$
$$= 35 - 1.55 = 33.45 \text{ or } 33 \text{ marks approximately.}$$

Now we take up the calculation of arithmetic mean for the same set of data using the step-deviation method. This is shown in Table 6.5.

Example 6.5

Table 6.5 Calculation of Arithmetic Mean by Step-deviation Method

Marks	*Mid-point*	*f*	*d*	*d′ = d/10*	*fd′*
0–10	5	4	–30	–3	–12
10–20	15	8	–20	–2	–16
20–30	25	11	–10	–1	–11
30–40	35	15	0	0	0
40–50	45	12	10	1	12
50–60	55	6	20	2	12
60–70	65	2	30	3	6
					$\Sigma fd' = -9$

where '*d*' is divided by the common factor (which is 10) and *C* is the common factor in the formula used below.

$$\bar{x} = A + \frac{\Sigma fd'}{n} \times C$$
$$= 35 + \left(\frac{-9 \times 10}{58}\right)$$
$$= 33.45 \text{ or } 33 \text{ marks approximately.}$$

It will be seen that the answer in each of the three cases is the same. The step-deviation method is the most convenient on account of simplified calculations. It may also be noted that if we select a different arbitrary mean and recalculate deviations from that figure, we would get the same answer.

Now that we have learnt how the arithmetic mean can be calculated by using different methods, we are in a position to handle any problem where calculation of the arithmetic mean is involved. Let us take one example.

Example 6.6 The mean of the following frequency distribution was found to be 1.46.

No. of Accidents	*No. of Days (Frequency)*
0	46
1	?
2	?
3	25
4	10
5	5
Total	**200 days**

Calculate the missing frequencies.

Solution Here we are given the total number of frequencies and the arithmetic mean. We have to determine the two frequencies that are missing.

Let us assume that the frequency against 1 accident is X and against 2 accidents is Y. If we can establish two simultaneous equations, then we can easily find the values of X and Y.

$$\text{Mean} = \frac{(0\times 46) + (1\times x) + (2\times y) + (3\times 25) + (4\times 10) + (5\times 5)}{200}$$

$$1.46 = \frac{x + 2y + 140}{200}$$

$$x + 2y + 140 = (200)\ (1.46)$$

$$x + 2y = 152 \qquad \text{(i)}$$

$$x + y = 200 - \{46 + 25 + 10 + 5\}$$

$$x + y = 200 - 86$$

$$x + y = 114 \qquad \text{(ii)}$$

Now subtracting equation (ii) from equation (i), we get

$$\begin{array}{r} x + 2y = 152 \\ x + y = 114 \\ - \quad - \\ \hline y = 38 \end{array}$$

Substituting the value of $y = 38$ in equation (ii) above, $x + 38 = 114$.

Therefore, $x = 114 - 38 = 76$.

Hence, the missing frequencies are:

Against accident 1 : 76

Against accident 2 : 38

Combined Mean

Sometimes, we are given means of two related groups. We are then asked to find the combined mean taking both the groups together. In a problem of this type, the following formula is used.

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of group 1 and group 2 respectively;
n_1 and n_2 are the number of items in group 1 and group 2 respectively; and
$\bar{x}_{12}$ is the combined mean of the two groups.

Let us take an example using this formula.

Example 6.7 The mean height of 25 male workers in a factory is 61 inches and the mean height of 35 female workers in the same factory is 58 inches. Find the combined mean height of 60 workers in the factory.

Solution

Combined mean $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

$$= \frac{(25 \times 61) + (35 \times 58)}{25 + 35}$$

$$= \frac{1525 + 2030}{60}$$

$$= 59.25 \text{ inches}$$

Let us take another example.

Example 6.8 The mean annual salary of employees of a company is Rs 30000. The mean annual salaries of male and female employees are Rs 35000 and Rs 23000, respectively. Find out the percentage of male and female employees working in the company.

Solution

Given: $\bar{x}_1 = \text{Rs } 35000$, $\bar{x}_2 = \text{Rs } 23000$ and $\bar{x}_{12} = \text{Rs } 30000$

Now, $$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

or $$(\bar{x}_{12})(n_1 + n_2) = n_1\bar{x}_1 + n_2\bar{x}_2$$

or $$30000(n_1 + n_2) = n_1 (35000) + n_2 (23000)$$

or $$30000\, n_1 + 30000\, n_2 = 35000 n_1 + 23000\, n_2$$

or $$30000\, n_1 - 35000\, n_1 = 23000\, n_2 - 30000\, n_2$$

or $$-5000\, n_1 = -7000\, n_2$$

or $$n_1/n_2 = -5000/-7000$$

$$= 5/7$$

Hence, the percentage of male employees is

$$5/7 \times 100 = 500/7 = 71.43$$

and the percentage of female employees is

$$1 - 5/7 = 2/7$$

$$2/7 \times 100 = 200/7 = 28.57$$

It may be noted that the combined mean can be calculated not only for 2 related groups but for any number of groups.

Characteristics of the Arithmetic Mean

Some of the important *characteristics* of the arithmetic mean are:

1. The sum of the deviations of the individual items from the arithmetic mean is always zero. This means $\Sigma (x - \bar{x}) = 0$, where x is the value of an item and $\bar{x}$ is the arithmetic mean. 'Since the sum of the deviations in the positive direction is equal to the sum of the deviations in the negative direction, *the arithmetic mean is regarded as a measure of central tendency*.'
2. The sum of the squared deviations of the individual items from the arithmetic mean is always minimum. In other words, the sum of the squared deviations taken from any value other than the arithmetic mean will be higher.
3. As the arithmetic mean is based on all the items in a series, a change in the value of any item will lead to a change in the value of the arithmetic mean.

4. In the case of highly skewed distribution, the arithmetic mean may get distorted on account of a few items with extreme values. In such a case, it may cease to be the representative characteristic of the distribution.

Advantages and Disadvantages of Arithmetic Mean

Advantages

1. The calculation of arithmetic mean is quite simple, and is unique. This is because any data set can have only one mean.
2. Arithmetic mean is calculated by taking all the values in the given data set. This is obviously an advantage as none of the observations is ignored.
3. Arithmetic mean is least affected by sample size. This implies that when several samples are drawn from the same population, the variations in the arithmetic mean for the different samples would be by the least possible amount.
4. Another advantage of arithmetic mean is that it can be put to algebraic treatment.
5. The arithmetic mean is useful while making comparisons between several data sets. This will be taken up in the subsequent chapter where hypothesis test is performed for having two or more sample data.

Disadvantages

1. A major limitation of arithmetic mean is that, in some cases, it cannot be calculated. In case a frequency distribution has open-ended, or unequal class intervals, either at the beginning or end, it is not possible to calculate arithmetic mean accurately.
   It will be seen the frequency distribution table has an open-ended class, 91 and above. In this case, we cannot calculate the arithmetic mean as we do not know the exact width of the class. The figure can be 91, 92 or any other value. This clearly demonstrates the limitation of the arithmetic mean.
2. Another limitation of arithmetic mean is that if there are outliers, that is, very extreme values, in the data set, it ceases to be representative of the data set. This becomes obvious from a simple example. Suppose, monthly incomes of five employees of a company are Rs 1000, Rs 1200, Rs 1800, Rs 2500 and Rs 6000. In this case, the arithmetic mean is Rs 2500, which is an inflated figure, as incomes of three employees are much less than Rs 2500. The same is true when an observation is too small. In the above example, instead of Rs 6000, if the salary of the employee is, say, Rs 600, then the average salary would be Rs 1420 per month.
3. Another disadvantage is that in case of a large number of observations in the data set, it becomes tedious to calculate, as every observation point is included. However, this problem can be overcome by the short-cut method of using grouped data, though this would give an approximate mean.

The following table shows individual observations pertaining to daily wages of 8 workers in a factory:

Workers	1	2	3	4	5	6	7	8	9	10
Daily wages (Rs)	53	60	47	70	80	66	49	92	82	86

The same data, in the form of a frequency distribution, is shown below:

Daily wages (Rs)	*Number of workers*
40–50	2
51–60	2

61–70	2
71–80	1
81–90	2
91 & above	1

6.3 THE MEDIAN

Median is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude. Thus, in an ungrouped frequency distribution if the n values are arranged in ascending or descending order of magnitude, the median is the middle value if n is odd. When n is even, the median is the mean of the two middle values.

Suppose we have the following series:

15, 19, 21, 7, 10, 33, 25, 18 and 5

We have to first arrange it in either ascending or descending order. These figures are arranged in an ascending order as follows:

5, 7, 10, 15, 18, 19, 21, 25, 33

Now as the series consists of odd number of items, to find out the value of the middle item, we use the formula

$$\frac{n+1}{2}$$

where n is the number of items. In this case, n is 9, as such $\frac{n+1}{2} = 5$, that is, the size of the 5^{th} item is the median.

This happens to be 18.

Suppose the series consists of one more item, 23. We may, therefore, have to include 23 in the above series at an appropriate place, that is, between 21 and 25. Thus, the series is now 5, 7, 10, 15, 18, 19, 21, 23, 25, 33. Applying the above formula, the median is the size of 5.5^{th} item. Here, we have to take the average of the values of 5^{th} and 6^{th} item. This means an average of 18 and 19, which gives the median as 18.5.

It may be noted that the formula $\frac{n+1}{2}$ itself is not the formula for the median; it merely indicates the position of the median, namely, the number of items we have to count until we arrive at the item whose value is the median. In the case of the even number of items in the series, we identify the two items whose values have to be averaged to obtain the median.

In the case of a grouped series, the median is calculated by linear interpolation with the help of the following formula:

$$M = l_1 + \frac{l_2 - l_1}{f}(m - c)$$

where M = the median

l_1 = the lower limit of the class in which the median lies

l_2 = the upper limit of the class in which the median lies

f = the frequency of the class in which the median lies

m = the middle item or $(n + 1)/2^{th}$, where n stands for total number of items

c = the cumulative frequency of the class preceding the one in which the median lies.

Let us take an example of a frequency distribution for which the median is to be calculated.

Example 6.9

Monthly Wages (Rs)	*No. of Workers*
800–1,000	18
1,000–1,200	25
1,200–1,400	30
1,400–1,600	34
1,600–1,800	26
1,800–2,000	10
Total	**143**

In order to calculate median in this case, we have to first provide cumulative frequency to the table. Thus, the table with the cumulative frequency is written as:

Monthly Wages	*Frequency*	*Cumulative Frequency*
800–1,000	18	18
1,000–1,200	25	43
1,200–1,400	30	73
1,400–1,600	34	107
1,600–1,800	26	133
1,800–2,000	10	143

$$M = l_1 + \frac{l_2 - l_1}{f}(m - c)$$

$$m = \frac{n+1}{2} = \frac{143+1}{2} = 72$$

It means median lies in the class-interval Rs 1,200 – 1,400.

Now, $$M = 1200 + \frac{1400 - 1200}{30}(72 - 43)$$

$$= 1200 + \frac{200}{30}(29)$$

$$= \text{Rs } 1393.3$$

At this stage, let us introduce two other concepts, viz. *quartile* and *decile*. To understand these, we should first know that the median belongs to a general class of statistical descriptions called *fractiles*. A fractile is a value below which lies a given fraction of a set of data. In the case of the median, this fraction is one-half (1/2). Likewise, a quartile has a fraction one-fourth (1/4).

The three quartiles Q_1, Q_2 and Q_3 are such that 25 per cent of the data fall below Q_1, 25 per cent fall between Q_1 and Q_2, 25 per cent fall between Q_2 and Q_3, and 25 per cent fall above Q_3.

It will be seen that Q_2 is the median. We can use the above formula for the calculation of quartiles as well. The only difference will be in the value of m. Let us calculate both Q_1 and Q_3 in respect of the table given in Example 6.9.

$$Q_1 = l_1 + \frac{l_2 - l_1}{f}(m - c)$$

Here, m will be $= \frac{n+1}{4} = \frac{143+1}{4} = 36$

$$Q_1 = 1{,}000 + \frac{1{,}200 - 1{,}000}{25}(36 - 18)$$

$$= 1{,}000 + \frac{200}{25}(18)$$

$$= \text{Rs } 1{,}144$$

In the case of Q_3, m will be $3\frac{(n+1)}{4} = \frac{3 \times 144}{4} = 108$

$$Q_3 = 1{,}600 + \frac{1{,}800 - 1{,}600}{26}(108 - 107)$$

$$= 1{,}600 + \frac{200}{26}(1)$$

$$= \text{Rs } 1{,}607.7 \text{ approx.}$$

In the same manner, we can calculate *deciles* (where the series is divided into 10 equal parts) and percentiles (where the series is divided into 100 equal parts).

Advantages of Median

1. Unlike arithmetic mean, median is not affected at all by extreme values as it is a positional average. As such, median is particularly very useful when a distribution happens to be skewed.
2. Another point that goes in favour of median is that it can be computed when a distribution has open-end classes.
3. Another merit of median is that when a distribution contains qualitative data, it is the only average that can be used. No other average is suitable in case of such a distribution.

Let us take a couple of examples to illustrate what has been said in favour of median.

Example 6.10 Calculate the most suitable average for the following data:

Size of the Item	Below 50	50–100	100–150	150–200	200 and above
Frequency	15	20	36	40	10

Solution Since the data have two open-end classes—one in the beginning (below 50) and the other at the end (200 and above), median should be the right choice as a measure of central tendency.

Table 6.6 Computation of Median

Size of Item	*Frequency*	*Cumulative Frequency*
Below 50	15	15
50–100	20	35
100–150	36	71
150–200	40	111
200 and above	10	121

$$\text{Median} = \text{Size of } \frac{n+1}{2} \text{ th item}$$

$$= \frac{121+1}{2} = 61^{st} \text{ item}$$

Now, 61^{st} item lies in the 100–150 class.

$$\text{Median} = l_1 + \frac{l_2 - l_1}{f}(m - c)$$

$$= 100 + \frac{150 - 100}{36}(61 - 35)$$

$$= 100 + 36.11 = 136.11 \text{ approx.}$$

Let us take another example.

Example 6.11 The following data give the savings bank accounts balances of nine sample households selected in a survey. The figures are in rupees.

745 2,000 1,500 68,000 461 549 3,750 1,800 4,795

(a) Find the mean and the median for these data.

(b) Do these data contain an outlier? If so, exclude this value and recalculate the mean and median. Which of these summary measures has a greater change when an outlier is dropped?

(c) Which of these two summary measures is more appropriate for this series?

Solution

(a) $$\text{Mean} = \text{Rs } \frac{745 + 2,000 + 1,500 + 68,000 + 461 + 549 + 3,750 + 1,800 + 4,795}{9}$$

$$= \text{Rs } \frac{83,600}{9} = \text{Rs } 9,289$$

$$\text{Median} = \text{Size of } \frac{n+1}{2} \text{ th item}$$

$$= \frac{9+1}{2} = 5^{th} \text{ item}$$

Arranging the data in an ascending order, we find that the median is Rs 1,800.

(b) An item of Rs 68,000 is excessively high. Such a figure is called an 'outlier'. We exclude this figure and recalculate both the mean and the median.

$$\text{Mean} = \text{Rs} \frac{83,600 - 68,000}{8}$$

$$= \text{Rs} \frac{15,600}{8} = \text{Rs } 1,950$$

$$\text{Median} = \text{Size of } \frac{n+1}{2} \text{ th item}$$

$$= \frac{8+1}{2} = 4.5^{th} \text{ item}$$

$$= \frac{1,500 + 1,800}{2} = \text{Rs } 1,650$$

It will be seen that the mean shows a far greater change than the median when the outlier is dropped from the calculations.

(c) As far as these data are concerned, the median will be a more appropriate measure than the mean.

Determining Median Graphically

In Chapter 4, while concluding the discussion on the ogive curve, it was mentioned that we could determine the median graphically. As such, we illustrate here how median can be determined graphically.

Example 6.12 Suppose we are given the following series:

Class interval	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Frequency	6	12	22	37	17	8	5

We are asked to draw both types of ogives from these data and to determine the median.

Solution First of all, we transform the given data into two cumulative frequency distributions, one based on 'less than' and another on 'more than' methods.

Table 6.7

	Frequency
Less than 10	6
Less than 20	18
Less than 30	40
Less than 40	77
Less than 50	94
Less than 60	102
Less than 70	107

Table 6.8

	Frequency
More than 0	107
More than 10	101
More than 20	89
More than 30	67
More than 40	30
More than 50	13
More than 60	5

Figure 6.1, below, shows the two ogives.

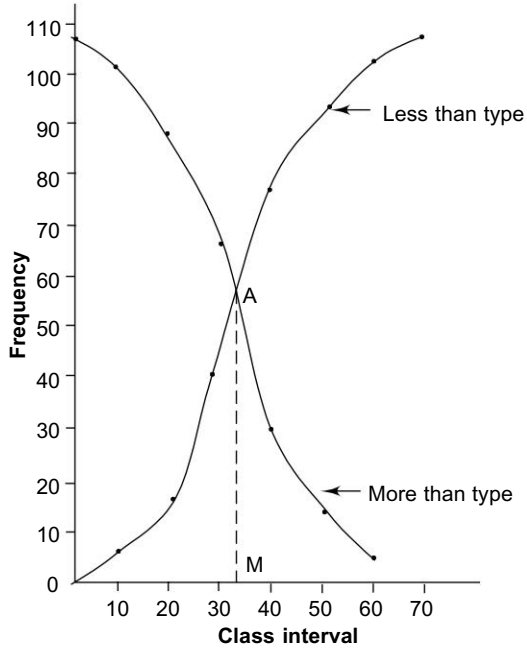


Fig. 6.1 **'Less than' and 'More than' Ogives**

It may be noted that the point of intersection of the two ogives gives the value of the median. From this point of intersection A, we draw a straight line to meet the X-axis at M. Thus, from the point of origin to the point at M gives the value of the median, which comes to 34, approximately. If we calculate the median by applying the formula, then the answer comes to 33.8, or 34, approximately.

It may be pointed out that even a single ogive can be used to determine the median. As we have determined the median graphically, so also we can find the values of quartiles, deciles or percentiles graphically. For example, to determine Q_3 we have to take size of $\{3(n + 1)\}/4 = 81^{st}$ item. From this point on the Y-axis, we can draw a perpendicular to meet the 'less than' ogive from which another straight line is to be drawn to meet the X-axis. This point will give us the value of the upper quartile. In the same manner, other values of Q_1, and deciles and percentiles can be determined.

Characteristics of Median

Some of the important *characteristics* of the median are:

1. Unlike the arithmetic mean, the median can be computed from open-ended distributions. This is because it is located in the median class-interval, which would not be an open-ended class.
2. The median can also be determined graphically whereas the arithmetic mean cannot be ascertained in this manner.
3. As it is not influenced by the extreme values, it is preferred in case of a distribution having extreme values.
4. In case of the qualitative data where the items are not counted or measured but are scored or ranked, it is the most appropriate measure of central tendency.

6.4 THE MODE

The mode is another measure of central tendency. It is the value at the point around which the items are most heavily concentrated. As an example, consider the following series:

8, 9, 11, 15, 16, 12, 15, 3, 7, 15

There are ten observations in the series wherein the figure 15 occurs maximum number of times—three. The mode is, therefore, 15. The series given above is a discrete series; as such, the variable cannot be in fraction. If the series were continuous, we could say that the mode is approximately 15, without further computation.

In the case of grouped data, mode is determined by the following formula:

$$\text{Mode} = l_1 + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

where l_1 = the lower value of the class in which the mode lies
f_1 = the frequency of the class in which the mode lies
f_0 = the frequency of the class preceding the modal class
f_2 = the frequency of the class succeeding the modal class
i = the class-interval of the modal class

While applying the above formula, we should ensure that the class-intervals are uniform throughout. If the class-intervals are not uniform, then they should be made uniform on the assumption that the frequencies are evenly distributed throughout the class. In the case of inequal class-intervals, the application of the above formula will give misleading results.

Example 6.13 Let us take the following frequency distribution:

Class-intervals (1)	*Frequency (2)*
30–40	4
40–50	6
50–60	8
60–70	12
70–80	9
80–90	7
90–100	4

We have to calculate the mode in respect of this series.

Solution We can see from column (2) of the table that the maximum frequency of 12 lies in the class-interval of 60–70. This suggests that the mode lies in this class-interval. Applying the formula given earlier, we get:

$$\text{Mode} = 60 + \frac{12 - 8}{(12 - 8) + (12 - 9)} \times 10$$

$$= 60 + \frac{4}{4 + 3} \times 10$$

$$= 65.7 \text{ approx.}$$

In several cases, just by inspection one can identify the class-interval in which the mode lies. One should see which is the highest frequency and then identify to which class-interval this frequency belongs. Having done this, the formula given for calculating the mode in a grouped frequency distribution can be applied.

At times, it is not possible to identify by inspection the class where the mode lies. In such cases, it becomes necessary to use the method of grouping. This method consists of two parts: **(i)** preparation of a grouping table, and **(ii)** preparation of an analysis table.

(i) *Preparation of a grouping table* A grouping table has six columns, the first column showing the frequencies as given in the problem. Column 2 shows frequencies grouped in two's, starting from the top. Leaving the first frequency, column 3 shows frequencies grouped in two's. Column 4 shows the frequencies of the first three items, then second to fourth item and so on. Column 5 leaves the first frequency and groups the remaining items in three's. Column 6 leaves the first two frequencies and then groups the remaining in three's. Now, the maximum total in each column is marked and shown either in a circle or in a bold type.

(ii) *Preparation of an analysis table* After having prepared a grouping table, an analysis table is prepared. On the left-hand side, provide the first column for column numbers and on the right-hand side the different possible values of mode. The highest values marked in the grouping table are shown here by a bar or by simply entering 1 in the relevant cell corresponding to the values they represent. The last row of this table will show the number of times a particular value has occurred in the grouping table. The highest value in the analysis table will indicate the class-interval in which the mode lies. The procedure of preparing both the grouping and analysis tables to locate the modal class will be clear by the following example.

Example 6.14 The following table gives some frequency data:

Size of Item	*Frequency*
10–20	10
20–30	18
30–40	25
40–50	26
50–60	17
60–70	4

Solution **Grouping Table**

Size of item	*1*	*2*	*3*	*4*	*5*	*6*
10–20	10					
		28				
20–30	18			53		
			43			
30–40	25				69	
		51				
40–50	26					68
			43			
50–60	17			47		
		21				
60–70	4					

Analysis Table

Col. No.	*Size of Item*				
	10–20	*20–30*	*30–40*	*40–50*	*50–60*
1				1	
2			1	1	
3		1	1	1	1
4	1	1	1		
5		1	1	1	
6			1	1	1
Total	**1**	**3**	**5**	**5**	**2**

This is a bimodal series as is evident from the analysis table, which shows that the two classes 30–40 and 40–50 have occurred five times each in the grouping. In such a situation, we may have to determine mode indirectly by applying the following formula:

Mode = 3 median – 2 mean

Median = Size of $(n + 1)/2^{th}$ item, that is, $101/2 = 50.5^{th}$ item. This lies in the class 30–40. Applying the formula for the median, as given earlier, we get

$$= 30 + \frac{40 - 30}{25} (50.5 - 28)$$
$$= 30 + 9 = 39$$

Now, arithmetic mean is to be calculated. This is shown in the following table.

Class-interval	*Frequency*	*Mid-points*	*d*	*d′ = d/10*	*fd′*
10–20	10	15	–20	–2	–20
20–30	18	25	–10	–1	–18
30–40	25	35	0	0	0
40–50	26	45	10	1	26
50–60	17	55	20	2	34
60–70	4	65	30	3	12
Total	**100**				**34**

Deviation is taken from arbitrary mean = 35

$$\text{Mean} = A + \frac{\Sigma fd'}{n} \times i$$
$$= 35 + \frac{34}{100} \times 10$$
$$= 38.4$$

Mode = 3 median – 2 mean
$$= (3 \times 39) - (2 \times 38.4)$$
$$= 117 - 76.8$$
$$= 40.2$$

This formula, Mode = 3 Median – 2 Mean, is an empirical formula only. And it can give only approximate results. As such, its frequent use should be avoided. However, when mode is ill-defined or the series is bimodal (as is the case in the present example) it may be used.

In the next two examples, the above formula is not used as the mode is well defined.

Example 6.15 The regional transport authority is concerned about the speed of motor bikes. College students are driving on a section of the road. Following are the speeds of 45 drivers (speed in km/hr).

15	31	44	56	38	32	48	42	58	29
45	49	38	48	62	46	56	52	47	49
42	52	55	52	69	39	39	58	37	18
68	48	62	64	61	47	69	58	29	55
18	61	48	55	49					

Calculate the mode from the above data.

Solution The given data are first transformed into frequency distribution.

Class	*Frequency*
15–24	3
25–34	4
35–44	8
45–54	14
55–64	14
65–74	2
Total	45

As can be seen from the table, the last two classes, viz., 55–64 and 65–74, give the speed in excess of 55 km/hr. The frequency of these two classes adds to 16. Hence, $\frac{16}{45} \times 100 = 35.6\%$.

In order to calculate the mode, grouping table and analysis table have been prepared below:

Grouping Table

Class	*Frequency*			
	1	*2*	*3*	*4*
15–24	3	7		
25–34	4		12	15
35–44	8	22		
45–54	14		28	30
55–64	14	16		
65–74	2			

Analysis Table

Col. No.	*Class*					
	15–24	*25–34*	*35–44*	*45–54*	*55–64*	*65–74*
1				1	1	
2			1	1		
3				1	1	
4				1	1	1
Total	—	—	1	4	3	1

From the above table, it is clear that the mode lies in 45–54 class. Applying the formula,

$$\text{Mode} = l_1 + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

$$= 45 + \frac{14 - 8}{(14 - 8) + (14 - 14)} \times 9$$

$$= 45 + \frac{6}{6 + 0} \times 9$$

$$= 45 + \frac{6}{6} \times 9 = 45 + 9 = 54$$

Example 6.16 The expenditure details of 1000 families is given below:

Expenditure in Rs	40–59	60–79	80–99	100–119	120–139
No. of families	50		500		50

The median of distribution is Rs 87. Calculate the missing frequencies and, for the complete distribution table, calculate mode.

Solution Since the median is given as 87, we can straightaway apply the formula.

$$\text{Median} = l_1 + \frac{l_2 - l_1}{f}(m - c) = 87$$

It is obvious that the median lies in the class 80–99. Applying the values in the above formula.

$$\text{Median} = 80 + \frac{99 - 80}{500}(500.5 - c) = 87$$

or $$87 = 80 + \frac{19}{500}(500.5 - c)$$

or $$87 = 80 + 19.019 - \frac{19c}{500}$$

or $$87 = 99.019 - \frac{19c}{500}$$

or $$87 = \frac{49509.5 - 19c}{500}$$

or $$43500 = 49509.5 - 19c$$

or $$43500 - 49509.5 = -19c$$

or $$-19c = -6009.5$$

$$\therefore \quad c = \frac{-6009.5}{-19} = 316.289$$

As the table does not contain any fraction, this figure may be taken as 316.

It may be noted that 316 is the cumulative frequency of the class preceding the median class, that is, 60–79. Hence, the actual frequency of 60–79 class would be 316 – 50 (frequency of 40–59 class) = 266.

Now, as the total number of families is 1000, the missing frequency, x, can be calculated as follows:

$$50 + 266 + 500 + x + 50 = 1000$$

or $$x = 1000 - 816$$

or $$x = 134$$

Thus, the complete table is as follows:

Expenditure in Rs	40–59	60–79	80–99	100–119	120–139
No. of families	50	266	500	134	50

As the maximum frequency, 500, is in 80–99 class, the mode lies in this class. Applying the formula,

$$\text{Mode} = l_1 + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

$$= 80 + \frac{500 - 266}{(500 - 266) + (500 - 134)} \times 19$$

$$= 80 + \frac{234}{234 + 366} \times 19$$

$$= 80 + \frac{234}{600} \times 19$$

$$= 80 + 7.41$$

$$= 87.41 \text{ or } 87 \text{ approx.}$$

6.5 COMPARISON OF THE MEAN, MEDIAN AND MODE

Having discussed mean, median and mode, we now turn to the relationship amongst these three measures of central tendency. We shall discuss the relationship assuming that there is an unimodal frequency distribution.

When a distribution is symmetrical, the mean, median and mode are the same, as is shown below in Fig. 6.2.

In case, a distribution is skewed to the right, then mean > median > mode. Generally, income distribution is skewed to the right where a large number of families have relatively low income and a small

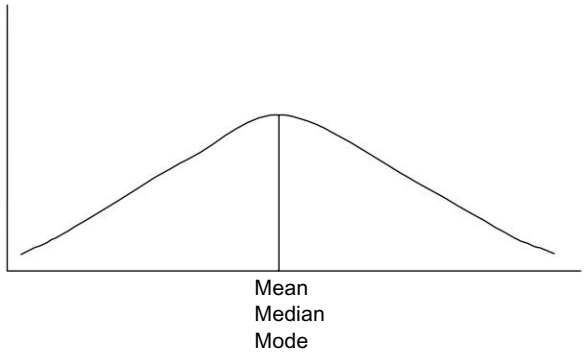


Fig. 6.2 A Symmetrical Distribution

number of families have extremely high income. In such a case, the mean is pulled up by the extreme high incomes and the relation among these three measures is as shown in Fig. 6.3. Here, we find that mean > median > mode.

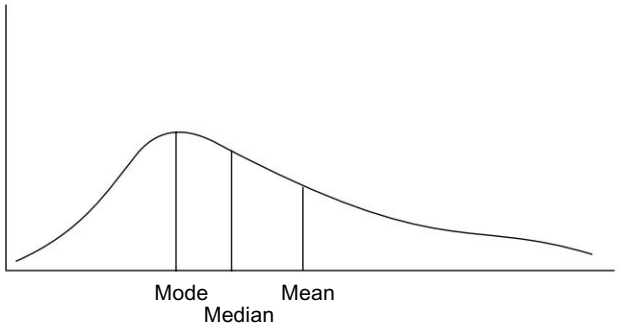


Fig. 6.3 A Positively Skewed Distribution

When a distribution is skewed to the left, then mode > median > mean. This is because here mean is pulled down below the median by extremely low values. This is shown in Fig. 6.4.

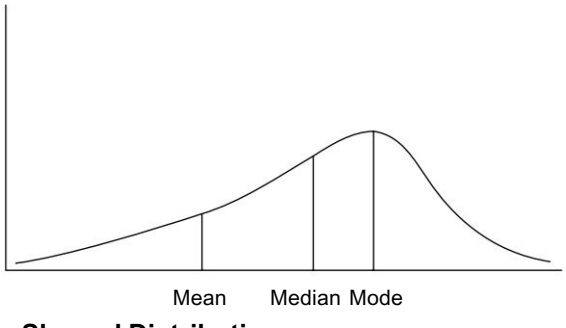


Fig. 6.4 A Negatively Skewed Distribution

Given the mean and median of a unimodal distribution, we can determine whether it is skewed to the right or left. When mean > median, it is skewed to the right; when median > mean, it is skewed to the left. It may be noted that the median is always in the middle between mean and mode.

Which of the Three Measures is the Best?

As mentioned earlier, in a symmetrical distribution, the mean, median and mode have the same value. As such, the problem of choosing a particular measure of central tendency does not arise in such a distribution. When the distribution is skewed either positively or negatively, a decision has to be made as to which measure of central tendency is the most appropriate. There is no simple answer to this question. It is because these three measures are based upon different concepts. The arithmetic mean is the sum of the values divided by the number of observations in the series. The median is the value of the middle observation that divides the series into two equal parts. Mode is the value around which the observations tend to concentrate. As such, the use of a particular measure will largely depend on the purpose of the study and the nature of the data. For example, when we are interested in knowing the consumers' preferences for different brands of television sets or different kinds of advertising the choice should go in favour of mode. The use of mean and median would not be proper. When the data are qualitative in nature or when the data have extreme values, the mean will not be proper. Instead, the median will be most appropriate measure. A point to note is that in a skewed series, median is always in the middle between mean and mode.

Further, it is not influenced by extreme values in the data. This apart, it can be used in an open-ended series where mean cannot be used.

The mean is, no doubt, the most frequently used measure. As the least square criterion is frequently used in statistics, the mean is consistent with several statistical techniques. However, when the data have one or two outliers, the use of mean would give a distorted figure. In such a case, the median is preferable.

6.6 THE GEOMETRIC MEAN

Apart from the three measures of central tendency as discussed above, there are two other means that are used sometimes in business and economics. These are the geometric mean and the harmonic mean. The geometric mean is more important than the harmonic mean. We discuss below both these means. First, we take up the geometric mean.

Geometric mean is defined at the n^{th} root of the product of n *observations of a distribution.*

Symbolically, $GM = \sqrt[n]{x_1 \cdot x_2 \ldots x_n}$

If we have only two observations, say, 4 and 16 then GM = $\sqrt{4 \times 16} = \sqrt{64} = 8$. Similarly, if there are three observations, then we have to calculate the cube root of the product of these three observations; and so on. When the number of items is large, it becomes extremely difficult to multiply the numbers and to calculate the root. To simplify calculations, logarithms are used.

Example 6.17 If we have to find out the geometric mean of 2, 4 and 8, then we find

$$\log GM = \frac{\Sigma \log x_i}{n}$$

$$= \frac{\log 2 + \log 4 + \log 8}{3}$$

$$= \frac{0.3010 + 0.6021 + 0.9031}{3}$$

$$= \frac{1.8062}{3} = 0.60206$$

$$\text{GM} = \text{Antilog } 0.60206$$

$$= 4$$

When the data are given in the form of a frequency distribution, then the geometric mean can be obtained by the formula:

$$\log \text{GM} = \frac{f_1 \cdot \log x_1 + f_2 \cdot \log x_2 + \ldots + f_n \cdot \log x_n}{f_1 + f_2 + \ldots f_n}$$

$$= \frac{\Sigma f \cdot \log x}{n}$$

Then, $\text{GM} = \text{Antilog}\left(\frac{\Sigma f \cdot \log x}{n}\right)$

The geometric mean is most suitable in the following three cases:

1. Averaging rates of change
2. The compound interest formula
3. Discounting, capitalisation.

Let us first take averaging rates of change.

Example 6.18 A person has invested Rs 5,000 in the stock market. At the end of the first year the amount has grown to Rs 6,250; he has had a 25 per cent profit. If at the end of the second year his principal has grown to Rs 8,750, the rate of increase is 40 per cent for the year. What is the average rate of increase of his investment during the two years?

Here, the use of geometric mean will be most suitable.

Solution

$$\text{GM} = \sqrt{1.25 \times 1.40} = \sqrt{1.75} = 1.323$$

The average rate of increase in the value of investment is therefore 1.323 – 1 = 0.323, which if multiplied by 100, gives the rate of increase as 32.3 per cent.

Example 6.19 We can also derive a *compound interest formula* from the above set of data. This is shown below:

Solution Now, $1.25 \times 1.40 = 1.75$. This can be written as $1.75 = (1 + 0.323)^2$.

Let $P_2 = 1.75$, $P_0 = 1$, and $r = 0.323$, then the above equation can be written as $P_2 = (1 + r)^2$ or $P_2 = P_0 (1 + r)^2$.

Where P_2 is the value of investment at the end of the second year, P_0 is the initial investment and r is the rate of increase in the two years. This, in fact, is the familiar compound interest formula. This can be written in a generalised form as $P_n = P_0 (1 + r)^n$.

In our case P_0 is Rs 5,000 and the rate of increase in investment is 32.3 per cent. Let us apply this formula to ascertain the value of P_n, that is, investment at the end of the second year.

$$P_n = 5{,}000 (1 + 0.323)^2$$

$$= 5{,}000 \times 1.75$$

$$= \text{Rs } 8{,}750$$

It may be noted that in the above example, if the arithmetic mean is used, the resultant figure will be wrong. In this case, the average rate for the two years is $\frac{25+40}{2}$ per cent per year, which comes to 32.5.

Applying this rate, we get $P_n = \frac{165}{100} \times 5{,}000$

$= \text{Rs } 8{,}250$

This is obviously wrong as the figure should have been Rs 8,750.

Example 6.20 An economy has grown at 5 per cent in the first year, 6 per cent in the second year, 4.5 percent in the third year, 3 per cent in the fourth year and 7.5 per cent in the fifth year. What is the average rate of growth of the economy during the five years?

Solution

Year	*Rate of Growth (Per cent)*	*Value at the End of the Year x (in Rs)*	*Log x*
1	5	105	2.02119
2	6	106	2.02531
3	4.5	104.5	2.01912
4	3	103	2.01284
5	7.5	107.5	2.03141
			$\Sigma \log X = 10.10987$

$$\text{GM} = \text{Antilog}\left(\frac{\Sigma \log x}{n}\right)$$

$$= \text{Antilog}\left(\frac{10.10987}{5}\right)$$

$= \text{Antilog } 2.021974$

$= 105.19$

Hence, the average rate of growth during the five-year period is 105.19 – 100 = 5.19 per cent per annum. In case of a simple arithmetic average, the corresponding rate of growth would have been 5.2 per cent per annum.

Discounting and Capitalisation

The compound interest formula given above was

$$P_n = P_0 (1 + r)^n$$

This can be written as

$$P_0 = \frac{P_n}{(1+r)^n}$$

This may be expressed as follows:

If the future income is P_n rupees and the present rate of interest is 100 r per cent, then the present value of P_n rupees will be P_0 rupees. For example, if we have a machine that has a life of 20 years and

is expected to yield a net income of Rs 50,000 per year, and at the end of 20 years it will be obsolete and cannot be used, then the machine's present value is

$$\frac{50,000}{(1+r)}+\frac{50,000}{(1+r)^2}+\frac{50,000}{(1+r)^3}+\ldots+\frac{50,000}{(1+r)^{20}}$$

This process of ascertaining the present value of future income by using the interest rate is known as *discounting*.

In conclusion, it may be said that when there are extreme values in a series, geometric mean should be used as it is much less affected by such values. The arithmetic mean in such cases will give misleading results.

Before we close our discussion on the geometric mean, we should be aware of its advantages and limitations.

Advantages

1. Geometric mean is based on each and every observation in the data set.
2. It is rigidly defined.
3. It is more suitable while averaging ratios and percentages as also in calculating growth rates.
4. As compared to the arithmetic mean, it gives more weight to small values and less weight to large values. As a result of this characteristic of the geometric mean, it is generally less than the arithmetic mean. At times it may be equal to the arithmetic mean.
5. It is capable of algebraic manipulation. If the geometric mean of two or more series is known along with their respective frequencies, then a combined geometric mean can be calculated by using the logarithms.

Limitations

1. As compared to the arithmetic mean, geometric mean is difficult to understand.
2. Both computation of the geometric mean and its interpretation are rather difficult.
3. When there is a negative item in a series or one or more observations have zero value, then the geometric mean cannot be calculated.

In view of the limitations mentioned above, the geometric mean is not frequently used.

6.7 THE HARMONIC MEAN

The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of individual observations. Symbolically,

$$\text{HM} = \frac{n}{1/x_1+1/x_2+1/x_3+\ldots+1/x_n} = \text{Reciprocal}\,\frac{\Sigma 1/x}{n}$$

The calculation of harmonic mean becomes very tedious when a distribution has a large number of observations.

In the case of grouped data, the harmonic mean is calculated by using the following formula:

$$\text{HM} = \text{Reciprocal of } \sum_{i=1}^{n}\left(f_i \times \frac{1}{x_i}\right)$$

or $$\frac{n}{\sum_{i=1}^{n}\left(f_i \times \frac{1}{x_i}\right)}$$

where n is the total number of observations.

It is worth noting that the harmonic mean is always lower than the geometric mean, which is lower than the arithmetic mean. This is because the harmonic mean assigns lesser importance to higher values. Since the harmonic mean is based on reciprocals, it becomes clear that as reciprocals of higher values are lower than those of lower values, it is a lower average than the arithmetic mean as well as the geometric mean.

Example 6.21 ***Ungrouped Data—Individual Observations***

Suppose we have three observations—4, 8 and 16. We are required to calculate the harmonic mean. Reciprocals of 4, 8 and 16 are:

$$\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \text{ respectively.}$$

Since $$\text{HM} = \frac{n}{1/x_1 + 1/x_2 + 1/x_3}$$

$$= \frac{3}{1/4 + 1/8 + 1/16}$$

$$= \frac{3}{0.25 + 0.125 + 0.0625}$$

$$= 6.857 \text{ approx.}$$

Example 6.22 ***Calculation of the harmonic mean in the grouped frequency distribution***

Consider the following series:

Class-interval	2–4	4–6	6–8	8–10
Frequency	20	40	30	10

Let us set up the worksheet as follows:

Class-interval	*Mid-value*	*Frequency*	*Reciprocal of MV*	$f \times 1/x$
2–4	3	20	0.3333	6.6660
4–6	5	40	0.2000	8.0000
6–8	7	30	0.1429	4.2870
8–10	9	10	0.1111	1.1111
			Total	**20.0641**

$$\frac{1}{\text{HM}} = \frac{\sum_{i=1}^{n}\left(f_i \times \frac{1}{x_i}\right)}{n} = \frac{20.0641}{100}$$

$$\text{HM} = \frac{100}{20.0641} = 4.984 \text{ approx.}$$

Example 6.23 In a small company, two typists are employed. Typist A types one page in ten minutes while typist B takes twenty minutes for the same.

(i) Both are asked to type 10 pages. What is the average time taken for typing one page?

(ii) Both are asked to type for one hour. What is the average time taken by them for typing one page?

Solution Here Q-(i) is on arithmetic mean while Q-(ii) is on harmonic mean.

(i)
$$M = \frac{(10 \times 10) + (20 \times 10)\ \text{(minutes)}}{10 \times 2\ \text{(pages)}}$$
$$= 15 \text{ minutes}$$

(ii)
$$HM = \frac{60 \times 2\ \text{(minutes)}}{60/10 + 60/20\ \text{(pages)}}$$
$$= \frac{120}{\frac{120 + 60}{20}} = \frac{40}{3} = 13 \text{ minutes and } 20 \text{ seconds.}$$

Example 6.24 It takes ship A 10 days to cross the Pacific Ocean; ship B takes 15 days and ship C takes 20 days.

(i) What is the average number of days taken by a ship to cross the Pacific Ocean?

(ii) What is the average number of days taken by a cargo to cross the Pacific Ocean when the ships are hired for 60 days?

Here again Q-(i) pertains to simple arithmetic mean while Q-(ii) is concerned with the harmonic mean.

(i)
$$M = \frac{10 + 15 + 20}{3} = 15 \text{ days}$$

(ii)
$$HM = \frac{60 \times 3\ \text{(days)}}{60/10 + 60/15 + 60/20}$$
$$= \frac{180}{\frac{360 + 240 + 180}{60}}$$
$$= 13.8 \text{ days approx.}$$

Advantages of Harmonic Mean

1. It takes into consideration all observations in a series.
2. It is amenable to algebraic treatment.
3. It is most appropriate when greater weightage needs to be given to the small observations and less weightage to the large ones.
4. In problems involving time and rates, it provides better results than other averages.

Limitations of Harmonic Mean

1. It is not easily understood.
2. Its computation is rather difficult.

3. When a series consists of both positive and negative values or when one of the items is zero, it cannot be computed.
4. As it gives the largest weight to the smallest item, it is not appropriate in the analysis of economic data.

6.8 THE QUADRATIC MEAN

We have seen earlier that the geometric mean is the antilogarithm of the arithmetic mean of the logarithms, and the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. Likewise, the quadratic mean (Q) is the square root of the arithmetic mean of the squares.
Symbolically,

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$$

Instead of using original values, the quadratic mean can be used while averaging deviations when the standard deviation is to be calculated. This will be used in the next chapter on dispersion.

Relative Position of Different Means

The relative position of different means will always be:

$$Q > \bar{x} > G > H$$

provided that all the individual observations in a series are positive and all of them are not the same.

Composite Average or Average of Means

Sometimes, we may have to calculate an average of several averages. In such cases, we should use the same method of averaging, that was employed in calculating the original averages. Thus, we should calculate the arithmetic mean of several values of $\bar{x}$, the geometric mean of several values of GM, and the harmonic mean of several values of HM. It will be wrong if we use some other average in averaging of means.

Additional Examples

Example 6.25 In a factory, there are 100 skilled workers, 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 2 hours, by a semi-skilled worker in 3 hours and by an unskilled worker in 4 hours. After a training of two years, the semi-skilled workers are expected to become skilled and unskilled workers to become semi-skilled. How much less time will be required after two years of training for weaving the unit length of fabric by an average worker?

Solution

Worksheet

Workers	*No.*	*Time Taken*	*Total Time*	*After Training*	
				Time Taken	*Total Time*
Skilled	100	2	200	2	200
Semi-skilled	250	3	750	2	500
Unskilled	150	4	600	3	450
	500		1550		1150

Average time taken by a worker for providing one-unit length of fabric

$$M_1 = \frac{1550}{500} = 3.1 \text{ hours}$$

$$M_2 = \frac{1150}{500} = 2.3 \text{ hours}$$

$$M_1 - M_2 = 3.1 - 2.3 = 0.8 \text{ hour}$$

Hence, after two years of training, an average worker will take 0.8 hour less time for weaving the unit length of fabric.

Example 6.26 From the following data, determine **(a)** the lowest mark scored by the top 10% of the class; and **(b)** the highest mark scored by the lowest 20% of the class.

Marks	*No. of Students*	
90–100	9	120
80–89	32	111
70–79	43	79
60–69	21	36
50–59	11	15
40–49	3	4
30–39	1	1

Solution Highest 10% students would be 12 students.

(a) $l_1 + \frac{l_2 - l_1}{f_1}(m - c) = 80 + \frac{89 - 80}{32} \times (12 - 9)$

$$= 80 + \frac{9}{32} \times 3 = 80 + \frac{27}{32} = 81 \text{ approx.}$$

(b) Lowest 20% students means $\frac{\cancel{20}^{2}}{\cancel{100}} \times \cancel{120} = 24$ students

This lies in 60–69 class.

$$l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

$$= 60 + \frac{69 - 60}{21} \times (24 - 15)$$

$$= 60 + \frac{9}{21} \times (24 - 15)$$

$$= 60 + \frac{81}{21} = 60 + 3.86 = 64 \text{ approx.}$$

Example 6.27 A number of particular articles have been classified according to their weights. After drying for two weeks, the same articles have again been weighed and similarly classified. It is known that the median weight in the first weighing was 20.83 gm while in the second weighing it was 17.35 gm. Some frequencies a and b in the first weighing and x and y in the second are missing. It is known that $a = 1/3x$ and $b = 1/2y$. Find out the values of the missing frequencies.

Class	*Frequencies*	
	First weighing	*Second 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

Solution

Given $a = 1/3\ x$ and $b = 1/2\ y$

Median weight 1 = 20.83 gm

Median weight 2 = 17.35 gm

The formula for calculating median is $l_1 + \frac{l_2 - l_1}{f_1}(m - c)$.

Since median is given as 20.83 in the first weighing, it is clear that it lies in the 20–25 class. Applying the formula given above and taking $(n + 1)/2$ as the size of the middle item m, we get

$$20 + \frac{5}{75}\left[\left(\frac{161 + a + b}{2}\right) - (a + b + 63)\right] = 20.83$$

or $$\frac{1}{15}\left(\frac{161 + a + b - 2a - 2b - 126}{2}\right) = 20.83 - 20$$

or $$35 - a - b = 2(0.83 \times 15)$$

or $$35 - \frac{1}{3}x - \frac{1}{2}y = 24.9$$

or $$-\frac{1}{3}x - \frac{1}{2}y = 24.9 - 35$$

or $$\frac{1}{3}x + \frac{1}{2}y = 10.1 \qquad (1)$$

Now, we take the second weighing and proceed in the same manner.

$$15 + \frac{1}{10}\left[\left(\frac{149 + x + y}{2}\right) - (x + y + 40)\right] = 17.35$$

or $$\frac{1}{10}\left(\frac{149 + x + y - 2x - 2y - 80}{2}\right) = 17.35 - 15$$

or $$\frac{69 - x - y}{2} = 2.35 \times 10$$

or $$69 - x - y = 47$$

or $$-x - y = 47 - 69$$

or $$x + y = 22 \qquad (2)$$

Multiplying equation (1) by 3, we get

$$x + 1.5y = 30.3 \quad (3)$$

$$x + y = 22 \quad (2)$$

Subtracting equation (2) from equation (3), we get

$$0.5y = 8.3$$

Hence $y = 16.6$

Substituting the value of $y = 16.6$ in (2)

$$x + 16.6 = 22$$

Hence $x = 22 - 16.6 = 5.4$

Now, series A, $a = \frac{1}{3}x$ or $\frac{5.4}{3} = 1.8$

$$b = \frac{1}{2}y \quad \text{or} \quad \frac{16.6}{2} = 8.3$$

Rounding them off to the nearest integer, we get

$a = 2, b = 8$ and $x = 6, y = 16$

Let us take an example involving the calculation of the geometric mean of a frequency distribution.

Example 6.28 Calculate the geometric mean from the following data:

x	2	4	8	10	32
f	1	4	6	4	1

Solution

x	$\log x$	f	$f(\log x)$
2	0.3010	1	0.3010
4	0.6020	4	2.4080
8	0.9030	6	5.4180
16	1.2041	4	4.8164
32	1.5051	1	1.5051
			$\Sigma f(\log x) = 14.4485$

$$\log G = \frac{14.4485}{16} = 0.9030$$

$$GM = \text{Antilog } 0.9030$$

$$= 7.998 \quad \text{or} \quad 8 \text{ approx.}$$

Let us take another example involving group frequency distribution.

Example 6.29 Calculate the geometric mean from the following data:

Income Per Day	*Frequency*
40–60	14
60–80	54
80–100	26
100–120	4
120–140	2

Worksheet

Income (Rs)	*x* *Mid.V.*	*log x*	*f*	*f (log x)*
40–60	50	1.6990	14	23.7860
60–80	70	1.8451	54	99.6354
80–100	90	1.9542	26	50.8092
100–120	110	2.0414	4	8.1656
120–140	130	2.1139	2	4.2278
			$n = \Sigma f = 100$	$\Sigma f (\log x)$ $= 186.6240$

$$\log G = \frac{\sum f(\log x)}{n}$$

$$= \frac{186.6240}{100} = 1.866240$$

$$GM = \text{Antilog } 1.866240 = 73.49$$

Hence, geometric mean income = Rs 73.49.

Example 6.30 Cities A, B and C are equidistant from each other. A motorist travels from A to B at 30 km/h; from B to C at 40 km/h and from C to A at 50 km/h. Determine the average speed for the entire trip.

Solution A motorist travels from A to B at 30 km/h
from B to C at 40 km/h and
from C to A at 50 km/h

Here, we have to use harmonic mean to find out the average speed.

$$HM = \frac{3}{\frac{1}{30} + \frac{1}{40} + \frac{1}{50}}$$

$$= \frac{3}{\frac{20 + 15 + 12}{600}}$$

$$= \frac{3}{\frac{47}{600}}$$

$$= \frac{3 \times 600}{47} = \frac{1800}{47} = 38.3 \text{ km/h}$$

Example 6.31 If A, G and H be the arithmetic mean, geometric mean and harmonic mean respectively of the two positive numbers, then show by taking a hypothetical example that:

(i) $A \geq G \geq H$. When does the equality sign hold?

(ii) $G^2 = AH$.

Solution

(i) Suppose we take two positive numbers 4 and 16.

$$\text{Arithmetic mean} = \frac{4+16}{2} = \frac{20}{2} = 10$$

$$\text{GM} = \sqrt{4 \times 16} = \sqrt{64} = 8$$

$$\text{HM} = \frac{2}{\frac{1}{4}+\frac{1}{16}} = \frac{2}{\frac{(4+1)}{16}} = \frac{2}{\frac{5}{16}} = \frac{32}{5} = 6.4$$

This shows that A > G > H.

The equality sign holds only if all the numbers $X_1, X_2, \ldots, X_n$ are identical.

(ii) Taking the same figures 4 and 16

$$G^2 = (8)^2 = 64$$

$$\text{AH} = \text{Arith. Mean} \times \text{Harmonic Mean}$$
$$= 10 \times 6.4$$
$$= 64$$

Hence, $G^2 = \text{AH}$

GLOSSARY

Arithmetic mean	A measure of central tendency calculated by dividing the sum of all observations by the number of observations in the data set.
Bimodal distribution	A distribution that has two modes.
Deciles	Fractiles that divide the data into ten equal parts.
Geometric mean	A measure of central tendency computed by taking the n^{th} root of the product of n observations.
Harmonic mean	The reciprocal of the arithmetic mean of the reciprocals of individual observations.
Lower quartile or first quartile	The value in a ranked data set such that one-fourth of the measurements are below this value and three-fourths are above it.
Measures of central tendency	Measures that describe the centre of a distribution. The mean, median and mode are three measures of central tendency.
Median	The value of the middle item in a data set arranged in an ascending or a descending order. It divides the data set into two equal parts.
Mode	The value that has the maximum frequency in the data set.
Multimodal distribution	A distribution that has more than two modes.
Outliers	Values that are either very small or very large as compared to the majority of the values in a data set.
Percentiles	Fractiles that divide a ranked data set into hundred equal parts.
Quadratic mean	It is the square root of the arithmetic mean of the squares.
Quartiles	Fractiles that divide a ranked data set into four equal parts.

Second quartile	It is the same as the median that divides a ranked data set into two equal parts.
Unimodal	A distribution that has only one mode.
Upper quartile or third quartile	Third of the three quartiles that divide a ranked data set into four equal parts. About three-fourths of the values in a data set are smaller than the value of the third quartile and about one-fourth above it.
Weighted mean	An average in which each item in the data is weighted depending on its importance in the total series.

LIST OF FORMULAE

1. Population arithmetic mean of individual observations: $\mu = \frac{\Sigma X}{N}$

2. Sample arithmetic mean of individual observations: $\bar{x} = \frac{\Sigma x}{n}$

3. Sample arithmetic mean of a discrete series: $\bar{x} = \frac{\Sigma fx}{n}$

4. Sample arithmetic mean of grouped data: $\bar{x} = \frac{\Sigma fm}{n}$

where m stands for mid-points.

5. Sample arithmetic mean by the short-cut method: $\bar{x} = A + \frac{\Sigma fd}{n}$

where A stands for the arbitrary mean and d stands for the deviation from arbitrary mean.

6. Sample arithmetic mean by the step-deviation method : $\bar{x} = A + \frac{\Sigma fd'}{n} \times C$

where d' stands for the deviations divided by the common factor C, which is used to simplify calculations.

7. Weighted mean: $\bar{x}_w = \frac{\Sigma wx}{\Sigma w}$ where w stands for the weight.

8. Combined mean of two series: $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

This can be generalised for any number of series.

9. Geometric mean of individual observations: $GM = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

$$= \text{Antilog}\left(\frac{\log x_1 + \log x_2 + \ldots + \log x_n}{n}\right)$$

10. Geometric mean of a frequency distribution: $GM = \text{Antilog}\left(\frac{\Sigma f \log x}{n}\right)$, where x is mid-point

11. Harmonic mean for grouped data: $\text{HM} = \dfrac{n}{\Sigma\,(f \times 1/x)}$

12. Median in a data array: M = Size of $\left(\dfrac{n+1}{2}\right)^{\text{th}}$ item

where n is the number of items in the data array. Where the series consists of an even number of items, the median is the average of the two middle items.

13. Median in a grouped series: $M = l_1 + \dfrac{l_2 - l_1}{f}(m - c)$

where m is the size of the middle item, i.e., $[(n + 1) \div 2)]^{\text{th}}$ item, l_2 and l_1 are respectively the upper and lower limits of the class in which the median lies, f is the frequency of the class in which the median lies, and c is the cumulative frequency of the preceding class in which the median lies.

14. Lower quartile: $\text{Q}_1 = l_1 + \dfrac{l_2 - l_1}{f}\left(\dfrac{n+1}{4} - c\right)$

15. Upper quartile: $\text{Q}_3 = l_1 + \dfrac{l_2 - l_1}{f}\left(\dfrac{3(n+1)}{4} - c\right)$

16. Decile, say, 2nd decile: $\text{D}_2 = l_1 + \dfrac{l_2 - l_1}{f}\left(\dfrac{2(n+1)}{10} - c\right)$

17. Percentile, say, 10th percentile: $\text{P}_{10} = l_1 + \dfrac{l_2 - l_1}{f}\left(\dfrac{10(n+1)}{100} - c\right)$

18. Mode: $\text{Mo} = f_1 + \dfrac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$

where f_1 = frequency of the class in which the mode lies
f_0 = frequency of the class preceding the modal class
f_2 = frequency of the class succeeding the modal class
i = class-interval

19. Compound interest formula $P_n = P_0\,(1 + r)^n$

where P_n = value of investment at the end of the n^{th} year
P_0 = initial investment
r = annual rate of interest
n = number of years

QUESTIONS

6.1 Given below are twelve statements. Indicate in each case whether the statement is true or false:

(a) While calculating median, every individual item in the data is taken into consideration.
(b) The arithmetic mean tends to be the most repeated value in a series.
(c) Extreme values in a given series strongly influence the median.

(d) If a series is arranged in an ascending or descending order, the median tends to be the value of the middle item.
(e) The arithmetic mean of two positive numbers is always equal to or greater than their geometric mean.
(f) The geometric mean of two positive numbers is always less than their harmonic mean.
(g) Median is always equidistant between the upper quartile and the lower quartile.
(h) Geometric mean is the easiest to calculate amongst all the means.
(i) A mean calculated from grouped data is seldom exact.
(j) Mode is not influenced by extreme values in a series.
(k) It is possible to calculate for a frequency distribution with open ends.
(l) In case of a symmetrical distribution, the mean, median and mode are equal.

Multiple Choice Questions (6.2 to 6.13)

6.2 The sum of deviations taken from arithmetic mean is
(a) minimum (b) zero (c) maximum (d) none of the above

6.3 The sum of squares of deviations from arithmetic mean is
(a) zero (b) maximum (c) minimum (d) none of the above

6.4 When calculating the average growth of economy, the correct mean to use is
(a) weighted mean (b) geometric mean
(c) arithmetic mean (d) either (a) or (c)

6.5 When calculating mean from grouped data, an implict assumption is
(a) Each value in a class is equal to the mid-point.
(b) No value is repeated in the data set.
(c) All values are discrete.
(d) The number of the frequency in each class is the same.

6.6 When calculating median of a data set, the first step is
(a) Calculate the mean of the middle two items in the data set.
(b) Arrange the data in an ascending or descending order.
(c) Determine the relative weights of the data values in terms of importance.
(d) None of these.

6.7 Which of the following notations is a parameter?
(a) n (b) $\bar{x}$ (c) $\bar{p}$ (d) μ

6.8 The geometric mean of 4, 16 and 8 is
(a) 6 (b) 10 (c) 12 (d) 8

6.9 The relative position of arithmetic, geometric and harmonic mean is
(a) G.M. < H.M. < $\bar{x}$ (b) $\bar{x}$ < G.M. < H.M.
(c) $\bar{x}$ > H.M. > G.M. (d) $\bar{x}$ > G.M. > H.M.

6.10 Which of the following statements is *not* correct?
(a) Extreme values in the data affect the mean.
(b) Geometric mean is always less or equal to the mean.
(c) Some data sets do not have means.
(d) At times weighted mean is much better than the simple mean.

6.11 The geometric mean in the data set having a negative observation is
(a) zero (b) negative
(c) positive (d) cannot be calculated

6.12 For the given set of observations 7, 8, 9, 9 and 17
(a) mean is greater than median (b) mode is greater than mean
(c) median is greater than made (d) none of these

6.13 When an observation in the data set is zero, then its geometric mean is
(a) negative (b) zero
(c) positive (d) cannot be calculated.

6.14 What are the desiderata (requirements) of a good average? Compare the mean, the median and the mode in the light of these desiderata? Why are averages called measures of central tendency?

6.15 "Every average has its own peculiar characteristics. It is difficult to say which average is the best." Explain with examples.

6.16 What do you understand by 'Central Tendency'? Under what conditions is the median more suitable than other measures of central tendency?

6.17 Three teachers conducted a test in Statistics in their classes. The average marks obtained in their classes were 65, 59 and 76. There were 30, 42 and 28 students respectively in their classes. What were the average marks for all the classes?

6.18 The average monthly salary paid to all employees in a company was Rs 8,000. The average monthly salaries paid to male and female employees of the company were Rs 10,600 and Rs 7,500 respectively. Find out the percentages of males and females employed by the company.

6.19 Calculate the arithmetic mean from the following data:

Class	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
Frequency	2	4	9	11	12	6	4	2

6.20 A travelling salesman made five trips in two months. The record of sales is given below:

Trip	*No. of Days*	*Value of Sales (Rs)*	*Sales Per Day (Rs)*
1	5	3,000	600
2	4	1,600	400
3	3	1,500	500
4	7	3,500	500
5	6	4,200	700
Total	**25**	**13,800**	**2,700**

The sales manager criticised the salesman's performance as not very good, since his mean daily sales were only Rs 540 (2,700/5). The salesman called this an unfair statement, for his daily mean sales were as high as Rs 552 (13,800/25). What does each average here mean? Which mean seems to be more appropriate in this case?

6.21 A college management wanted to give scholarships to B.Sc. students securing 60 per cent and above marks in the following manner:

Percentage of Marks	*Monthly Scholarship in Rs*
60–65	100
65–70	150
70–75	200
75–80	250
80–85	300

The marks of 25 students, who were eligible for scholarship are given below:
74, 62, 84, 72, 61, 83, 72, 81, 64, 71, 63, 61, 60, 67, 74, 66, 64, 79, 75, 76, 69, 68, 78, 67 and 73.
Calculate the monthly scholarship paid to the students.

6.22 From the following table, calculate the average yield:

Yield in lbs	*Number of Plots*
Over 0	216
Over 60	210
Over 120	156
Over 180	98
Over 240	57
Over 300	31
Over 360	13
Over 420	7
Over 480	2
Up to 540	216

6.23 Find the mean, median and mode for the following set of numbers:
(a) 3, 5, 2, 6, 5, 9, 5, 2, 8 and 6
(b) 51.6, 48.7, 50.3, 49.5 and 48.9

6.24 Given the following data, calculate the arithmetic mean by the 'short-cut' method:

X	462	480	498	516	534	552	570	588	606	624
f	98	75	56	42	30	21	15	11	6	2

6.25 Calculate the mean, median and mode from the following data:

Height in Inches	*Number of Persons*
62–63	2
63–64	6
64–65	14
65–66	16
66–67	8
67–68	3
68–69	1
Total	**50**

6.26 The distribution of weights of 152 students is as follows:

Weight in kg	*Frequency*
30–40	18
40–50	38
50–60	46
60–70	27
70–80	15
80–90	8

Calculate Q_1, median, D_7 and P_{80}.

6.27 The net profits earned by 100 companies for one year are as under:

Net Profit (Rs in lakh)	*Number of Companies*
20–30	5
30–40	7
40–50	19
50–60	29
60–70	16
70–80	9
80–90	8
90–100	7

Calculate Q_1, median, D_4 and P_{60} and interpret their values.

6.28 The following data pertain to marks obtained by 120 students in their final examination in mathematics:

Marks	*Number of Students*
90–100	9
80–89	32
70–79	43
60–69	21
50–59	11
40–49	3
30–39	1
Total	**120**

Calculate the lower and upper quartiles as also the median.

6.29 During one year the ratio of milk prices per litre to bread prices per loaf was 3.00, whereas during the next year the ratio was 2.00. (a) Find the arithmetic mean of these ratios for the two-year period. (b) Find the arithmetic mean of the ratios of bread prices to milk prices for the two-year period. (c) Discuss the advisability of using the arithmetic mean for averaging ratios. (d) Discuss the suitability of the geometric mean for averaging ratios.

6.30 Calculate the geometric mean of (a) 4, 8 and 16; (b) 12, 16 and 20.

6.31 The population of a town was 10,000 in 1990. It was 20,000 in the year 2000. (i) What is the rate of growth? (ii) If the present rate of growth continues, what is likely to be the population in 2005?

6.32 The rates of increase in population of a country during the last three decades are 10 per cent, 20 per cent and 30 per cent. Find the average rate of growth during the last three decades.

6.33 A sum of Rs 20,000 is invested at 10 per cent annual rate of interest. What will be the total amount after eight years if the original principal is not withdrawn?

6.34 The geometric mean of two numbers is 18. If, by mistake, one figure is taken as 12, instead of 21, find the correct geometric mean.

6.35 If the price of a commodity doubles in a period of six years, what is the average percentage increase per year?

6.36 A man travels from A to B at an average speed of 30 km/h and returns from B to A along the same route at an average speed of 60 km/h. Find the average speed for the entire trip.

6.37 A vehicle when climbing up a gradient, consumes petrol @ 8 km per litre. While coming down it runs 12 km per litre. Find its average consumption for to and from travel between two places situated at the two ends of 25 km long gradient.

6.38 Give a specific example of your own in which:
(i) the arithmetic mean would be more suitable instead of the mode and the median.
(ii) the median would be preferred instead of the mode and vice versa.

6.39 For a pair of observations, show that
$AM \times HM = (GM)^2$, where AM, GM and HM stand respectively for arithmetic, geometric and harmonic means of the observations.

6.40 The frequency distribution of weight in grams of mangoes of a given variety is as given below:

Weight in grams	*Number of mangoes*
410–419	14
420–429	20
430–439	42
440–449	54
450–459	45
460–469	18
470–479	7

Calculate the arithmetic mean, median and mode.

6.41 The following table gives the number of firms as well as the number of workers in various income groups in a certain industry. Calculate the average salary of the workers.

Income groups (Rs)	*No. of firms*	*No. of workers*
600–900	18	50
900–1200	22	46
1200–1600	16	40
1600–2000	12	44
2000–2500	10	42
2500–3000	9	36
3000–3500	9	30
3500–4000	4	12

6.42 A professor, teaching in a business school has decided to award the final grades (marks), based on attendance, term paper and mid-term test. The data are given below:

Name of Student	*Attendance*	*Term Paper*	*Mid-term Test*
A	22	25	42
B	20	23	40
C	19	16	37
D	24	15	25
E	25	18	30
F	18	20	32
G	16	23	45
H	21	19	38
I	25	15	30
J	23	17	27

He has assigned 20% for attendance, 30% for term paper and 50% for the mid-term test. Calculate the final marks for each of the ten students, and the average final marks, taking the group as a whole.

The McGraw-Hill Companies

CHAPTER 7

MEASURES OF DISPERSION

Learning Objectives

By the end of your work on this chapter, you should be able to

- differentiate between average and dispersion
- calculate different measures of dispersion—the range, quartile deviation, mean deviation, standard deviation and the coefficient of variation
- select the most appropriate measure of dispersion for a given set of data, and justify your choice.

Chapter Prerequisites

Before starting work on this chapter, you should ensure that you are conversant with the topics covered in the chapter on Measures of Central Tendency.

7.1 INTRODUCTION

In the preceding chapter, we have seen different types of means and learnt how they can be calculated in varying types of distributions. The means are just the measures of central tendency and do not indicate the extent of variability in a distribution. The main theme of this chapter is *dispersion* or *variability*, which provides us one more step in increasing our understanding of the pattern of the data.

Let us explain the concept of dispersion with the help of a diagram. Figure 7.1 shows two curves with different means and same variability. In contrast, Figure 7.2 shows two curves with different variabilities and same mean. Curve A has least variability as most of its data are closely centred around the mean. Curve B has wider variability than curve A. In fact, both the concepts—mean and dispersion—are necessary for studying a variety of problems in Business Statistics.

7.2 IMPORTANCE OF DISPERSION

A study of dispersion/variability enables us to get additional information about the composition of data. Confining merely to mean will not provide us this vital information.

A point worth noting is that a high degree of uniformity (i.e. low degree of dispersion) is a desirable quality. Many problems emerge from wide variability in data. Unless we know the extent of variability in the data, we would not be able to handle the problem properly.

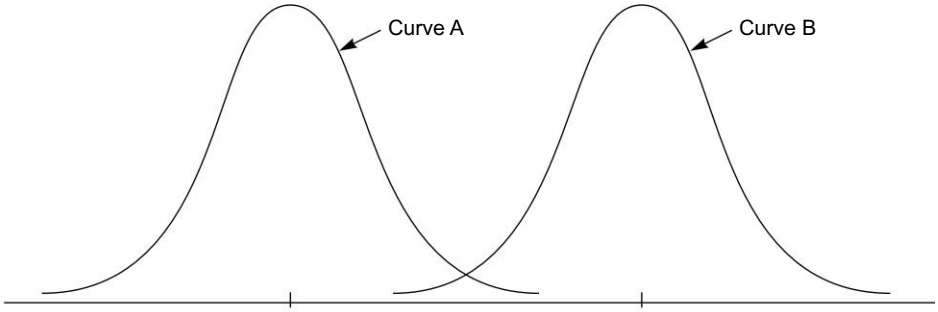


Fig. 7.1 **Curves A and B with Different Means and Same Variability**

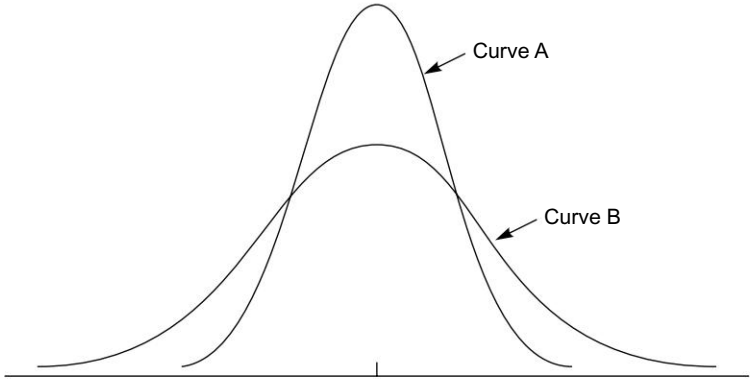


Fig. 7.2 **Curves A and B with Equal Means and Different Variability**

We give below a few examples where the study of dispersion would be most appropriate.

1. Suppose, of the two applicants, a firm is planning to appoint one worker. It may assign certain work to each of them. Though the average output of both the workers is the same, the output of the first worker is more stable as compared to the second worker, whose output shows considerable fluctuation. Therefore, the firm should appoint the first worker, as his output has less variability.
2. Suppose, an investor is looking for a suitable equity share for investment. While examining the movement of share prices, he should avoid shares with highly fluctuating price, that is, at times very high and at other times very low. Such extreme fluctuations mean that these shares involve high risk. The investor should, therefore, prefer those shares where risk is not so high.
3. In the sphere of quality control, dispersion, i.e., variability, is more relevant than average measurement. A manufacturing company may first fix a standard for a prospective product and then ascertain as to the extent to which it has deviated from the previously-set standard. This would enable the company to take necessary steps to ensure that the output is in conformity with the laid down standard.
4. As will be seen in subsequent chapters, various analytical tools are used in statistics. For example, a number of hypotheses can be used for studying variability. By using such analytical tools, one can go into the depth of the given problem, which will not be possible by using average.

There are four measures of dispersion:

1. The range
2. The interquartile range or the quartile deviation
3. The mean deviation
4. The standard deviation

All these are mathematical models. We shall discuss each of these methods, giving suitable examples.

7.3 THE RANGE

The simplest measure of dispersion is the range, which is the difference between the maximum value and the minimum value of data.

Example 7.1 Find the range for the following three sets of data:

Set 1	15	15	15	15	15	15	15	15	15	5
Set 2	8	7	15	11	12	5	13	11	15	9
Set 3	5	5	5	5	5	15	15	15	15	15

Solution In each of these three sets, the highest number is 15 and the lowest number is 5. Since the range is the difference between the maximum value and the minimum value of the data, it is 10 in each case. But the range fails to give any idea about the dispersal or spread of the series between the highest and the lowest value. This becomes evident from the above data.

In a frequency distribution, range is calculated by taking the difference between the upper limit of the highest class and the lower limit of the lowest class.

Example 7.2 Find the range for the following frequency distribution:

Size of Item	*Frequency*
20–40	7
40–60	11
60–80	30
80–100	17
100–120	5
Total	**70**

Solution Here, the upper limit of the highest class is 120 and the lower limit of the lowest class is 20. Hence, the range is 120 – 20 = 100. Note that the range is not influenced by the frequencies. Symbolically, the range is calculated by the formula $L–S$, where L is the largest value and S is the smallest value in a distribution. The coefficient of range is calculated by the formula:

$$\frac{L-S}{L+S}$$

This is the relative measure. The coefficient of the range in respect of the earlier example having three sets of data is

$$\frac{L-S}{L+S}=\frac{15-5}{15+5}=\frac{10}{20}=0.5$$

The coefficient of range is more appropriate for purposes of comparison as will be evident from the following example.

Example 7.3 Calculate the coefficient of range separately for the two sets of data given below:

Set 1	8	10	20	9	15	10	13	28
Set 2	30	35	42	50	32	49	39	33

Solution It can be seen that the range in both the sets of data is the same:

Set 1 $28 - 8 = 20$

Set 2 $50 - 30 = 20$

Coefficient of range in Set 1 is: $\frac{28-8}{28+8} = \frac{20}{36} = 0.55$

Coefficient of range in Set 2 is: $\frac{50-30}{50+30} = \frac{20}{80} = 0.25$

Limitations

There are some major *limitations* of range:

1. It is based only on two items and does not cover all the items in a distribution.
2. It is subject to wide fluctuations from sample to sample based on the same population.
3. It fails to give any idea about the pattern of distribution. This was evident from the data given in Examples 7.1 and 7.3.
4. Finally, in the case of open-ended distributions, it is not possible to compute the range.

Despite these limitations of the range, it has certain advantages also.

Advantages

1. It is mainly used in situations where one wants to quickly have some idea of the variability of a set of data.
2. When the sample size is very small, the range is considered quite adequate measure of the variability. Thus, it is widely used in quality control where a continuous check on the variability of raw materials or finished products is needed.
3. The range is also a suitable measure in weather forecast. The meteorological department uses the range by giving the maximum and the minimum temperatures. This information is quite useful to the common man as he can know the extent of possible variation in the temperature on a particular day.

7.4 THE INTERQUARTILE RANGE OR THE QUARTILE DEVIATION

The interquartile range or the quartile deviation is a better measure of variation in a distribution than the range. Here, the middle 50 per cent of the distribution is used by avoiding the 25 per cent of the distribution at both the ends. In other words, the interquartile range denotes the difference between the third quartile and the first quartile.

Symbolically, interquartile range = $Q_3 - Q_1$

Many times the interquartile range is reduced in the form of semi-interquartile range or quartile deviation as shown below:

Semi-interquartile range or Quartile deviation = $\frac{Q_3 - Q_1}{2}$.

When quartile deviation is small, it means that there is a small deviation in the central 50 per cent items. In contrast, if the quartile deviation is high, it shows that the central 50 per cent items have a large variation. It may be noted that in a symmetrical distribution, the two quartiles, that is, Q_3 and Q_1 are equidistant from the median. Symbolically,

$$M - Q_1 = Q_3 - M$$

However, this is seldom the case as most of the business and economic data are asymmetrical. But, one can assume that approximately 50 per cent of the observations are contained in the interquartile range.

It may be noted that interquartile range or the quartile deviation is an absolute measure of dispersion. It can be changed into a relative measure of dispersion as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The computation of a quartile deviation is very simple, involving the computation of upper and lower quartiles.

Example 7.4 Let us take an example to calculate the quartile deviation. Example 6.13 in the preceding chapter gives the following frequency distribution.

Class-interval	*Frequency*	*Cumulative frequency*
30–40	4	4
40–50	6	10
50–60	8	18
60–70	12	30
70–80	9	39
80–90	7	46
90–100	4	50

In order to calculate quartiles, Q_1 and Q_3, it is necessary that we should use the cumulative frequencies. These are shown in the above table.

First, we have to find out the class in which Q_1 lies. For this, we use the formula:

$$Q_1 = \text{Size of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$

$$= \frac{50+1}{4} = 12.75$$

This figure lies in 50–60 class. Now we apply the formula:

$$Q_1 = l_1 + \frac{l_2 - l_1}{f}(m - c)$$

$$= 50 + \frac{60-50}{8}\ (12.75 - 10)$$

$$= 50 + \frac{10}{8} \times 2.75$$

$$= 50 + 3.4375$$

$$= 53.44 \text{ approx.}$$

$$Q_3 = \text{Size of } \frac{3(n+1)}{4} \text{ item}$$

$$= \frac{3(50+1)}{4} = 38.25^{\text{th}} \text{ item}$$

$$Q_3 = l_1 + \frac{l_2 - l_1}{f}\ (\text{m} - \text{c})$$

$$= 70 + \frac{80-70}{9}\ (38.25 - 30)$$

$$= 70 + \frac{10}{9} \times 8.25$$

$$= 70 + 9.17 = 79.17$$

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{79.17 - 53.44}{2}$$

$$= \frac{25.73}{2}$$

$$= 12.865$$

We can also calculate the coefficient of QD.

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{79.17 - 53.44}{79.17 + 53.44}$$

$$= \frac{25.73}{132.61}$$

$$= 0.19$$

Merits of Quartile Deviation

1. As compared to range, it is considered a superior measure of dispersion.
2. In the case of open-ended distribution, it is quite suitable.

3. Since it is not influenced by the extreme values in a distribution, it is particularly suitable in highly skewed or erratic distributions.

Limitations of Quartile Deviation

1. Like the range, it fails to cover all the items in a distribution.
2. It is not amenable to mathematical manipulation.
3. It varies widely from sample to sample based on the same population.
4. Since it is a positional average, it is not considered as a measure of dispersion. It merely shows a distance on scale and not a scatter around an average.

In view of the above-mentioned limitations, the interquartile range or the quartile deviation has a limited practical utility.

7.5 INTERFRACTILE RANGE

In the preceding chapter, we have seen that the median divides a frequency distribution into two halves. This can be said in another way: the median is the 0.5 fractile, which suggests that half of the given data is below or equal to this value.

The interfractile range indicates the spread of the given data between two fractiles in a frequency distribution. It may be noted that different names are given to fractiles on the basis of the fraction or division of the data they measure. For example, when the data are divided into four equal parts, the parts are called quartiles. Quartile 1 is one-fourth part of the given data, Quartile 2 happens to be median and Quartile 3 shows the three-fourth part of the given data. Likewise, when the given data are divided into ten equal parts, then these are known as deciles. If there are hundred equal parts of the given data, then they are called percentiles.

While the fractiles can be used to measure any equal fractions or proportions of the given data, quartiles, deciles and percentiles are commonly used.

Suppose we are given the following data:

10, 12, 8, 7, 15, 16, 20, 13, 11, 18, 22, 26

We are asked to find the 50^{th} and the 70^{th} percentiles.

First, we have to arrange the data either in an ascending or descending order. Let us take the ascending order:

7, 8, 10, 11, 12, 13, 15, 16, 18, 19, 22, 26

For P_{50}, the formula would be

$$P_{50} = \frac{(n+1)P}{100}$$

$$= \frac{(12+1)50}{100} = \text{size of } 6.5^{th} \text{ item}$$

This means we have to take the average of the 6^{th} and 7^{th} items:

$$\frac{13+15}{2} = 14$$

$$P_{70} = \frac{(12+1)\,70}{100} = \frac{910}{100} = 9.1$$

Here 9th item is 18 and 10th item is 19. Hence, the 70th percentile is a point lying 0.1 of the way from 18 to 19, that is 18.1.

Example 7.5 Decile divides the data into ten equal parts. Let us take an example.

The following observations are given and we are asked to find the value of the 3rd and the 8th deciles.

10, 15, 16, 5, 18, 11, 9, 25, 19, 23, 17, 13, 22, 8, 26, 7, 12, 24, 28

Solution We have to first arrange these figures in the ascending order.

5, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 22, 23, 24, 25, 26, 28

$$3^{rd}\ \text{Decile} = \text{Size of}\ \frac{3(n+1)}{10} = \frac{3(19+1)}{10} = \frac{60}{10} = 6$$

This means the size of the 6th item, which happens to be 11.

$$8^{th}\ \text{Decile} = \text{Size of}\ \frac{8(n+1)}{10} = \frac{8\times 20}{10} = \frac{160}{10} = 16$$

that is, size of 16th item happens to be 24.

Hence, the 8th decile is 24.

It may be noted that the 5th decile happens to be the median as it divides the series into two equal parts.

We have considered individual observations in these examples. We can calculate deciles and percentiles in case of a frequency distribution, as we do in case of quartiles and median. We have to interpolate the actual value of the decile and the percentile after ascertaining the size of the item.

7.6 THE MEAN DEVIATION

The mean deviation is also known as the average deviation. As the name implies, it is the average of absolute amounts by which the individual items deviate from the mean. Since the positive deviations from the mean are equal to the negative deviations, while computing the mean deviation, we ignore positive and negative signs. Symbolically,

$$\text{MD} = \frac{\Sigma |x|}{n}$$

where MD = mean deviation

$|x|$ = deviation of an item from the mean,* ignoring positive and negative signs

n = the total number of observations.

* Occasionally, deviations are taken from the median.

Let us take an example:

Example 7.6

Size of Item	*Frequency*
2–4	20
4–6	40
6–8	30
8–10	10

Solution We set up the worksheet for calculating the mean deviation.

Size of Item	*Mid-points (m)*	*Frequency (f)*	*mf*	*d from* $\bar{x}$	$f\|d\|$
2–4	3	20	60	– 2.6	52
4–6	5	40	200	– 0.6	24
6–8	7	30	210	1.4	42
8–10	9	10	90	3.4	34
	Total	**100**	**560**		**152**

$$\bar{x} = \frac{\Sigma fm}{n} = \frac{560}{100} = 5.6$$

$$\text{MD } (\bar{x}) = \frac{\Sigma f \mid d \mid}{n} = \frac{152}{100} = 1.52$$

Merits of Mean Deviation

1. A major advantage of mean deviation is that it is simple to understand and easy to calculate.
2. It takes into consideration each and every item in the distribution. As a result, a change in the value of any item will have its effect on the magnitude of mean deviation.
3. The values of extreme items have less effect on the value of the mean deviation.
4. As deviations are taken from a central value, it is possible to have meaningful comparisons of the formation of different distributions.

Limitations of Mean Deviation

1. It is not capable of further algebraic treatment.
2. At times it may fail to give accurate results. The mean deviation gives best results when deviations are taken from the median instead of from the mean. But in a series, which has wide variations in the items, median is not a satisfactory measure.
3. Strictly on mathematical considerations, the method is wrong as it ignores the algebraic signs when the deviations are taken from the mean.

In view of these limitations, it is seldom used in business studies. A better measure known as the standard deviation is more frequently used.

7.7 THE STANDARD DEVIATION

The fourth method of dispersion to be considered is the standard deviation. It is similar to the mean deviation in that here too the deviations are measured from the mean. At the same time, the standard deviation is preferred to the mean deviation or the quartile deviation or the range because it has desirable mathematical properties.

Before defining the concept of the standard deviation, we introduce another concept, viz., variance. Consider the individual items given in the following table:

Example 7.7

x	$x - \mu$	$(x - \mu)^2$
20	20 – 18 = 2	4
15	15 – 18 = –3	9
19	19 – 18 = 1	1
24	24 – 18 = 6	36
16	16 – 18 = –2	4
14	14 – 18 = –4	16
108	**Total**	**70**

$$\text{Mean} = \frac{108}{6} = 18$$

The second column shows the deviations from the mean. The third or the last column shows the squared deviations, the sum of which is 70. The arithmetic mean of the squared deviations is:

$$\frac{\Sigma(x-\mu)^2}{N} = \frac{70}{6} = 11.67 \text{ approx.}$$

This mean of the squared deviations is known as the variance. It may be noted that this variance is described by different terms that are used interchangeably: the variance of the distribution X; the variance of X; the variance of the distribution; and just simply, the variance.

Symbolically, $$\text{Var}(X) = \frac{\Sigma(x-\mu)^2}{N}$$

It is also written as $$\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{N}$$

where σ^2 (called sigma squared) is used to denote the variance.

Although the variance is a measure of dispersion, the unit of its measurement is $(\text{points})^2$. If a distribution relates to income of families, then the variance is $(\text{Rs})^2$ and not rupees. Similarly, if another distribution pertains to marks of students, then the unit of variance is $(\text{marks})^2$. To overcome this inadequacy, the square root of variance is taken, which yields a better measure of dispersion known as the standard deviation.

Taking our earlier example of individual observations, we take the positive square root of the positive variance

$$\text{SD or } \sigma = \sqrt{\text{variance}} = \sqrt{11.67} = 3.42 \text{ points}$$

Symbolically,

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

In applied Statistics, the standard deviation is more frequently used than the variance. This can also be written as:

$$\sigma = \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N}}{N}}$$

We use this formula to calculate the standard deviation from the individual observations given earlier.

Example 7.8

x	x^2
20	400
15	225
19	361
24	576
16	256
14	196
108	**2014**

$\Sigma x_i^2 = 2014 \quad \Sigma x_i = 108 \quad N = 6$

$$\sigma = \sqrt{\frac{2014 - \frac{(108)^2}{6}}{6}} \qquad \sigma = \sqrt{\frac{2014 - \frac{11664}{6}}{6}}$$

$$\sigma = \sqrt{\frac{12084 - 11664}{6}} \qquad \sigma = \sqrt{\frac{\frac{420}{6}}{6}}$$

$$\sigma = \sqrt{\frac{70}{6}} \qquad \sigma = \sqrt{11.67}$$

$$\sigma = \sqrt{11.67}$$

The figure of $\sqrt{11.67}$ is the same as arrived earlier.

Example 7.9 ***Grouped Data***

Let us take the following distribution relating to marks obtained by students in an examination:

Marks	*Number of Students*
1–10	1

(*Contd.*)

(*Contd.*)

20–30	6
30–40	10
40–50	12
50–60	11
60–70	6
70–80	3
80–90	2
90–100	1

Worksheet

Marks	*Frequency (f)*	*Mid-points*	*Deviations (d) ÷ 10 = d'*	*fd'*	*fd'²*
0 – 10	1	5	–5	–5	25
10 – 20	3	15	– 4	–12	48
20 – 30	6	25	–3	–18	54
30 – 40	10	35	–2	–20	40
40 – 50	12	45	–1	–12	12
50 – 60	11	55	0	0	0
60 – 70	6	65	1	6	6
70 – 80	3	75	2	6	12
80 – 90	2	85	3	6	18
90 –100	1	95	4	4	16
Total	**55**		**Total**	**– 45**	**231**

In the case of frequency distribution where the individual values are not known, we use the mid-points of the class-intervals. Thus, the formula used for calculating the standard deviation is as given below:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{K} f_i \, (m_i - \mu)^2}{N}}$$

where m_i is the mid-point of the class-intervals; μ is the mean of the distribution; f_i is the frequency of each class; N is the total number of frequency and K is the number of classes. This formula requires that the mean μ be calculated and that deviations $(m_i - \mu)$ be obtained for each class. To avoid this inconvenience, the above formula can be modified as:

$$\sigma = C\sqrt{\frac{\sum_{i=1}^{K} f_i \, d_i^2}{N} - \left(\frac{\sum_{i=1}^{K} fd_i}{N}\right)^2}$$

where C is the width of the class-interval; f_i is the frequency of the i^{th} class and d_i is the deviation of the i^{th} item from an assumed origin; and N is the total number of observations.

Applying this formula for the table given earlier,

$$\sigma = 10\sqrt{\frac{231}{55} - \left(\frac{-45}{55}\right)^2}$$

$$= 10\sqrt{4.2 - 0.669421}$$

$$= 18.8 \text{ marks}$$

When it becomes clear that the actual mean would turn out to be in fraction, calculating deviations from the mean would be too cumbersome. In such cases, an assumed mean is used and the deviations from it are calculated. While mid-point of any class can be taken as an assumed mean, it is advisable to choose the mid-point of that class that would make calculations least cumbersome. Guided by this consideration, in Example 7.9 we have decided to choose 55 as the mid-point and, accordingly, deviations have been taken from it. It will be seen from the calculations that they are considerably simplified.

Combined Standard Deviation

If two groups contain n_1 and n_2 observations with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2 respectively, then the standard deviation (σ) of the combined group is given by

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where σ_{12} *is the combined standard deviation of the two groups,* $d_1 = \bar{x}_{12} - \bar{x}_1$, $d_2 = \bar{x}_{12} - \bar{x}_2$ *and* $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

The formula given above can be extended to find out the standard deviation of three or more groups. Let us take an example.

Example 7.10 The mean and the standard deviation of two distributions of 100 and 150 items are 50 and 5, and 40 and 6, respectively. Find the mean and the standard deviation of all the 250 items taken together.

Solution Combined mean of 250 items

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{(100 \times 50) + (150 \times 40)}{100 + 150}$$

$$= \frac{5000 + 6000}{250} = \frac{11000}{250} = 44$$

Combined standard deviation of 250 items

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

On the basis of the data given in the problem,

$$d_1 = \bar{x}_{12} - \bar{x}_1 = 44 - 50$$

$$d_2 = \bar{x}_{12} - \bar{x}_2 = 44 - 40$$

Applying the values in the above formula:

$$\sigma_{12} = \sqrt{\frac{\left[100\left((5)^2 + (44-50)^2\right)\right] + \left[150\left((6)^2 + (44-40)^2\right)\right]}{100+150}}$$

$$= \sqrt{\frac{100(25+36) + 150(36+16)}{250}}$$

$$= \sqrt{\frac{6100 + 7800}{250}}$$

$$= \sqrt{\frac{13900}{250}} = \sqrt{55.6} = 7.46$$

Let us take another example.

Example 7.11 Given the set of numbers 2, 5, 8, 11, 14 and 2, 8, 14, find **(a)** the mean of each set, **(b)** the variance of each set, **(c)** the mean of the combined or 'pooled' sets, and **(d)** the variance of combined or pooled sets.

Solution

	1^{st} set			2^{nd} set	
	d	d^2		d	d^2
2	–6	36	2	–6	36
5	–3	9	8	0	0
8	0	0	14	6	36
11	3	9			
14	6	36			
40		$\Sigma d^2 = 90$	24		72

$\bar{x}$ = 40/5 = 8 $\bar{x}$ = 24/3 = 8

Variance of 1st set = $\frac{\sum d^2}{n} = \frac{90}{5} = 18$

Variance of 2nd set = $\frac{\sum d^2}{n} = \frac{72}{3} = 24$

Mean of combined sets

$$\frac{n_1 x_1 + n_2 x_2}{n_1 + n_2} = \frac{(8 \times 5) + (3 \times 8)}{5+3}$$

$$= \frac{40+24}{8} = \frac{64}{8} = 8$$

Variance of combined sets

$$\sigma_{12}^2 = n_1 \sigma_1^2$$

$$\sigma_{12}^2 = \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}$$

$$= \frac{5\left[18 + (8-8)^2\right] + 3\left[24 + (8-8)^2\right]}{5+3}$$

$$= \frac{90+72}{8}$$

$$= \frac{162}{8} = 20.25$$

Uses of the Standard Deviation

The standard deviation is a frequently used measure of dispersion. It enables us to determine as to how far individual items in a distribution deviate from its mean. In a symmetrical, bell-shaped curve such as the one given below:

(i) About 68 per cent of the values in the population fall within ± 1 standard deviation from the mean.
(ii) About 95 per cent of the values will fall within ± 2 standard deviations from the mean.
(iii) About 99 per cent of the values will fall within ± 3 standard deviations from the mean.

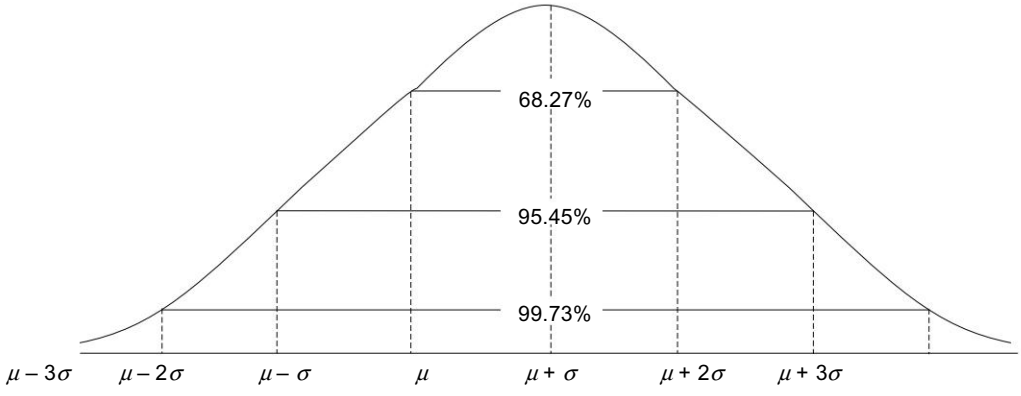


Fig. 7.3 **A Symmetrical Curve**

Two additional examples are given below:

Example 7.12 Given the weights of five persons:
120, 140, 150, 160 and 180 lbs

(a) Find the variance.
(b) Find the standard deviation.
(c) Subtract 30 lbs from each of the weights given above, and then calculate the variance and the standard deviation.

Solution

x	$x - \bar{x}$	$(x - \bar{x})^2$
120	–30	900
140	–10	100
150	0	0
160	10	100
180	30	900
	Total	**2,000**

$$\bar{x} = \frac{120 + 140 + 150 + 160 + 180}{5}$$

$$= \frac{750}{5} = 150$$

$$\text{Var}(X) = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

$$= \frac{2,000}{5} = 400$$

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{400} = 20$$

After subtracting 30 from each of the five figures, we have the series as follows:

$X - 30 = X'$	$X' - \bar{x}$	$(X' - \bar{x})^2$
90	–30	900
110	–10	100
120	0	0
130	10	100
150	30	900
	Total	**2,000**

$$\bar{x} = \frac{90 + 110 + 120 + 130 + 150}{5}$$

$$= \frac{600}{5} = 120$$

$$\text{Var}(X') = \frac{2,000}{5} = 400$$

$$\sigma = \sqrt{\text{Var}(X')} = \sqrt{400} = 20$$

We find that both the variance and the standard deviation are the same as obtained earlier. This is because the reduced figures show the same summation of deviations in the numerator as earlier; as such, the resultant figure remains the same.

Example 7.13 Given the following data, find out the combined mean and the standard deviation:

	Series A	*Series B*
No. of observations	100	500
Mean	50	60
Standard deviation	10	11

Solution

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

where n_1 and n_2 are number of observations in series A and B, respectively and $\bar{x}_1$ and $\bar{x}_2$ are means of series A and B, respectively.

$$= \frac{(100 \times 50) + (500 \times 60)}{100 + 500}$$

$$= 58.3 \text{ approx.}$$

$$d_1 = 50 - 58.3 = -8.3$$

$$d_2 = 60 - 58.3 = 1.7$$

The formula for the combined standard deviation is:

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$$\sigma_{12} = \sqrt{\frac{100\,[(10)^2 + (-8.3)^2] + 500\,[(11)^2 + (1.7)^2]}{100 + 500}}$$

$$\sigma_{12} = \sqrt{\frac{100\,(100 + 68.89) + 500\,(121 + 2.89)}{600}}$$

$$\sigma_{12} = \sqrt{\frac{16889 + 61945}{600}}$$

$$= 11.46 \text{ approx.}$$

7.8 RELATIVE DISPERSION: THE COEFFICIENT OF VARIATION

The standard deviation is an absolute measure of dispersion as it measures variation in the same units as the original data. As such, it cannot be a suitable measure while comparing two or more distributions. For this purpose, we should use a relative measure of dispersion. One such measure of relative dispersion is the coefficient of variation, which relates the standard deviation and the mean such that the standard deviation is expressed as a percentage of mean. Thus, the specific unit in which the standard deviation is measured is done away with and the new unit becomes per cent.

Symbolically, CV (coefficient of variation) = $\frac{\sigma}{\mu} \times 100\%$

Let us take an example.

Example 7.14 In a small business firm, two typists are employed—typist A and typist B. Typist A types out, on an average, 30 pages per day with a standard deviation of 6. Typist B, on an average, types out 45 pages with a standard deviation of 10. Which typist shows greater consistency in his output?

Solution

$$\text{Coefficient of variation for A} = \frac{\sigma}{\mu} \times 100\%$$

$$= \frac{6}{30} \times 100 = 20 \text{ per cent}$$

$$\text{Coefficient of variation for B} = \frac{\sigma}{\mu} \times 100\%$$

$$= \frac{10}{45} \times 100 = 22.2 \text{ per cent}$$

These calculations clearly indicate that although typist B types out more pages, there is a greater variation in his output as compared to that of typist A. We can say this in a different way: Though typist A's daily output is much less, he is more consistent than typist B.

The usefulness of the coefficient of variation becomes clear in comparing two groups of data having different means as has been the case in the above example.

7.9 STANDARDISED VARIABLE, STANDARD SCORES

The variable $Z = (x - \bar{x})/s$ or $(x - \mu)/\sigma$, which measures the deviation from the mean in units of the standard deviation, is called a standardised variable. Since both the numerator and the denominator are in the same units, a standardised variable is independent of units used.

If deviations from the mean are given in units of the standard deviation, they are said to be expressed in standard units or standard scores. Through this concept of standardised variable, proper comparisons can be made between individual observations belonging to two different distributions whose compositions differ. This will be clear from Example 7.15.

Example 7.15 A student has scored 68 marks in Statistics for which the average marks were 60 and the standard deviation was 10. In the paper on Marketing, he scored 74 marks for which the average marks were 68 and the standard deviation was 15. In which paper, Statistics or Marketing, was his relative standing higher?

Solution The standardised variable $Z = (x - \bar{x}) \div s$ measures the deviation of x from the mean in terms of standard deviation s. For Statistics, $Z = (68 - 60) \div 10 = 0.8$

For Marketing, $Z = (74 - 68) \div 15 = 0.4$

Since the standard score is 0.8 in Statistics as compared to 0.4 in Marketing, his relative standing was higher in Statistics.

Example 7.16 Convert the set of numbers 6, 7, 5, 10 and 12 into standard scores:

Solution

X	X^2
6	36
7	49
5	25
10	100
12	144
$\Sigma X = 40$	$\Sigma X^2 = 354$

$$\bar{x} = \Sigma x \div 5 = 40 \div 5 = 8$$

$$\sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma X)^2}{N}}{N}} \qquad \sigma = \sqrt{\frac{354 - \frac{(40)^2}{5}}{5}}$$

$$= \sqrt{\frac{354 - 320}{5}} \qquad = 2.61 \text{ approx.}$$

$$Z = \frac{x - \bar{x}}{\sigma} = \frac{6 - 8}{2.61} = -0.77 \text{ (Standard score)}$$

Applying this formula to other values:

(i) $\frac{7 - 8}{2.61} = -0.38$

(ii) $\frac{5 - 8}{2.61} = -1.15$

(iii) $\frac{10 - 8}{2.61} = 0.77$

(iv) $\frac{12 - 8}{2.61} = 1.53$

Thus, the standard scores for 6, 7, 5, 10 and 12 are –0.77, –0.38, –1.15, 0.77 and 1.53, respectively.

Additional Examples

Example 7.17 For a frequency distribution of marks in Econometrics of 200 candidates (grouped in intervals of 0–5, 5–10, … etc.), the mean and the standard deviation were found to be 40 and 15, respectively. Subsequently, it was found that the score 43 was wrongly taken as 53 in obtaining the frequency distribution. What should be the corrected mean and the standard deviation?

Solution

Given: $n = 200$

Class-intervals 0–5, 5–10 ….

Mean 40 SD 15

Instead of score 43, it was taken as 53. This means that the total of marks column should be less by 10. As a result the corrected mean would be

$$\frac{(40 \times 200) - 10}{200} = \frac{7990}{200} = 39.95$$

Similarly, corrected SD would be

$\sigma = \sqrt{\frac{\sum d^2}{n}}$. Since it is class-interval series

$\sigma = \sqrt{\frac{\sum fd^2}{n}}$. This means the revised Σfd^2 will be $(-10)^2$ i.e., 100 less than the earlier one.

$$\sqrt{\frac{100}{200}} = \sqrt{0.5} = 0.71$$

Hence, corrected SD is 15 – 0.71 = 14.29

Example 7.18 Find the missing information in the following table:

	Group			
	A	*B*	*C*	*Combined*
Number	50	—	90	200
Standard deviation	6	7	—	7.746
Mean	113	—	115	116

Solution There are three missing informations in the problem, they are to be found.
Combined number 200 – 50 – 90 = 60
Hence, the number for group B is 60.

Mean of group B is not given while combined mean is 116. Assuming x as the mean of group B.

Then $$\frac{(113 \times 50) + (115 \times 90) + (x \times 60)}{50 + 90 + 60} = 116$$

or $$\frac{16000 + 60x}{200} = 116$$

or $$16000 + 60x = 23200$$

or $$60x = 23200 - 16000$$

or $$x = \frac{7200}{60} = 120$$

Thus, the mean of group B is 120.

Computation of missing standard deviation:

$$d_1 = \bar{x}_{123} - \bar{x}_1 = 116 - 113 = 3$$

$$d_2 = \bar{x}_{123} - \bar{x}_2 = 116 - 120 = -4$$

$$d_3 = \bar{x}_{123} - \bar{x}_3 = 116 - 115 = 1$$

We assume x as the missing standard deviation.

$$\sigma_{123} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(x^2 + d_3^2)}{n_1 + n_2 + n_3}} = 7.746$$

$$= \sqrt{\frac{50(6^2 + 3^2) + 60(7^2 + (-4)^2) + 90(x^2 + 1^2)}{50 + 60 + 90}} = 7.746$$

$$= \sqrt{\frac{(50 \times 45) + (60 \times 65) + 90x^2 + 90}{200}} = 7.746$$

$$= \sqrt{\frac{2250 + 3900 + 90x^2 + 90}{200}} = 7.746$$

$$= \sqrt{\frac{6240 + 90x^2}{200}} = 7.746$$

$$= \frac{6240 + 90x^2}{200} = (7.746)^2$$

$$= \frac{6240 + 90x^2}{200} = 60$$

$$= 6240 + 90x^2 = 12000$$

$$= 90x^2 = 12000 - 6240$$

$$= x^2 = \frac{5760}{90}$$

$$= x^2 = 64$$

$$\therefore \quad x = \sqrt{64} = 8$$

Hence, standard deviation for group C is 8.

Example 7.19 The arithmetic mean and standard deviation of a series of 20 items were computed as 20 cm and 5 cm, respectively. While calculating these, an item 13 cm was misread as 30. Find the correct mean and standard deviation.

Solution

Given $\bar{x} = 20$ and $n = 20$

Now, $\bar{x} = \frac{\sum x}{n}$

$$20 = \frac{\sum x}{20} \quad \text{or} \quad \Sigma x = 20 \times 20 = 400$$

Corrected $\Sigma x = 400 - 30 + 13 = 383$

Corrected mean $= \frac{383}{n} = \frac{383}{20} = 19.15$

Corrected sum of squares (Σx^2) = Uncorrected sum of squares – sum of squares wrongly recorded observation + sum of squares of corrected observation.

$$\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = 5$$

or $$\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = 25$$

or $$\frac{\sum x^2}{n} - (20)^2 = 25$$

or $$\frac{\sum x^2}{n} - 400 = 25$$

or $$\Sigma x^2 = 425 \times 20$$
$$= 8500$$

Corrected sum of squares

$$\Sigma x^2 = 8500 - (30)^2 + (13)^2$$
$$= 8500 - 900 + 169$$
$$= 7769$$

Corrected standard deviation

$$\sigma^2 = \frac{7769}{20} - \left(\frac{383}{20}\right)^2$$
$$= 388.45 - 366.72$$
$$= 21.73$$

$$\sigma = \sqrt{21.73}$$
$$= 4.6615 \quad \text{or} \quad 4.66$$

Correct mean = 19.15.

Correct standard deviation 4.66.

Example 7.20 Suppose that a prospective buyer tests bursting pressure of samples of polythene bags received from two manufacturers A and B. The tests reveal the following results:

Bursting Pressure (lbs)	*5–10*	*10–15*	*15–20*	*20–25*	*25–30*
Number of bags – A	2	9	29	54	6
Number of bags – B	9	15	30	32	14

Which manufacturer's bags, judging from these two samples, have the higher average bursting pressure? Which of them is more uniform in bursting pressure?

Solution

Worksheet

Bursting Pressure (lbs)	*No. of Bags A*	*MV*	*d from 17.5*	*d/5*	*fd′*	*fd′²*
5–10	2	7.5	–10	–2	–4	8
10–15	9	12.5	–5	–1	–9	9
15–20	29	17.5	0	0	0	0
20–25	54	22.5	5	1	54	54
25–30	6	27.5	10	2	12	24
	100				53	95

Series A

$$\text{Mean} = A + \frac{\sum fd'}{n} \times c$$

$$= 17.5 + \frac{53}{100} \times 5$$

$$= 17.5 + 2.65 = 20.15$$

$$\sigma = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times c = \sqrt{\frac{95}{100} - \left(\frac{53}{100}\right)^2} \times 5$$

$$= \sqrt{0.95 - 0.2809} \times 5$$

$$= \sqrt{0.6691} \times 5 = 0.818 \times 5 = 4.09$$

$$\text{C.V.} = \frac{4.09}{20.15} = 0.2$$

Worksheet

Bursting Pressure (lbs)	*No. of Bags B*	*MV*	*d. from 17.5*	*d/s*	*fd′*	*fd′²*
5–10	9	7.5	–10	–2	–18	36
10–15	15	12.5	–5	–1	–15	15
15–20	30	17.5	0	0	0	0
20–25	32	22.5	5	1	32	32
25–30	14	27.5	10	2	28	56
	100				27	139

Series B

$$\text{Mean} = A + \frac{\sum fd'}{n} \times c = 17.5 + \left(\frac{27}{100} \times 5\right) = 17.5 + 1.35 = 18.85$$

$$\sigma = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times c = \sqrt{\frac{139}{100} - \left(\frac{27}{100}\right)^2} \times 5$$

$$= \sqrt{1.39 - (0.27)^2} \times 5$$

$$= \sqrt{1.39 - 0.0729} \times 5$$

$$= \sqrt{1.3171} \times 5$$

$$= 1.147 \times 5$$

$$= 5.735$$

$$\text{C.V.} = \frac{5.735}{18.85}$$

$$= 0.3$$

Conclusion

Manufacturer A has the higher average as well as more uniformity in bursting pressure.

Example 7.21 From the data given in the following table, find out which of the two series is more consistent?

Variable	10–20	20–30	30–40	40–50	50–60	60–70
Series A	10	16	34	38	24	18
Series B	18	22	38	34	20	8

Solution In order to find out which of the two series is more consistent, we have to compare the coefficient of variation of both the series.

Worksheet

Variable	*Mid-points*	d	d'	*Series A*			*Series B*		
				f	fd'	fd'^2	f	fd'	fd'^2
10–20	15	–20	–2	10	–20	40	18	–36	72
20–30	25	–10	–1	16	–16	16	22	–22	22
30–40	35	0	0	34	0	0	38	0	0
40–50	45	10	1	38	38	38	34	34	34
50–60	55	20	2	24	48	96	20	40	80
60–70	65	30	3	18	54	162	8	24	72
				140	104	352	140	40	280

For Series A

$$\bar{x} = \text{Arbitrary Mean} + \frac{\sum fd'}{n} \times c$$

$$= 35 + \left(\frac{104}{140} \times 10\right)$$

$$= 35 + 7.43 = 42.43$$

$$\sigma = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \times 10$$

$$= \sqrt{\frac{352}{140} - \left(\frac{104}{140}\right)^2} \times 10$$

$$= \sqrt{2.51 - (0.743)^2} \times 10$$

$$= \sqrt{2.51 - 0.55} \times 10$$

$$= \sqrt{1.96} \times 10 = 1.4 \times 10 = 14$$

Coefficient of Variation (V) $= \frac{\sigma}{\bar{x}} \times 100$

$$= \frac{14}{42.43} \times 100 = 0.3299 \times 100 = 33\%$$

For Series B

$$\bar{x} = \text{Arbitrary Mean} + \frac{\sum fd'}{n} \times c$$

$$= 35 + \frac{40}{140} \times 10$$

$$= 35 + 2.86$$

$$= 37.86$$

$$\sigma = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \times c$$

$$= \sqrt{\frac{280}{140} - \left(\frac{40}{140}\right)^2} \times 10$$

$$= \sqrt{2 - (0.286)^2} \times 10$$

$$= \sqrt{2 - 0.08} \times 10$$

$$= \sqrt{1.92} \times 10$$

$$= 1.385 \times 10 = 13.85$$

Coefficient of Variation (V) $= \frac{\sigma}{\bar{x}} \times 100$

$$= \frac{13.85}{37.86} \times 100$$

$$= 0.3658 \times 100$$

$$= 36.58\%$$

Since Series A has a smaller coefficient of variation than Series B, Series A is more consistent.

Example 7.22 A series comprising five individual observations is as follows:
8, 5, 2, 6 and 7
Convert these numbers into standard scores.

Solution The numbers given are:
8, 5, 2, 6 and 7

	d	d^2
8	2.4	5.76
5	–0.6	0.36
2	–3.6	12.96
6	0.4	0.16
7	1.4	1.96
28		21.2

$$\mu = \frac{28}{5} = 5.6$$

$$\sigma = \sqrt{\frac{\sum d^2}{n}} = \sqrt{\frac{21.2}{5}} = \sqrt{4.24} = 2.06$$

$$Z = \frac{x - \mu}{\sigma}$$

For ob. 8 $\quad \frac{8 - 5.6}{2.06} = \frac{2.4}{2.06} = 1.17$

5 $\quad \frac{5 - 5.6}{2.06} = \frac{-0.6}{2.06} = -0.29$

2 $\quad \frac{2 - 5.6}{2.06} = \frac{-3.6}{2.06} = -1.75$

6 $\quad \frac{6 - 5.6}{2.06} = \frac{0.4}{2.06} = 0.19$

7 $\quad \frac{7 - 5.6}{2.06} = \frac{1.4}{2.06} = 0.68$

The total of these standard scores, as it should be, is zero.

GLOSSARY

Coefficient of variation	A measure of relative variability that expresses the standard deviation as a percentage of the mean.
Dispersion	The spread or variability in a set of data.
Interquartile range	The difference between the values of the first and the third quartiles.

Mean deviation	A measure of dispersion that gives the average absolute differences (i.e. ignoring plus and minus signs) between each item and the mean.
Measures of dispersion	Measures that give the spread of a distribution.
Quartile deviation or Semi-interquartile range	A measure of dispersion, that is obtained by dividing the difference between the upper and the lower quartiles by two.
Range	Difference between the largest and the smallest values in a data set.
Standard deviation	The square root of the variance in a series. It shows how the data are spread out.
Standard score	The transformation of an observation by subtracting the mean and then dividing by the standard deviation. Thus, an observation is expressed in standard deviation units above or below the mean.
Standardised variable	A variable that expresses the x value of interest in terms of the number of standard deviations it is away from (that is, above or below) the mean. It is also known as the standardised normal variable.
Statistic	A summary measure calculated for sample data.
Variance	Average squared deviation between the mean and each item in a series.

LIST OF FORMULAE

1. Range = $L - S$
where L = value of the largest item and S = value of the smallest item.

2. Coefficient of range = $\dfrac{L-S}{L+S}$

3. Interquartile range = $Q_3 - Q_1$, where Q_3 and Q_1 are upper and lower quartiles, respectively.

4. Semi-interquartile range or Quartile deviation = $\dfrac{Q_3 - Q_1}{2}$

5. Coefficient of semi-interquartile range or Quartile deviation = $\dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

6. Mean deviation = $\dfrac{\sum |x|}{N}$
where $|x|$ stands for deviations from the mean ignoring plus and minus signs

7. Mean deviation in a grouped frequency = $\dfrac{\sum f|d|}{N}$
where $|d|$ stands for deviations from the mean ignoring plus and minus signs.

8. Variance of x = $\dfrac{\sum (x-\mu)^2}{N}$

9. Standard deviation σ of ungrouped data

$$= \sqrt{\frac{\sum (x-\mu)^2}{N}} = \sqrt{\frac{\sum d^2}{N}}$$

where d = deviation from the mean

10. Standard deviation in a grouped series

$$= \sqrt{\frac{\sum f(x-\mu)^2}{N}} = \sqrt{\frac{\sum fd^2}{N}}$$

where $d = x - \mu$

11. Standard deviation using the arbitrary mean

$$= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

where $d = x - A$ (arbitrary mean)

12. Standard deviation by the step-deviation method

$$= \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

where d' stands for deviations divided by C, the class-interval. C is used to simplify the calculations.

13. Combined standard deviation of two series

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where σ_1 = standard deviation of first group; σ_2 = standard deviation of second group; n_1 and n_2 being the number of observations in group 1 and group 2, respectively; $d_1 = (\bar{x}_1 - \bar{x}_{12})$ and $d_2 = (\bar{x}_2 - \bar{x}_{12})$.

14. Coefficient of variation

$$\text{CV} = \frac{\sigma}{\mu}(100)\%$$

A relative measure of dispersion, which is free from unit of measurement. It enables us to compare two or more distributions having different units of measurement.

15. Standardised variable

$$Z = \frac{x - \mu}{\sigma}$$

This is the standard score of an observation x in which we are interested. This shows the number of standard deviations the observation lies below or above the mean.

QUESTIONS

7.1 Given below are ten statements. Indicate in each case whether the statement is true or false:

(a) The difference between the largest and the smallest observations is called the quartile–range.

(b) The dispersion in a series indicates the reliability of the measure of central tendency.

(c) The interquartile range is based on only two values contained in a series.

(d) The square root of the variance gives the standard deviation.

(e) The coefficient of variation is not a relative measure of dispersion.

(f) In measuring dispersion, the standard deviation is more frequently used than the mean deviation.

(g) A major limitation of the range is that it ignores the large number of observations in a series.

(h) Even for an open-ended distribution, it is possible to measure the range.

(i) While calculating variance, every observation in a series is considered.

(j) The coefficient of variation is measured in the same units as the observations in a series.

Multiple Choice Questions (7.2 to 7.11)

7.2 Which of the following statements is *not* correct in respect of the range as a measure of dispersion?

(a) It is difficult to calculate.
(b) Only two points in the data set determine it.
(c) It is affected by extreme values.
(d) There may be considerable change in it from one sample to another.

7.3 If the first and third quartiles are 20.58 and 60.38, respectively, then the quartile deviation is

(a) 39.8 (b) 30.3
(c) 19.9 (d) None of the above

7.4 A series has its mean as 15 and its coefficient of variation as 20, its standard deviation is

(a) 5 (b) 10 (c) 3 (d) 7

7.5 If the first and third quartiles in a series are 15 and 35, then the semi-inter-quartile range is

(a) 30 (b) 20
(c) 10 (d) None of the above

7.6 Which of the following is a relative measure of dispersion?

(a) variance (b) coefficient of variance
(c) standard deviation (d) all of these

7.7 Which calculating a sample variance

(a) N is replaced by n (b) μ is replaced by $\bar{x}$
(c) N is replaced by n – 1 (d) (b) and (c), but not (a)
(e) (a) and (b) but not (c)

7.8 The square of the variance of a distribution is the

(a) absolute deviation
(b) mean
(c) standard deviation
(d) none of these

7.9 The standard deviation of a population divided by its mean and multiplied by 100 results into
(a) standard score
(b) variance
(c) coefficient of variation
(d) none of these

7.10 When calculating the population variance, the differences from the mean are squared because
(a) some differences are positive while other differences are negative.
(b) to avoid extreme values influencing the calculation.
(c) it is possible that total number of observations is quite small.
(d) none of these.

7.11 Which of the following is *not* true in respect of mean deviation?
(a) It is simple to understand.
(b) It considers each and every item in a series.
(c) It is capable of further algebraic treatment.
(d) The extreme items have less effect on its magnitude.

7.12 What do you mean by dispersion? What are the different measures of dispersion?

7.13 "Variability is not an important factor because even though the outcome is more certain, you still have an equal chance of falling either above or below the median. Therefore, on an average, the outcome will be the same." Do you agree with this statement? Give reasons for your answer.

7.14 Distinguish between measures of central tendency and measures of dispersion, pointing out the importance of studying dispersion.

7.15 Discuss the relative merits and demerits of any two measures of dispersion.

7.16 "Measures of central tendency and measures of dispersion are complementary to each other in highlighting the characteristics of a frequency distribution."

7.17 Discuss the range and the quartile deviation as measures of dispersion, using hypothetical data. Also point out their relative advantages and limitations.

7.18 What are the characteristics of a good measure of dispersion? How far quartile deviation and standard deviation satisfy these characteristics?

7.19 Do you think that the standard deviation is the best measure of dispersion amongst all the measures? Are there any limitations of the standard deviation?

7.20 Differentiate between an absolute measure and a relative measure of dispersion. Which one of these is preferable and why?

7.21 Comment on the validity of the following statement:
"The standard deviation of the heights measured in inches will be less than the standard deviation measured in feet for the same group of individuals."

7.22 Why is the standard deviation the most widely used measure of dispersion? Explain.

7.23 The mean and the standard deviation of two distributions of 100 and 150 items are 50 and 5; and 40 and 6, respectively. Find the mean and the standard deviation of all the 250 items taken together.

7.24 The mean and the standard deviation of 100 items are found by a student as 50 and 5. If at the time of calculation, two items are wrongly taken as 40 and 50 instead of 60 and 30, find the correct mean and the standard deviation.

7.25 The following table gives the height of students in a class. Find out the quartile deviation.

Height in Inches	50–53	53–56	56–59	59–62	62–65	65–68
No. of Students	2	7	24	27	13	3

7.26 Find the mean deviation of each set of numbers:
(a) 12, 6, 7, 3, 15, 10, 18 and 5
(b) 9, 3, 8, 8, 9, 8, 9 and 18

7.27 Find the range of weights of 100 students from the data given below:

Weight (kg)	60–62	63–65	66–68	69–71	72–74
No. of Students	5	18	42	27	8

7.28 The mean of five observations is 4.4 and the variance is 8.24. If three of the five observations are 1, 2 and 6, find the other two.

7.29 The following is a record of the number of bricks laid each day by two masons A and B:
A : 700, 675, 725, 675, 800, 650, 675, 625, 700 and 650
B : 600, 625, 675, 575, 650, 625, 600, 625, 550 and 700
Calculate the coefficient of variation in each case and discuss the relative consistency of the two masons. If the figures for A were in every case 20 more and those of B in every case 10 more than the figures given above, how would the answer be affected?

7.30 The following table gives the marks of 59 students in economics. Calculate the semi–interquartile range and its coefficient.

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

7.31 Given the set of numbers 2, 5, 8, 11, 14 and 2, 8, 14, find (a) the mean of each set, (b) the variance of each set, (c) the mean of the combined or 'pooled' sets, and (d) the variance of combined or pooled sets.

7.32 Work the preceding problem for the sets of numbers 2, 5, 8, 11, 14 and 10, 16, 22.

7.33 Prove that for any frequency distribution the total percentage of cases falling in the interval $\frac{1}{2}(Q_1 + Q_3) \pm \frac{1}{2}(Q_3 - Q_1)$ is 50 per cent. Is the same true for the interval $Q_2 \pm \frac{1}{2}(Q_3 - Q_1)$? Explain your answer.

7.34 Let the variance of individual observations 1, 2, 3, 4 and 5 be σ_1^2. Add two 3 s, that is, we now have 1, 2, 3, 3, 3, 4 and 5. Let the variance of these individual observations be σ_2^2. Will σ_2^2 be larger or smaller than σ_1^2?

7.35 **(a)** By adding 5 to each of the numbers in the set 3, 6, 2, 1, 7, 5, we obtain the set 8, 11, 7, 6, 12, 10. Show that the two sets have the same standard deviation but different means. How are the means related?
(b) By multiplying each of the numbers 3, 6, 2, 1, 7, 5 by 2 and then adding 5, we obtain the set 11, 17, 9, 7, 19, 15. What is the relationship between the standard deviations and the means for the two sets?
(c) What properties of the mean and the standard deviation are illustrated by the particular sets of numbers in (a) and (b)?

7.36 For the following data set, calculate the variance, the standard deviation and the coefficient of variation.

Frequency distribution of weight (lbs) for 100 persons

Weight	*Number of Plots*
130.5–140.5	10
140.5–150.5	20
150.5–160.5	30
160.5–170.5	20
170.5–180.5	10
180.5–190.5	10
Total	**100**

7.37 **(a)** If n_1, n_2 are the sizes; $\bar{x}_1, \bar{x}_2$, the means and σ_1, σ_2 the standard deviations of the two series, find the standard deviation σ of the combined series of size $n_1 + n_2$.

(b) Prove that for any discrete distribution, the standard deviation is not less than the mean deviation from the mean.

7.38 How is the standard deviation affected if **(a)** every item in a frequency distribution is increased by, say, 7; and **(b)** every item is multiplied by 7? Illustrate by taking a short hypothetical frequency distribution.

7.39 The arithmetic mean and standard deviation of a series of 20 items were computed as 20 cm and 5 cm, respectively. While calculating these, an item 13 cm was misread as 30. Find the correct mean and standard deviation.

7.40 The mean and the standard deviation of a continuous series are 31 and 15.9, respectively. The distribution after taking step deviations is as given below:

Table

x_i	–3	–2	–1	0	1	2	3
f_i	10	15	25	25	10	10	5

Determine the actual class intervals.

7.41 Using the following four numbers 1, 2, 4 and 5, **(i)** show that the variance of the standardised variable is unity. **(ii)** Show that when the variables are transformed into $(x - \bar{x})/\sigma$, the distribution becomes a unit distribution.

7.42 In two factories A and B engaged in similar type of industry, the average weekly wages and standard deviations are as given below:

	Factory A	*Factory B*
Average weekly wages (Rs)	460	490
Standard deviation of weekly wages (Rs)	50	40
Number of wage–earners	100	80

(i) Which, factory A or factory B, pays larger amount as weekly wages?
(ii) Which factory shows greater variability in the distribution of weekly wages?
(iii) What is the mean and the standard deviation of all the workers in two factories taken together?

7.43 The following data are given for two companies. Combining the data for groups of male and female employees, find out (i) which company has a higher average productivity per employee, and (ii) which company has a smaller dispersion of the productivity?

Productivity per Employee	*Company A*		*Company B*	
	Males	*Females*	*Males*	*Females*
Mean	30	20	27	32
Variance	8	3	12	5
No. of employees	40	10	20	30

7.44 For a group containing 100 observations, the arithmetic mean and the standard deviation are 8 and $\sqrt{10.5}$, respectively. For 50 observations selected from these observations, the mean and the standard deviation are 10 and 2, respectively. Find the mean and the standard deviation of the remaining 50 observations.

7.45 The following figures relate to weights in lbs of five persons: 120, 140, 150, 160 and 180. Calculate the variance and the standard deviation of the weights.

Subtract 40 lbs from each of the weights in the problem given above and then calculate the variance and the standard deviation. Do you find any change in your answers as compared to those of the preceding problem?

7.46 Given below are the daily wages paid to workers in two factories X and Y.

	Number of workers	
Daily Wages in Rs	*Factory X*	*Factory Y*
20–30	15	25
30–40	30	40
40–50	44	60
50–60	60	35
60–70	60	20
70–80	14	15
80–90	7	5

Using arithmetic mean and standard deviation, answer the following questions:
(a) Which factory pays higher average wages? By how much?
(b) In which factory are wages more variable?
(c) Which factory has to pay more wages in a month assuming that both work for 25 days in a month?

7.47 The first of two groups has 100 observations with mean 15 and standard deviation 5. If the entire group has 250 observations with mean 15.6 and variance 13.44, find the standard deviation of the second group.

7.48 Calculate standard deviation for average life of a particular band of T.V. sets

Life in years	*Number of sets*
0–2	5
2–4	16
4–6	13
6–8	7
8–10	5
10–12	4

7.49 Calculate standard deviation for the following distribution of net profits of 100 firms in a certain industry.

Net profit as per cent of sales	*Number of firms*
0–5	8
5–10	42
10–15	36
15–20	10
20–25	4

7.50 The distribution of wages of workers in two factories A and B is given below. Determine the factory in which total wages paid to all the workers is more, and the factory in which the wages are more variable.

Wages (in Rs)	*Number of workers*	
	A	*B*
50–100	2	6
100–150	9	11
150–200	29	18
200–250	54	32
250–300	11	27
300–350	5	11

The McGraw-Hill Companies

CHAPTER 8

SKEWNESS, MOMENTS AND KURTOSIS

Learning Objectives

By the end of your work on this chapter, you should be able to

- define skewness and distinguish it both from the mean and the dispersion
- calculate skewness by different methods
- define and calculate moments and kurtosis for a given set of data, and interpret the result.

Chapter Prerequisites

Before starting work on this chapter, you should be able to recollect the different measures of dispersion, along with the formulae used in their calculation.

8.1 INTRODUCTION

In Chapter 3, we discussed frequency distributions in detail. It may be reiterated here that frequency distributions differ in three ways:

1. Average value
2. Variability or dispersion, and
3. Shape

Since the first two, that is, average value and variability or dispersion have already been discussed in separate Chapters (6 and 7), here our main focus will be on the shape of frequency distribution. Taken together, all the three aspects will give us a more comprehensive idea about the given series. The discussion will pertain to relationships between shapes of frequency distributions and averages.

8.2 SKEWNESS

When the mean, median and mode do not have the same value in a distribution, then it is known as a *skewed distribution*. Skewness indicates lack of symmetry in a distribution. When a frequency distribution is elongated to the right, that is, having a longer tail to the right, it is said to be positively skewed. In contrast, if the distribution has a longer tail to the left, it is said to be negatively skewed. While discussing the relationship amongst mean, median and mode in Chapter 6 (See Section 6.5), Fig. 6.2 showed the curve of symmetrical distribution and Figs. 6.3 and 6.4 showed positively skewed

distribution and negatively skewed distribution, respectively. Since these curves have already been shown in that chapter, they are not given here.

It may be recalled that the mean is a centre of gravity or balance point, the median is that value that divides the distribution into equal areas, and the mode is the value indicating the largest frequency, that is, the maximum ordinate of the distribution. It will be seen that if mean > median > mode, the skewness is positive. Conversely, if mean < median < mode, the skewness is negative. It may be noted that a symmetrical distribution is not skewed. In contrast, a skewed distribution is unsymmetrical. In a symmetrical distribution, mean, median and mode all have the same value.

8.3 MEASURES OF SKEWNESS

There are four measures of skewness, each divided into absolute and relative measures. The relative measure is known as the coefficient of skewness and is more frequently used than the absolute measure of skewness. Further, when a comparison between two or more distributions is involved, it is the relative measure of skewness which is used.

The measures of skewness are:

(i) Karl Pearson's measure
(ii) Bowley's measure
(iii) Kelly's measure
(iv) Moment's measure

These measures are discussed briefly below:

Karl Pearson's Measure

The formula for measuring skewness as given by Karl Pearson is as follows:

$$\text{Skewness} = \text{Mean} - \text{Mode}$$

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

In case the mode is indeterminate, the coefficient of skewness is

$$Sk_P = \frac{\text{Mean} - (3\,\text{Median} - 2\,\text{Mean})}{\text{Standard deviation}}$$

$$Sk_P = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

Now this formula is equal to the earlier one.

$$= \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

or 3 Mean – 3 Median = Mean – Mode
or Mode = Mean – 3 Mean + 3 Median
or Mode = 3 Median – 2 Mean

The direction of skewness is determined by ascertaining whether the mean is greater than the mode or less than the mode. If it is greater than the mode, then skewness is positive. But when the mean is less than the mode, it is negative. The difference between the mean and mode indicates the extent of departure from symmetry. It is measured in standard deviation units, which provide a measure

independent of the unit of measurement. It may be recalled that this observation was made in the preceding chapter while discussing standard deviation.

The value of coefficient of skewness is zero, when the distribution is symmetrical. Normally, this coefficient of skewness lies between ± 1.

Example 8.1 Given the following data, calculate the Karl Pearson's coefficient of skewness:

$\Sigma x = 452$ $\Sigma x^2 = 24270$ Mode = 43.7 and $N = 10$

Solution Pearson's coefficient of skewness is

$$Sk_P = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$\text{Mean } (\bar{x}) = \frac{\Sigma x}{N} = \frac{452}{10} = 45.2$$

$$\text{SD } (\sigma) = \sqrt{\frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2}$$

$$= \sqrt{\frac{24270}{10} - \left(\frac{452}{10}\right)^2}$$

$$= \sqrt{2427 - (45.2)^2} = 19.59$$

Applying the values of mean, mode and standard deviation in the above formula,

$$Sk_P = \frac{45.2 - 43.7}{19.59}$$

$$= 0.08$$

This shows that there is a positive skewness though the extent of skewness is marginal.

Example 8.2 From the following data, calculate the measure of skewness using the mean, median and standard deviation:

X	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	18	30	40	55	38	20	16

Solution

Worksheet

x	*MV x*	d_x	*f*	fd_x	fd_{x^2}	*cf*
10–20	15	–3	18	– 54	162	18
20–30	25	–2	30	– 60	120	48
30–40	35	–1	40	– 40	40	88
40–50	45	0	55	0	0	143
50–60	55	1	38	38	38	181
60–70	65	2	20	40	80	201
70–80	75	3	16	48	144	217
		Total	**217**	**–28**	**584**	

a = Arbitrary mean = 45
cf = Cumulative frequency
d_x = Deviation from arbitrary mean
i = 10

$$\bar{x} = a + \frac{\Sigma fdx}{N} \times i$$

$$= 45 - \frac{28}{217} \times 10 = 43.71$$

$$\text{Median} = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

where $m = (N + 1)/2^{\text{th}}$ item

$$= (217 + 1)/2 = 109^{\text{th}} \text{ item}$$

$$\text{Median} = 40 + \frac{50 - 40}{55}(109 - 88)$$

$$= 40 + \frac{10}{55} \times 21$$

$$= 43.82$$

$$\text{SD} = \sqrt{\frac{\Sigma fd_x^2}{\Sigma f} - \left(\frac{\Sigma fd_x}{\Sigma f}\right)^2} \times 10 = \sqrt{\frac{584}{217} - \left(\frac{-28}{217}\right)^2} \times 10$$

$$= \sqrt{2.69 - 0.016} \times 10 = 16.4$$

$$\text{Skewness} = 3\ (\text{Mean} - \text{Median})$$
$$= 3\ (43.71 - 43.82)$$
$$= 3 \times -0.11$$
$$= -0.33$$

Coefficient of skewness

$$= \frac{\text{Skewness}}{\text{SD}}$$

$$= \frac{-0.33}{16.4}$$

$$= -0.02$$

The result shows that the distribution is negatively skewed, but the extent of skewness is extremely negligible.

Bowley's Measure

Bowley developed a measure of skewness, which is based on quartile values. The formula for measuring skewness is:

$$\text{Skewness} = (Q_3 - Q_2) - (Q_2 - Q_1)$$

where Q_3 and Q_1 are upper and lower quartiles.

The value of this skewness varies between ± 1. In the case of open-ended distribution as well as where extreme values are found in the series, this measure is particularly useful.

In a symmetrical distribution, skewness is zero. This means that Q_3 and Q_1 are positioned equidistantly from Q_2, that is, the median. In symbols, $Q_3 - Q_2 = Q_2 - Q_1$. In contrast, when the distribution is skewed, then $Q_3 - Q_2$ will be different from $Q_2 - Q_1$. When $Q_3 - Q_2$ exceeds $Q_2 - Q_1$, then skewness is positive. As against this, when $Q_3 - Q_2$ is less than $Q_2 - Q_1$, then skewness is negative. Bowley's measure of skewness can be written as:

$$\text{Skewness} = (Q_3 - Q_2) - (Q_2 - Q_1)$$

This can be written as $Q_3 - Q_2 - Q_2 + Q_1$

or $Q_3 + Q_1 - 2Q_2$ ($2Q_2$ is $2M$)

However, this is an absolute measure of skewness. As such, it cannot be used while comparing two distributions where the units of measurement are different. In view of this limitation, Bowley suggested a relative measure of skewness as given below:

$$\text{Relative skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

$$= \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Example 8.3 For a distribution, Bowley's coefficient of skewness is –0.56, $Q_1 = 16.4$ and Median = 24.2. What is the coefficient of quartile deviation?

Solution Bowley's coefficient of skewness is:

$$\text{Sk}_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Substituting the values in the above formula,

$$\text{Sk}_B = \frac{Q_3 + 16.4 - (2 \times 24.2)}{Q_3 - 16.4}$$

or $$-0.56 = \frac{Q_3 + 16.4 - 48.4}{Q_3 - 16.4}$$

or $$-0.56\,(Q_3 - 16.4) = Q_3 - 32$$

or $$-0.56\,Q_3 + 9.184 = Q_3 - 32$$

or $$-0.56\,Q_3 - Q_3 = -32 - 9.184$$

or $$-1.56\,Q_3 - = -41.184$$

$\therefore$ $$Q_3 = \frac{-41.184}{-1.56} = 26.4$$

Now, we have the values of both the upper and the lower quartiles.

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{26.4 - 16.4}{26.4 + 16.4} = \frac{10}{42.8} = 0.234 \text{ approx.}$$

Example 8.4 Calculate an appropriate measure of skewness from the following data:

Value in Rs	*Frequency*
Less than 50	40
50–100	80
100–150	130
150–200	60
200 and above	30

Solution It should be noted that the series given in the question is an open-ended series. As such, Bowley's coefficient of skewness, which is based on quartiles, would be the most appropriate measure of skewness in this case. In order to calculate the quartiles and the median, we have to use the cumulative frequency. The table is reproduced below with the cumulative frequency.

Value in Rs	*Frequency*	*Cumulative Frequency*
Less than 50	40	40
50–100	80	120
100–150	130	250
150–200	60	310
200 and above	30	340

$$Q_1 = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

Now $$m = \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item} = \frac{341}{4} = 85.25, \text{ which lies in 50–100 class}$$

$$Q_1 = 50 + \frac{100 - 50}{80}(85.25 - 40) = 78.28$$

$$m = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item} = \frac{341}{2} = 170.5, \text{ which lies in 100–150 class}$$

$$M = 100 + \frac{150 - 100}{130}(170.5 - 120) = 119.4$$

$$Q_3 = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

where $$m = 3(341) \div 4 = 255.75$$

$$Q_3 = 150 + \frac{200 - 150}{60}(255.75 - 250) = 154.79$$

Bowley's coefficient of skewness is:

$$\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{154.79 + 78.28 - (2 \times 119.4)}{154.79 - 78.28} = \frac{-5.73}{76.51}$$

$$= -0.075 \text{ approx.}$$

This shows that there is a negative skewness, which has a very negligible magnitude.

Example 8.5 In a frequency distribution, the coefficient of skewness based on quartiles is 0.5. If the sum of the upper and the lower quartiles is 28 and the median is 11, find the values of the lower and upper quartiles.

Solution Bowley's coefficient of skewness is;

$$\text{Sk}_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Applying the values given in the question in the above formula, we get

$$0.5 = \frac{Q_3 + Q_1 - (2 \times 11)}{Q_3 - Q_1}$$

or $$0.5 = \frac{28 - 22}{Q_3 - Q_1}$$

or $$0.5\,(Q_3 - Q_1) = 6$$

or $$Q_3 - Q_1 = 6/0.5 = 12$$

since $$Q_3 + Q_1 = 28 \quad \text{(i)}$$

and $$Q_3 - Q_1 = 12 \quad \text{(ii)}$$

By adding (i) and (ii), we get $2Q_3 = 40$

$\therefore$ $$Q_3 = 20$$

and $$Q_1 = 28 - 20 = 8$$

Hence, the lower and the upper quartiles are 8 and 20, respectively.

Kelly's Measure

Kelly developed another measure of skewness, which is based on percentiles.

The formula for measuring skewness is as follows:

$$\text{Coefficient of skewness} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

or

$$\frac{D_1 + D_9 - 2M}{D_9 - D_1}$$

where P and D stand for percentile and decile, respectively.

In order to calculate the coefficient of skewness by this formula, we have to ascertain the values of 10^{th}, 50^{th} and 90^{th} percentiles. Somehow, this measure of skewness is seldom used. All the same, we give an example to show how it can be calculated.

Example 8.6

Class-intervals	*f*	*cf*
10–20	18	18
20–30	30	48
30–40	40	88
40–50	55	143
50–60	38	181
60–70	20	201
70–80	16	217

Use Kelly's measure to calculate skewness.

Solution Now we have to calculate P_{10}, P_{50} and P_{90}.

$$P_{10} = l_1 + \frac{l_2 - l_1}{f_1}(m - c), \text{ where } m = (n+1)/10^{\text{th}} \text{ item}$$

$$= \frac{217+1}{10} = 21.8^{\text{th}} \text{ item}$$

This lies in the 20–30 class.

$$= 20 + \frac{30-20}{30}(21.8 - 18) \qquad = 20 + \frac{10 \times 3.8}{30} = 21.27 \text{ approx.}$$

P_{50} (median): where $m = (n+1)/2^{\text{th}}$ item $= \frac{217+1}{2} = 109^{\text{th}}$ item

This lies in the class 40–50. Applying the above formula

$$= 40 + \frac{50-40}{55}(109 - 88) \qquad = 40 + \frac{10 \times 21}{55} \times 21 = 43.82 \text{ approx.}$$

P_{90}: here $m = 90\ (217+1)/100^{\text{th}}$ item $= 196.2^{\text{th}}$ item

This lies in the class 60–70. Applying the above formula

$$= 60 + \frac{70-60}{20}(196.2 - 181) = 60 + \frac{10 \times 15.2}{20} = 67.6 \text{ approx.}$$

Kelly's skewness

$$\text{Sk}_K = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

$$= \frac{67.6 - (2 \times 43.82) + 21.27}{67.6 - 21.27}$$

$$= \frac{88.87 - 87.64}{46.33}$$

$$= 0.027$$

This shows that the series is positively skewed though the extent of skewness is extremely negligible. It may be recalled that if there is a perfectly symmetrical distribution, then the skewness will be zero. One can see that the above answer is very close to zero.

8.4 MOMENTS

In mechanics, the term *moment* is used to denote the rotating effect of a force. In Statistics, it is used to indicate peculiarities of a frequency distribution. The utility of moments lies in the sense that they indicate different aspects of a given distribution. Thus, by using moments, we can measure the central tendency of a series, dispersion or variability, skewness and the peakedness of the curve.

The moments about the actual arithmetic mean are denoted by μ. The first four moments about mean or *central moments* are as follows:

First moment $\qquad \mu_1 = \frac{1}{N}\Sigma(x_i - \bar{x})$

Second moment $\mu_2 = \frac{1}{N}\Sigma(x_i - \bar{x})^2$

Third moment $\mu_3 = \frac{1}{N}\Sigma(x_i - \bar{x})^3$

Fourth moment $\mu_4 = \frac{1}{N}\Sigma(x_i - \bar{x})^4$

These moments are in relation to individual items. In the case of a frequency distribution, the first four moments will be

First moment $\mu_1 = \frac{1}{N}\Sigma f_i(x_i - \bar{x})$

Second moment $\mu_2 = \frac{1}{N}\Sigma f_i(x_i - \bar{x})^2$

Third moment $\mu_3 = \frac{1}{N}\Sigma f_i(x_i - \bar{x})^3$

Fourth moment $\mu_4 = \frac{1}{N}\Sigma f_i(x_i - \bar{x})^4$

It may be noted that the first central moment is zero, that is, $\mu_1 = 0$.

The second central moment is $\mu_2 = \sigma^2$, indicating the variance.

The third central moment μ_3 is used to measure skewness. The fourth central moment gives an idea about the Kurtosis.

Karl Pearson suggested another measure of skewness, which is based on the third and second central moments as given below:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

Example 8.7 Find the (**a**) first, (**b**) second, (**c**) third and (**d**) fourth moments for the set of numbers 2, 3, 4, 5 and 6.

Solution

(a) $\bar{x} = \frac{\Sigma x}{N} = \frac{2+3+4+5+6}{5} = \frac{20}{5} = 4$

(b) $\bar{x}^2 = \frac{\Sigma x^2}{N} = \frac{2^2+3^2+4^2+5^2+6^2}{5}$

$= \frac{4+9+16+25+36}{5} = 18$

(c) $\bar{x}^3 = \frac{\Sigma x^3}{N} = \frac{2^3+3^3+4^3+5^3+6^3}{5}$

$= \frac{8+27+64+125+216}{5} = 88$

(d) $\bar{x}^4 = \dfrac{\Sigma x^4}{N} = \dfrac{2^4 + 3^4 + 4^4 + 5^4 + 6^4}{5}$

$= \dfrac{16 + 81 + 256 + 625 + 1296}{5} = 454.8$

Example 8.8 Using the same set of five figures as given in Example 8.7, find the **(a)** first, **(b)** second, **(c)** third and **(d)** fourth moments about the mean.

Solution

$$m_1 = (x - \bar{x}) = \frac{\Sigma(x - \bar{x})}{N} = \frac{(2-4) + (3-4) + (4-4) + (5-4) + (6-4)}{5}$$

$$= \frac{-2 - 1 + 0 + 1 + 2}{5} = 0$$

$$m_2 = (x - \bar{x})^2 = \frac{\Sigma(x - \bar{x})^2}{N} = \frac{(2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2}{5}$$

$$= \frac{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2}{5}$$

$$= \frac{4 + 1 + 0 + 1 + 4}{5} = 2$$

It may be noted that m_2 is the variance.

$$m_3 = (x - \bar{x})^3 = \frac{\Sigma(x - \bar{x})^3}{N} = \frac{(2-4)^3 + (3-4)^3 + (4-4)^3 + (5-4)^3 + (6-4)^3}{5}$$

$$= \frac{(-2)^3 + (-1)^3 + 0^3 + 1^3 + 2^3}{5}$$

$$= \frac{-8 - 1 + 0 + 1 + 8}{5} = 0$$

$$m_4 = (x - \bar{x})^4 = \frac{\Sigma(x - \bar{x})^4}{N} = \frac{(2-4)^4 + (3-4)^4 + (4-4)^4 + (5-4)^4 + (6-4)^4}{5}$$

$$= \frac{(-2)^4 + (-1)^4 + 0^4 + 1^4 + 2^4}{5}$$

$$= \frac{16 + 1 + 0 + 1 + 16}{5} = 6.8$$

8.5 KURTOSIS

Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess. While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution.

Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are *mesokurtic*, *leptokurtic* and *platykurtic*. These three types of curves are shown in Fig. 8.1.

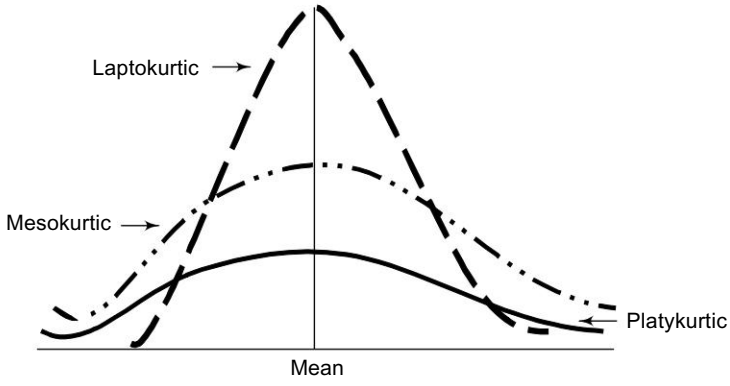


Fig. 8.1 Types of Curves

It will be seen from Fig. 8.1 that mesokurtic curve is neither too much flattened nor too much peaked. In fact, this is the frequency curve of a normal distribution. Leptokurtic curve is a more peaked than the normal curve. In contrast, platykurtic is a relatively flat curve.

The coefficient of kurtosis as given by Karl Pearson is $\beta_2 = \mu_4/\mu_2^2$. In case of a normal distribution, that is, mesokurtic curve, the value of $\beta_2 = 3$. If β_2 turns out to be > 3, the curve is called a laptokurtic curve and is more peaked than the normal curve. Again, when $\beta_2 < 3$, the curve is called a platykurtic curve and is less peaked than the normal curve.

The measure of kurtosis is very helpful in the selection of an appropriate average. For example, for normal distribution, mean is most appropriate; for a leptokurtic distribution, median is most appropriate; and for platykurtic distribution, the quartile range is most appropriate.

Example 8.9 From the data given below, calculate the percentile coefficient of kurtosis.

Daily Wages in Rs	*Number of Workers*	*cf*
50–60	10	10
60–70	14	24
70–80	18	42
80–90	24	66
90–100	16	82
100–110	12	94
110–120	6	100
Total	**100**	

Solution It may be noted that the question involved first two columns and in order to calculate quartiles and percentiles, cumulative frequencies have been shown in column three of the above table.

$$Q_1 = l_1 + \frac{l_2 - l_1}{f_1}(m - c), \text{ where } m = (n + 1)/4^{\text{th}} \text{ item, which is} = 25.25^{\text{th}} \text{ item}$$

This falls in 70–80 class-interval.

$$= 70 + \frac{80 - 70}{18} (25.25 - 24) = 70.69$$

$$Q_3 = l_1 + \frac{l_2 - l_1}{f_1} (m - c), \text{ where } m = 75.75$$

This falls in 90 – 100 class-interval

$$= 90 + \frac{100 - 90}{16} (75.75 - 66) = 96.09$$

$$P_{10} = l_1 + \frac{l_2 - l_1}{f_1} (m - c), \text{ where } m = 10.1$$

This falls in 60 – 70 class-interval.

$$= 60 + \frac{70 - 60}{14} (10.1 - 10) = 60.07$$

$$P_{90} = l_1 + \frac{l_2 - l_1}{f_1} (m - c), \text{ where } m = 90.9$$

This falls in 100 – 110 class-interval.

$$= 100 + \frac{110 - 100}{12} (90.9 - 82) = 107.41$$

$$K = \frac{Q_D}{P_{90} - P_{10}}$$

$$= \frac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}}$$

$$= \frac{\frac{1}{2}(96.09 - 70.69)}{107.41 - 60.07}$$

$$= 0.268$$

It will be seen that the above distribution is very close to normal distribution as the value of K is 0.263 in a normal distribution.

Additional Examples

Example 8.10 In a frequency distribution, coefficient of skewness based on quartiles is 0.6. If the sum of the upper and lower quartiles is 100 and the median is 38, find the value of the lower and upper quartiles.

Solution

$$\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = 0.6$$

$$= \frac{100 - (2 \times 38)}{Q_3 - Q_1} = 0.6$$

$$= \frac{100 - 76}{Q_3 - Q_1} = 0.6$$

By cross multiplication

$$(0.6)\,(Q_3 - Q_1) = 24$$

or $$Q_3 - Q_1 = \frac{24}{0.6} = 40$$

Now $$Q_3 - Q_1 = 40$$

$$Q_3 + Q_1 = 100$$

$$2Q_3 = 140$$

$$\therefore \quad Q_3 = 70$$

$$\therefore \quad Q_1 = 100 - Q_3 = 100 - 70 = 30$$

Example 8.11 Calculate the Pearson's measure of skewness on the basis of mean, mode and standard deviation:

Mid-value	14.5	15.5	16.5	17.5	18.5	19.5	20.5	21.5
Frequency	35	40	48	100	125	87	43	22

Solution

Worksheet

Mid-value	*F*	*Dev. from 17.5*	*fd*	fd^2
14.5	35	–3	–105	315
15.5	40	–2	–80	160
16.5	48	–1	–48	48
17.5	100	0	0	0
18.5	125	1	125	125
19.5	87	2	174	348
20.5	43	3	129	387
21.5	22	4	88	352
	500		283	1735

$$\text{Mean} = A + \frac{\Sigma fd}{n}$$

$$= 17.5 + \frac{283}{500}$$

$$= 17.5 + 0.57 = 18.07$$

Mode is against the largest frequency of 125, i.e., in the class 18–19.

$$\text{Mode} = l_1 + \frac{f_1}{(f_1 - f_0) + (f_1 - f_2)}$$

$$= 18 + \frac{125}{(125 - 100) + (125 - 87)}$$

$$= 18 + \frac{125}{25 + 38}$$

$$= 18 + \frac{125}{63}$$

$$= 18 + 1.98 = 19.98$$

$$\text{SD} = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$$

$$= \sqrt{\frac{1735}{500} - \left(\frac{283}{500}\right)^2}$$

$$= \sqrt{3.47 - 0.32}$$

$$= \sqrt{3.15} = 1.77$$

$$\text{Sk} = \frac{\text{Mean} - \text{Mode}}{\text{SD}}$$

$$= \frac{18.07 - 19.98}{1.77}$$

$$= -\frac{1.91}{1.77}$$

$$= -1.08 \quad \text{Negative skewness}$$

Example 8.12 Given the following figures: 3, 5, 3, 5, 10, 5, 15, 10, 12, 12. Calculate a measure of relative kurtosis.

Solution As a measure of relative kurtosis is

$$\beta_2 = \mu_4/\mu_2^2$$

We have to calculate μ_2 and μ_4.

Worksheet

Observations	*d from 8*	d^2	d^4
3	–5	25	625
5	–3	9	81
3	–5	25	625
5	–3	9	81
10	2	4	16
5	–3	9	81
15	7	49	2401
10	2	4	16
12	4	16	256
12	4	16	256
80		166	4438

Mean = 80/10 = 8

$$\mu_2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{166}{10} = 16.6$$

Hence $\mu_2^2 = (16.6)^2 = 275.56$

$$\mu_4 = \frac{\sum(x_i - \bar{x})^4}{n} = \frac{4438}{10} = 443.8$$

$$\beta_2 = \mu_4/\mu_2^2 = \frac{443.8}{275.56} = 1.61$$

Since β_2 is less than 3, this series is platykurtic, i.e., more widened.

Example 8.13 For a distribution, the first four moments about zero are 1, 7, 38 and 155 respectively. **(i)** Compute the moment coefficients of skewness and kurtosis. **(ii)** Is the distribution mesokurtic? Give reason.

Solution Given:

First four moments about zero

1, 7, 38 and 155

We have first to calculate moments about mean

$$\mu_1 = \mu_1' - \mu_1'$$
$$= 1 - 1 = 0$$
$$\mu_2 = \mu_2' - \mu_1'^2$$
$$= 7 - (1)^2 = 7 - 1 = 6$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$
$$= 38 - (3 \times 7 \times 1) + (2 \times (1)^3)$$
$$= 38 - 21 + 2 = 19$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 155 - (4 \times 38 \times 1) + (6 \times 7 \times (1)^2) - (3 \times (1)^4)$$
$$= 155 - 152 + 42 - 3$$
$$= 42$$

Coefficient of skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(19)^2}{(6)^3} = \frac{361}{216} = 1.67$$

Coefficient of kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{42}{(6)^2} = \frac{42}{36} = 1.17$$

The distribution is not mesokurtic as β_2 is not equal to 3. As β_2 is $\angle 3$, it is platykurtic distribution.

Example 8.14 The first four moments of a distribution about the value 4 are 1, 4, 10 and 45. Obtain various characteristics of the distribution on the basis of the information given. Comment upon the nature of the distribution.

Solution

$$\mu_1' = 1$$

$$\mu_1 = \mu_1' - \mu_1' = 1 - 1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 4 - (1)^2 = 3$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$
$$= 10 - (3 \times 4 \times 1) + 2(1)^3$$
$$= 10 - 12 + 2$$
$$= 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 45 - (4 \times 10 \times 1) + (6 \times 4 \times (1)^2) - 3(1)^4$$
$$= 45 - 40 + 24 - 3$$
$$= 26$$

The second moment μ_2 shows variance.
The third moment μ_3 shows skewness.
The fourth moment μ_4 gives an idea of kurtosis.

As μ_3 is 0, hence $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ is 0. That means the related distribution is normal and there is no skewness.

$$\beta_2 = \mu_4/\mu_2^2$$
$$= \frac{26}{(3)^2} = \frac{26}{9} = 2.89$$

As it is less than 3, the distribution in platykurtic suggesting that it is a normal distribution with wide spread out.

Example 8.15 If $\beta_1 = 1$ and $\beta_2 = 4$ and variance = 9, find the values of μ_3 and μ_4 and comment upon the nature of the distribution.

Solution Given $\beta_1 = 1$, $\beta_2 = 4$ and variance = 9
We have to find the values of μ_3 and μ_4.

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$4 = \frac{\mu_4}{(9)^2}$$

$$\mu^4 = 4 \times (9)^2 = 4 \times 81 = 324$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$1 = \frac{\mu_3^2}{(9)^3}$$

$$\mu_3^2 = 1 \times (9)^3 = 729$$

$$\mu_3 = \sqrt{729} = 27$$

Since $\beta_2 > 3$ the distribution is leptokurtic, indicating it is a peaked normal distribution.

GLOSSARY

Bowley's measure of skewness	A measure of skewness based on quartile values. It varies between ±1.
Karl Pearson's measure of skewness	Difference between the mean and the mode divided by the standard deviation of a given data set.
Kelly's measure of skewness	A measure of skewness based on percentiles.
Kurtosis	The degree of 'peakedness' or 'flatness' of a frequency polygon.
Leptokurtic curve	A distribution in which most of the observations are concentrated near the mode and in tails.
Mesokurtic curve	A distribution that is less peaked than a leptokurtic curve.
Moments	A concept that indicates different aspects of a given distribution. By using moments, we can measure the central tendency of a series, dispersion or variability, skewness and the pickedness of the curve.
Negative skewness	When more observations lie to the right of the mean, the longer tail of the distribution extends to the left.
Platykurtic curve	A 'flat' distribution like a table or plateau.
Positive skewness	When more observations lie to the left of the mean, the longer tail of the distribution extends to the right.
Skewness	The extent of non-symmetry or 'lop-sidedness' of a distribution.
Symmetrical curve	A 'bell-shaped' curve.

LIST OF FORMULAE

1. Skewness = Mean – Mode

2. Coefficient of skewness (Karl Pearson's formula)

(i) $\dfrac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$ (ii) $\dfrac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$

The second formula is to be used when mode is ill-defined.

3. Skewness (Bowley's measure)

$$\text{Sk}_B = (Q_3 - Q_2) - (Q_2 - Q_1) \quad \text{where} \quad Q_1 = \text{Lower quartile}$$

or $Q_3 + Q_1 - 2Q_2$ $Q_3 = \text{Upper quartile}$

4. Coefficient of skewness

$$\text{Sk}_B = \frac{Q_3 + Q_1 - 2\text{M}}{Q_3 - Q_1}$$

5. Kelly's coefficient of skewness

$$\text{Sk}_K = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \quad \text{(Based on percentiles)}$$

where P_{90} = value of 90^{th} percentile
P_{10} = value of 10^{th} percentile
P_{50} = value of the median

6. Bowley's another measure of coefficient of skewness based on moments.

Coefficient of skewness $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$

7. Moments with mean as the origin

$$\mu_1 = \frac{\Sigma(x_i - \bar{x})}{N}$$

$$\mu_2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

$$\mu_3 = \frac{\Sigma(x_i - \bar{x})^3}{N}$$

$$\mu_4 = \frac{\Sigma(x_i - \bar{x})^4}{N}$$

8. Coefficient of kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where $\beta_2 = 3$, for mesokurtic distribution
$\beta_2 > 3$, for leptokurtic distribution
$\beta_2 < 3$, for platykurtic distribution

QUESTIONS

8.1 Given below are twelve statements. Indicate in each case whether it is true or false.

(a) In a skewed distribution, the mean, median and mode do not have the same value.

(b) In a frequency distribution, if a curve has a longer tail to the right, then it is negatively skewed.

(c) In a positively skewed curve, mean < median < mode.

(d) The median does not always lie between the mean and the mode in a skewed distribution.

(e) Karl Pearson's measure of skewness is the most frequently used measure.

(f) Kelly's measure of skewness is more frequently used as compared to Bowley's.

(g) Kelly's measure of skewness is based on percentiles.

(h) The second central moment does not indicate the variance.

(i) The first central moment is always zero.

(j) A leptokurtic curve is neither too much flattened nor too much peaked.

(k) A platykurtic curve is a relatively flat curve.

(l) For laptokurtic distribution, the median is the most appropriate average.

Multiple Choice Questions (8.2 to 8.8)

8.2 Which one is the formula for relative skewness?

(a) Mean – Mode (b) $(Q_3 - Q_2) - (Q_2 - Q_1)$

(c) $\dfrac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$ (d) None of the above

8.3 Which one of the following is a measure of skewness?

(a) First moment (b) Second moment

(c) Third moment (d) Fourth moment

(e) None of these

8.4 Which one of the curves is more peaked than the normal curve?

(a) Mesokurtic (b) Platykurtic (c) Laptokurtic (d) None of the above

8.5 A curve which is elongated to the right, shows that it is

(a) Skewed to the left (b) Symmetrical

(c) Positively skewed (d) None of these

8.6 For a distribution, mean = 43.71, median = 43.82 and standard deviation = –0.33, the coefficient of skewness is

(a) 0.05 (b) 0.2 (c) –0.02 (d) None of these

8.7 In a symmetrical series

(a) mode > mean (b) mean > median

(c) mean < mode (d) mean = mode = median

8.8 Which of the following relationship is valid in a symmetrical distribution?

(a) (Median – Q_1) < (Q_3 – Median) (b) (Median – Q_1) > (Q_3 – Median)

(c) Median – Q_1 = (Q_3 – Median) (d) None of these

8.9 What is skewness? How does it differ from dispersion?

8.10 What are the different measures of skewness? Which one is frequently used?

8.11 How would you differentiate between a symmetrical and a skewed distribution?

8.12 What is the significance of skewness? What are the objectives of measuring it?

8.13 What are Bowley's absolute and relative measures of skewness?

8.14 Explain Karl Pearson's measure of skewness, illustrating with a hypothetical example.

8.15 Distinguish between Karl Pearson's and Bowley's measures of skewness. Which one would you prefer and why?

8.16 How are the quartiles of a frequency distribution used in measuring skewness? Illustrate with a hypothetical example.

8.17 "Measures of dispersion and skewness are complimentary to one another in understanding a frequency distribution." Elucidate the statement.

8.18 Show different types of skewness by drawing different diagrams and indicate broadly the location of the mean, the median and the mode in each diagram.

8.19 What is kurtosis? How does it differ from skewness?

8.20 Explain the concept of moments in Statistics. How is this concept useful in determining the shape of a particular frequency distribution?

8.21 Define moments. Explain clearly how the moments help in describing the characteristics of a frequency distribution.

8.22 The performance of two teachers employed in a school is being evaluated, one of whom is to be retained in the service. You find that the evaluation gives the same mean performance score and variance in both the cases. Show how, in such a situation, skewness can be helpful in your decision.

8.23 Given the following figures:
3, 5, 3, 5, 10, 5, 15, 10, 12, 12
Calculate Karl Pearsonian coefficient of skewness and characterise the shape of the distribution.

8.24 From the data given in Question 8.23, calculate the moments about zero and then moments about the mean.

8.25 Calculate Karl Pearson's coefficient of skewness from the following data:

Weekly Sales (Rs in lakh)	*Number of Companies*
10–12	12
12–14	18
14–16	35
16–18	42
18–20	50
20–22	45
22–24	30
24–26	8

Comment on the value obtained.

8.26 From the following data of age of employees, calculate the coefficient of skewness and comment on the result:

Age in Years	*Number of Employees*
Below 25	8
Below 30	20
Below 35	40
Below 40	65
Below 45	80
Below 50	92
Below 55	100

8.27 Calculate the coefficient of skewness from the following distribution:

Scores	10–20	20–30	30–40	40–50	50–60	60–70
Frequency	12	18	26	32	14	8

8.28 The first four moments of a distribution about the value 5 are equal to 2, 20, 40 and 50. Obtain the mean, variance, β_1 and β_2 for the distribution.

8.29 The standard deviation of a symmetrical distribution is given to be 5. What must be the value of the fourth moment about the mean in order that the distribution is mesokurtic?

8.30 For a distribution the mean is 10, standard deviation is 4, and $\sqrt{\beta_1} = 1$ and $\beta_2 = 4$. Obtain the first four moments about the origin, that is, zero.

8.31 Find the coefficient of skewness from the following information:
Difference of two quartiles = 8, mode = 11
Sum of two quartiles = 22 and mean = 8.0

8.32 Calculate the first four moments about the mean from the following data. Also calculate the values of β_1 and β_2.

Marks	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of Students	5	12	18	40	15	7	3

8.33 Find the four moments about the mean from the following data and decide whether it is a platykurtic distribution.

Central Size of Items	*Frequency*
1	2
2	3
3	5
4	4
5	1

CHAPTER 9 PROBABILITY

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand some basic terminology in probability and define probability in a given situation using suitable empirical methods
- solve problems involving calculation of simple, joint and conditional probabilities
- construct decision tables and trees, and make use of expected values for decision-making
- know the limitations of this approach to decision-making.

Chapter Prerequisites

Before starting work on this chapter, make sure you are fully conversant with

1. handling fractions in your calculations
2. conversion of fractions to decimals or percentages, and vice versa
3. calculations of the arithmetic mean.

9.1 INTRODUCTION

Probability is a part of our daily lives though many of us may not be conscious of it. Consider the following statements:

1. It is likely to rain heavily this evening.
2. This year the postgraduate results of the university would be better than those of the last year.
3. The price of equity shares of company X is expected to increase significantly in the next few days.

All these statements are based on some probability. This shows that the concept of probability is commonly used in our daily life. In business, firms always face uncertain situations and yet they have to take decisions. Sometimes the stakes are too high. A wrong decision may involve huge loss to the firm. As such, great care has to be exercised before taking a particular decision. In this respect, the management is guided by the *theory of probability*.

What is Probability?

What is the meaning of the word 'probability'? Probability is the chance that a particular event will occur. What is the chance of getting a head when a coin is tossed? What is the chance of picking a red

card from a deck of playing cards? A company has launched an add campaign, what is the chance that it will be successful? All these examples give us an idea of probability. It can be defined as the likelihood or chance that a particular event will occur.

9.2 PROBABILITY THEORY

In Chapter 1, a brief description of descriptive statistics and inferential statistics was given. The theory of probability is considered the basis of inferential statistics. It is interesting to know that the theory of probability is often associated with the gamblers of Europe. They were interested to know about how one could win in a game of chance. A French gambler, Chevalier de Mere, sought the help of Pascal to ascertain the probability of winning at a certain game of chance. The role of European gamblers cannot be ignored in widening the interest in the theory of probability. In fact, gambling models provide good examples of probability and its assessment. This is because dice, cards or roulette wheels are usually involved in games of chance. If cheating is not resorted to by any participant in such a game, it is reasonable to assume that the mechanical devices would provide equally likely outcomes. In view of this, it is possible for us to compute probabilities of winning at these games.

Turning to the systematic development of the theory of probability, we find that some leading mathematicians of Europe have made outstanding contribution in this area. Mention may be made of Galileo (1560–1642), Pierre de Fermat (1601–1665), Blaise Pascal (1623–1662) and Abraham de Moivre (1667–1754). Today, we find that the theory of probability stands out as a very important tool in the analysis of situations where element of uncertainty exists.

In business, there are several situations where management does not have adequate information and, yet, it becomes necessary to decide about the problem confronting it. In other words, decisions are taken with some uncertainty. We can visualize certain situations where the theory of probability can be very helpful. Some such situations are listed below.

An entrepreneur is thinking of starting manufacture of a new product. He is not sure whether the product would be well received in the market. He would like to know as to what the chances of the product being successful are. For this, he needs relevant data on a number of factors, on the basis of which he fixes a certain probability for this new venture.

A business firm has to set sale targets for different states/cities in the country. Here, its sales manager would ascertain the average monthly sales for a certain territory. For this, he has to compute probability.

An ongoing factory is in a dilemma as to the amount of investment it has to make to set up a new plant. The management has to gather some relevant information about the prospective demand for the product in the next few years. If the prospects seem to be good, i.e., the probability of it being successful is high, it would set aside a large amount for investment. In contrast, if its assessment shows that prospects are just moderate, it will go in for a small investment. Hence, the role of probability becomes clear in such a situation.

In order to understand various aspects of probability, it is necessary for us to be familiar with some basic terminology that is frequently used in any discussion on probability.

9.3 BASIC TERMINOLOGY IN PROBABILITY

The basic terminology in probability are *Experiments*, *Events* and *Sample Space.*

Experiments and Events

An experiment is an activity that results in one and only one outcome out of a set of disjoint outcomes, where an outcome cannot be predicted with certainty. Consider an experiment of tossing a die. There are 6 outcomes and we do not know which outcome will come up when we throw it. We usually assign a probability of 1/6 to each of the 6 outcomes.

We must first agree on the possible outcomes of an experiment. Take another experiment of tossing a coin. Here, we know that there are two possible outcomes, viz. head and tail and exclude the possibility of the coin standing on its edge.

The possible outcomes of an experiment are known as events. Suppose we toss two coins, then there are four outcomes:

(i) (H, H) **(ii)** (H, T) **(iii)** (T, H) **(iv)** (T, T)

In other words, these are four events, which may be denoted by E_1, E_2, E_3, and E_4. Using this symbol, we can say that getting two heads is E_1 and two tails is E_4. Getting one head and one tail is indicated by E_2 and E_3. Let this be known as E_5. Thus, we find E_5 is composed of two events E_2 and E_3.

The events E_1, E_2, E_3, and E_4 are the examples of simple events. A point worth noting is that a simple event cannot be decomposed into a combination of other events. In contrast, E_5 which can be decomposed in two events, viz. E_2 and E_3 is called a compound event, which is an aggregate of simple events. Similarly, an aggregate of simple events gives us all possible outcomes of an experiment.

Sample Space

The sample space is a set of all possible outcomes of an experiment. If we toss a fair coin, the sample space is

$$S = [\text{head, tail}]$$

Let us call these simple events by the term sample points. Similarly, if we toss an unbaised die, there exist six possible outcomes. The listing of all possible outcomes of our experiment is also the sample space:

$$S\,[1, 2, 3, 4, 5, 6]$$

In an experiment of drawing a card from a deck of playing cards, as there are 52 possible outcomes, the sample space has 52 points.

9.4 THREE TYPES OF PROBABILITY

There are three different approaches that can be used to study probability theory:

1. The classical approach
2. The relative frequency approach
3. The subjective approach

These approaches are discussed below.

The Classical Probability Approach

There are three basic conditions of the classical approach that need to be explained.

1. Equally Likely Events This means that there is an assumption that there is symmetry and homogeneity in the occurrence of events. As the theory is based on abstract reasoning and not on experiments, such an assumption is necessary. For example, the assumption of a fair coin or a fair die is very basic without which we cannot say that the probability of getting a head or a tail is ½.

2. Collectively Exhaustive This means that both favourable events and unfavourable events together exhaust all the events. This implies that the number of favourable events or unfavourable events cannot be greater than the overall number of events.

3. Mutually Exclusive This means that one and only one event takes place at a time. Take the case of tossing a coin. There are two possible outcomes, head and tail. On any toss, we may get either a head or a tail but not both. Thus, the events head and tail are mutually exclusive. For example, if two coins are tossed what is the probability of getting at least one head? There are four possibilities—HH, HT, TH and TT.

The list of four events is collectively exhaustive and each one of them is mutually exclusive. Since the total number of events is four, each of these four events has a probability of ¼. Three events, excluding the fourth one TT, have at least one H. Hence, the probability of getting at least one head is ¾.

Classical probability is also called a priori probability because we can state the answer in advance (a priori) without even tossing a coin or rolling a die or drawing a card. This is based on our assumption of fair coins, unbiased dice or decks of cards. We are not required to perform any experiments before making any statements about fair coins, unbiased dice or decks of cards.

Unfortunately, the world is not so orderly as is assumed by the classical approach in case of fair coins, unbiased dice or decks of cards. The world situation, which is rather disorderly and unstructured, makes it difficult for decision-making in business. Thus, we find that the classical approach has a serious limitation so far as the business world is concerned. As such, we may have to find a better way to define probability than the classical approach.

The Relative Frequency Approach

This approach is based on statistical data. In the 19th century, British statisticians became interested in calculating risk involved in life insurance and commercial insurance. For this purpose, they used census data on births and deaths. This approach of using statistical data is now called the relative frequency approach. Here, probability is defined as the proportion of times an event occurs in the long run when the conditions are stable. Alternatively, probability is defined as the observed relative frequency of an event in a very large number of events.

Example 9.1 Consider an experiment of tossing a fair coin. There are two possible outcomes—head (H) and tail (T). If this experiment is repeated 300 times, which is a fairly large number, then the relative frequency tends to be stable. On the other hand, initially, there are large fluctuations but as the experiment continues the fluctuations decrease. This can be shown by the following figure.

From Fig. 9.1, we find that amplitude of fluctuations, which are initially high gradually decrease as n increases and, in our present case, tends to fluctuate around 0.5 value.

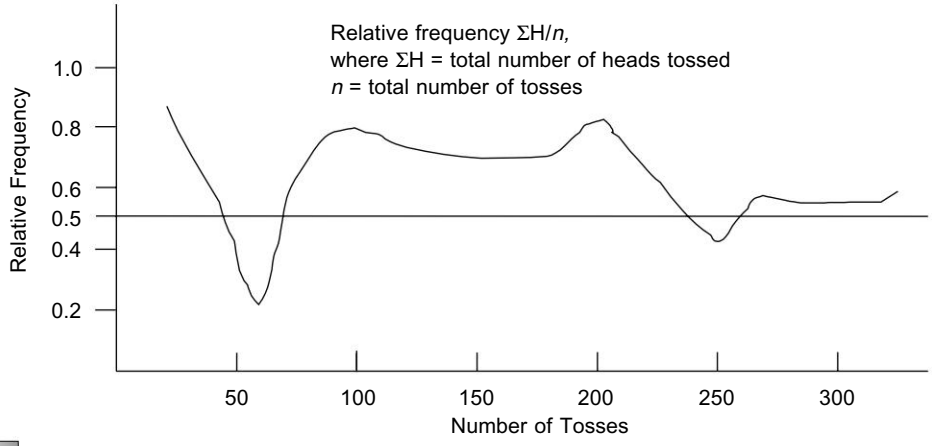


Fig. 9.1 **An Example of Relative Frequencies**

Based on this line of reasoning, we construct an abstract mathematical model of this experiment:

$$P(A) = m/n$$

where A is an event of getting head

m = Number of times the event occurs

n = Number of times the experiment is performed

It may be noted that the probability of event A, P(A), and the relative frequency of event A, m/n, are not the same thing. However, where n is large and P(A) is not known, m/n is used as an estimate of P(A), and is frequently called the probability of A. In any case, the relative frequency cannot exceed 1. Hence, we can write it $m/n \leq 1$.

When the number of occurrences of heads (H) is zero, then m becomes zero. This means $m/n = 0$.

We can now write

$$0 \leq m/n \leq 1$$

Hence, we can postulate that

$$0 \leq P(A) \leq 1$$

Problem with the Relative Frequency Approach A major problem that may arise in this approach is that sometimes people use it without evaluating a sufficient number of outcomes. A firm, for example, has developed a new consumer product. It is now contemplating to introduce it in the market nationwide. However, before it does so, it wants to test market the product. On the basis of extremely limited test market, it decides to manufacture it on a large scale. Later on, it realises that the market response turned out to be very poor, resulting in a heavy loss to the firm. There is, however, some risk involved in running a test market over a long period as some competitive firm may in the meantime introduce its own similar product in the market.

The Subjective Approach

The subjective approach is based on the personal belief of a person who is to make the probability estimate.

Professor Savage says: Personalistic views hold that probability measures the confidence that a particular individual has in the truth of a particular proposition, for example, the proposition that it will

rain tomorrow. This view postulates that the individual concerned is in some ways "reasonable". However, one cannot deny the possibility that two reasonable individuals faced with the same evidence may have different degrees of confidence in the truth of the same proposition. Professor Savage uses the term *personalistic* instead of *subjective*.

Of the three approaches to probability, this approach provides the greatest flexibility. One can use whatever evidence that is available and assign probability on the basis of his own perception of a situation. This approach gives sufficient freedom to decision-makers who assign probabilities subjectively when events occur only once or a couple of times. Suppose a firm is interested in appointing a national sales manager to look after the sales function all over the country and has shortlisted four persons from whom only one is to be selected. It may assign a subjective probability to each of the four candidates before making a final choice.

As can be seen, this subjective approach can be applied even to events that have not yet occurred or to a limited number of events. It does not require an experiment with a large number of trials. Further, it is very flexible and can be applied to widely varying situations.

Having discussed these three approaches to probability, we now turn to probability rules.

9.5 PROBABILITY AXIOMS

We now look at probability axioms. These are general probability rules that hold regardless of the particular situation or kind of probability (objective or subjective). Given an experiment:

1. Each elementary event or a combination of elementary events, must have associated with it a probability greater than or equal to zero but less than or equal to 1. Thus, if A is an event within a sample space, then
$$0 \leq P(A) \leq 1$$
2. The probability of an entire sample space is 1. Thus, if S represents a entire smaple space, then
$$P(S) = 1$$
3. The probability that one or the other or both of two mutually exclusive events will occur is equal to the sum of the individual probabilities of these events. Thus,
$$P(A \text{ or } B) = P(A) + P(B)$$
when A and B are mutually exclusive events.
4. The probability of an event that does not occur is equal to 1 minus the probability of the event that occurs. Thus,
$$P(\bar{A}) = 1 - P(A)$$
where $\bar{A}$ is the non-occurrence of event A.

Example 9.2 Suppose we have a box with 3 red, 2 black and 5 white balls. Each time a ball is drawn, it is returned to the box. What is the probability of drawing:

(a) either a red or a black ball?

(b) either a white or a black ball?

Solution The probabilities of drawing the specific colour ball are:

$$P(\text{red}) = 0.3 \quad P(\text{black}) = 0.2 \quad P(\text{white}) = 0.5$$

Applying the rule 2, we find

$$P(\text{red}) + P(\text{black}) + P(\text{white}) = 0.3 + 0.2 + 0.5 = 1$$

As we want to know the probability of drawing either a red or a black ball, then the answer will be probability P(red) + P(black) = 0.3 + 0.2 = 0.5. Likewise, the probability of getting either a white ball or a black ball will be

$$P(\text{white}) + P(\text{black}) = 0.5 + 0.2 = 0.7$$

Example 9.3 There are 20, 30 and 50 heavy smokers, moderate smokers and non-smokers, respectively. Construct a frequency table and illustrate the three rules of probability.

Solution

	f	*Relative Frequency*
Heavy (H)	20	0.2
Moderate (M)	30	0.3
Non-smokers (N)	50	0.5
Total	**100**	**1.0**

Rule 1

$P(H) = 0.2 > 0$
$P(M) = 0.3 > 0$
$P(N) = 0.5 > 0$

Rule 2

$P(H) + P(M) + P(N) = 0.2 + 0.3 + 0.5 = 1.0$

Rule 3

$P(H) \text{ or } P(M) = 0.2 + 0.3 = 0.5$
$P(H) \text{ or } P(N) = 0.2 + 0.5 = 0.7$
$P(M) \text{ or } P(N) = 0.3 + 0.5 = 0.8$

Example 9.4 An educational institution has offered admission to 100 students. On an average, the institution found 20 students secure grade A, 25 students grade B, 20 students grade C and 35 students grade D. What is the probability of selecting a student who has

(a) either grade A or B and
(b) either grade C or D?

Solution

There are four outcomes (events) and the probabilities of these events are:

$P(\text{Grade A}) = 20/100 = 0.2$
$P(\text{Grade B}) = 25/100 = 0.25$
$P(\text{Grade C}) = 20/100 = 0.2$
$P(\text{Grade D}) = 35/100 = 0.35$

(a) We are interested in knowing the probability of selecting a student who has either grade A or grade B. Since they are mutually exclusive events, P(Grade A or Grade B) = P(A) + P(B) = 0.2 + 0.25 = 0.45.

(b) We also want to know the probability of selecting a student either with grade C or grade D, then P(Grade C) + P(Grade D) = 0.2 + 0.35 = 0.55.

Addition Rule for Events Not Mutually Exclusive The foregoing discussion related to the mutually exclusive events. There are situations when we find that two events can occur together. Let us take an example to explain the method for calculating probability in such a case.

Example 9.5 In a group of 200 university students, 140 are full-time (80 females and 60 males) students and 60 part-time (40 females and 20 males) students. This break-up of students is shown below:

	200 University Students		
	Full-time	*Part-time*	*Total*
Males	60	20	80
Females	80	40	120

Two events pertaining to this selection are defined as below:

Event A: The student selected is full-time.

Event B: The student selected is part-time and male.

These two events A and B are mutually exclusive as no student can be both full-time and part-time. Either he or she is a full-time or a part-time student. Show these events in the Venn diagram.

Now, introduce another event C, which is defined as 'the student selected is female'. Are the events A and C mutually exclusive? Show this in another Venn diagram.

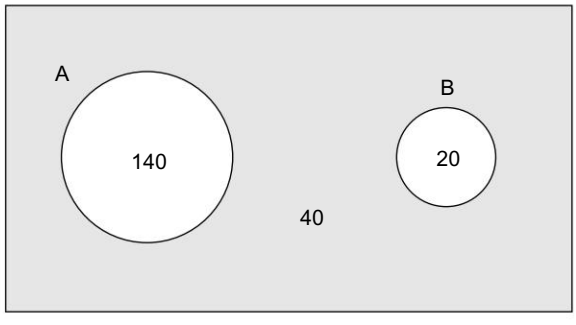


Fig. 9.2 Venn Diagram

Solution It will be seen from Fig. 9.2 that the two events A and B are shown in a sample space. As the total in the sample is 200 and as the two events account for 160 (140 full-time and 20 part-time and male), the remaining figure 40 is shown separately.

Let us now introduce another event, that is, event C, which is defined as 'the student selected is female'. Now, the question is: whether the events A and C are mutually exclusive or not? Since there are 80 full-time female students, the two events A and C are not mutually exclusive events. Figure 9.3 is the Venn diagram showing the 'intersection' of events A and C.

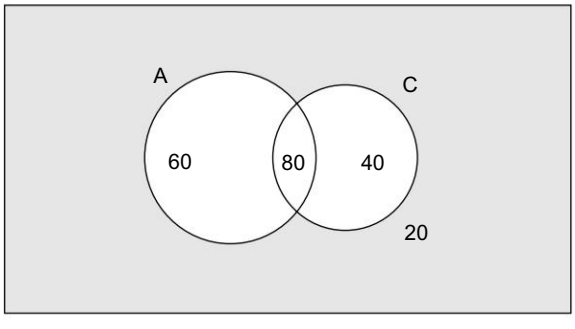


Fig. 9.3 Venn Diagram Showing Non-mutually Exclusive Events

Example 9.6 Using the sample space and event as defined in the preceding example, find the probability that the student selected is full-time or female, that is P(A or C).

Solutions

Referring to the sample space, we find that P(A) = 140 ÷ 200 = 0.7. Similarly, probability of female, that is, P(C) = 120 ÷ 200 = 0.6. Adding these two probabilities together, we get a figure of 1.3, which exceeds 1. At the same time, we know from the basic properties of probability mentioned earlier that probability numbers cannot be more than one. The question is: how has it happened? If we see more closely the sample space, we will come to know that there was double counting. We counted 80 of the 200 students twice. There are only 180 students who are full-time or female. Thus, the probability of A or C is

$$\text{P(A or C)} = \frac{n(\text{A or C})}{n(\text{S})} = \frac{180}{200} = 0.9$$

We can get the same answer as follows:

$$\text{P(A)} + \text{P(C)} - \text{P(A and C)} = \frac{140}{200} + \frac{120}{200} - \frac{80}{200}$$

$$= 0.7 + 0.6 - 0.4 = 0.9$$

We can now generalise the addition rule: Let A and B be two events defined in a sample space *S*.

$$\text{P(A or B)} = \text{P(A)} + \text{P(B)} - \text{P(A and B)}$$

It may be noted that for two mutually exclusive events, we have just to add the probability of each event A and B in order to calculate the occurrence of any one. Thus,

$$\text{P(A or B)} = \text{P(A)} + \text{P(B)}$$

This can be expanded to consider more than two mutually exclusive events:

$$\text{P(A or B or C or . . . or E)} = \text{P(A)} + \text{P(B)} + \text{P(C)} + \text{. . .} + \text{P(E)}$$

This means that when the two events are mutually exclusive, there is no double counting of sample points. In contrast, when the two events are not mutually exclusive, there will occur double counting when their individual probabilities are added together. This is shown in two Venn diagrams—Figs. 9.4 and 9.5.

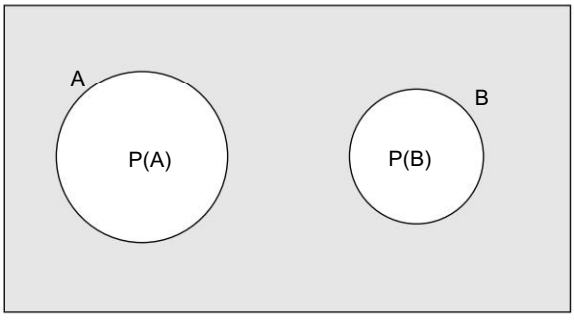


Fig. 9.4 **Mutually Exclusive Events**

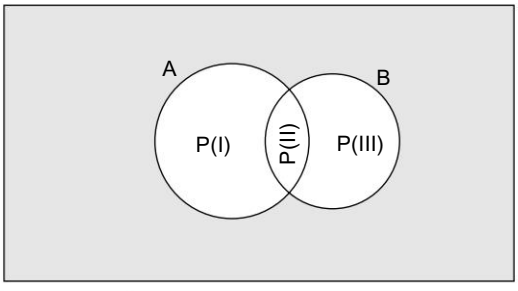


Fig. 9.5 **Non-mutually Exclusive Events**

In Fig. 9.4, events A and B are mutually exclusive. As mentioned earlier, simple addition is justified, as the total probability of the unshaded regions is sought.

In Fig. 9.5, events A and B are not mutually exclusive. The probability of the event 'A and B', that is, P(A and B) is shown in region II. This means P(A) = P(region I) + P(region II). Also P(B) = P(region II) + P(region III). Probability of A or B is the sum of the probabilities associated with the three regions. Thus,

$$P(A \text{ or } B) = P(I) + P(II) + P(III)$$

However, when we add P(A) and P(B) together, then there is double counting:

$$P(A) + P(B) = [P(I) + P(II)] + [P(II) + P(III)]$$
$$= P(I) + 2(PII) + P(III)$$

If we subtract one measure of region II from this total, we will obtain the correct value.

Example 9.7 X is a registered contractor with the government. Recently, X has submitted his tender for two contracts, A and B. The probability of getting the contract A is ¼, the contract B is ½ and both contracts A and B is 1/8. Find the probability that X will get contract A or B.

Solution

As getting contract A and contract B are mutually non-exclusive events, the required probability will be:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
$$= 1/4 + 1/2 - 1/8 = 5/8 = 0.625$$

Example 9.8 A business firm has invited applications for a managerial post. The probability that an applicant has a postgraduate qualification is 0.3 and that he has adequate work experience is 0.7, and that he has both the postgraduate qualification and work experience is 0.4. Assuming that 50 persons have applied for this managerial post in the company, find out how many applicants would have either a postgraduate degree or adequate work experience.

Solution

Let A be the event that an applicant has the postgraduate degree.

Let B be the event that the applicant has adequate work experience.

As A and B are mutually non-exclusive events, we use the addition rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
$$= 0.3 + 0.7 - 0.4 = 0.6$$

Since the number of applicants is 50 the number of applicants having postgraduate qualification or adequate work experience is $0.6 \times 50 = 30$.

9.6 PROBABILITY UNDER CONDITIONS OF STATISTICAL INDEPENDENCE

When a statistically independent event occurs, it does not have any effect on the happening of another event. There are three types of probabilities under statistical independence: **1.** Marginal, **2.** Joint, and **3.** Conditional.

Marginal Probability

Marginal probability is the simple probability of the occurrence of an event. Right in the beginning, we have given such examples pertaining to the tossing of a coin. When a coin is tossed, the probability

of getting a head is 0.5, so also in the case of getting a tail. These are known as marginal probabilities as a toss of a fair coin is a statistically independent event.

Joint Probabilities

The probability of two or more independent events occurring together is the product of their marginal probabilities.

Symbolically, $P(AB) = P(A) \times P(B)$
where $P(AB)$ = probability of events A and B occurring together or in succession
$P(A)$ = marginal probability of event A
$P(B)$ = marginal probability of event B

It will be seen that this is the multiplication rule for joint, independent events.

Example 9.9 Suppose we toss a fair coin twice. What is the probability of getting two successive heads?

Solution

$$P(H_1H_2) = P(H_1) \times P(H_2) = 0.5 \times 0.5 = 0.25$$

Obviously, the probability of getting two successive tails is also the same. Since P(T) = P(H) = 0.5. If there are three tosses of a fair coin, then the joint probability of getting three successive heads will be:

$$P(H_1H_2H_3) = P(H_1) \times P(H_2) \times P(H_3) = 0.5 \times 0.5 \times 0.5 = 0.125$$

Example 9.10 Suppose we have an unfair coin that has P(H) = 0.7 and P(T) = 0.3. What is the probability of getting three successive heads on tossing the coin three times?

Solution

$$P(H_1H_2H_3) = P(H_1) \times P(H_2) \times P(H_3) = 0.7 \times 0.7 \times 0.7 = 0.343$$

Again, we ask: what is the probability of getting three successive tails on tossing the coin three times?

$$P(T_1T_2T_3) = P(T_1) \times P(T_2) \times P(T_3) = 0.3 \times 0.3 \times 0.3 = 0.027$$

It may be noted that the two joint probabilities add up to 0.343 + 0.027 = 0.37 only and not to 1. This is because the events $H_1\ H_2\ H_3$ and $T_1\ T_2\ T_3$ do not form a collectively exhaustive list, though they are mutually exclusive in the sense that if one event occurs then the other event cannot occur.

Let us take a few more examples.

Example 9.11 What is the probability of getting tail, head, and tail on three successive tosses of a fair coin?

Solution

$$P(T_1H_2T_3) = P(T_1) \times P(H_2) \times P(T_3) = 0.5 \times 0.5 \times 0.5 = 0.125$$

We will get the same answer in the case of joint probability of $P(H_1T_2H_3)$.

Example 9.12 What is the probability of getting at least two tails on three successive tosses of a fair coin?

Solution

Now the sequence of three tosses could be one of the following:

$H_1\ H_2\ H_3$	$H_1\ T_2\ H_3$
$H_1\ T_2\ T_3$ (two tails)	$T_1\ H_2\ H_3$
$T_1\ T_2\ H_3$ (two tails)	$T_1\ T_2\ T_3$ (three tails)
$H_1\ H_2\ T_3$	$T_1\ H_2\ T_3$ (two tails)

Out of total eight outcomes, we find that at least two tails occur four times. As the probability of any of the three successive tosses is 0.5, probability of getting at least two tails is

$$P(H_1\ T_2\ T_3) + P(T_1\ T_2\ H_3) + P(T_1\ T_2\ T_3) + P(T_1\ H_2\ T_3) = 0.125 + 0.125 + 0.125 + 0.125 = 0.5$$

We will get the same answer in the case of the joint probability of at least two heads in three successive tosses.

Example 9.13 What is the probability of getting three heads or three tails on three successive tosses?

Solution

$$P(H_1\ H_2\ H_3 \text{ or } T_1\ T_2\ T_3) = P(H_1\ H_2\ H_3) + P(T_1\ T_2\ T_3)$$
$$= 0.125 + 0.125 = 0.25$$

Since there can be only eight outcomes of which only one can be three successive heads and one can be three successive tails, each outcome has a joint probability of 0.125 as the total eight outcomes must be equal to 1.

Probability Tree Diagrams In order to have a better understanding of these examples, we may construct a probability tree. Figure 9.6 shows the outcome of tossing a fair coin once. We can have only two possible outcomes—a head or a tail.

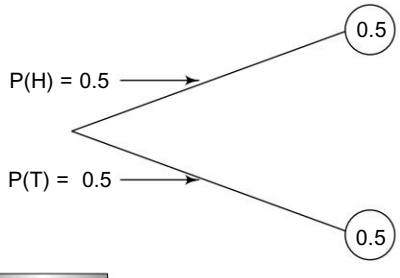


Fig. 9.6 **Example of Probability Tree**

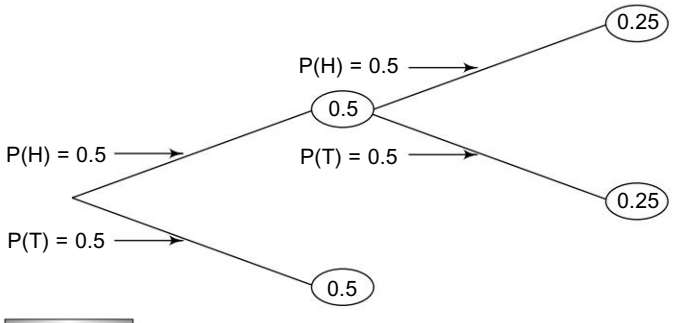


Fig. 9.7 **Probability Tree of a Partial Second Toss**

Assuming that the first toss is a head, then in the second toss the outcome could be either a head or a tail. This is shown in Fig. 9.7.

Now, we assume that the outcome of first toss is tail. In this situation, the second toss must originate from tail. This provides two more branches to the tree as shown in Fig. 9.8.

We may further extend the tree to depict the outcomes of the third toss. We repeat the same process, as a result we get what is depicted in Fig. 9.9.

It may be noted that when we toss once, we have two possible outcomes, when we toss a coin twice, we have four possible outcomes and when we toss it thrice, then we have eight possible outcomes.

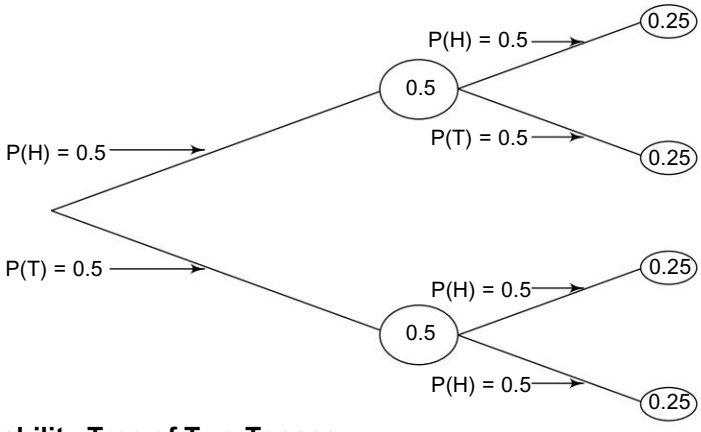


Fig. 9.8 **Probability Tree of Two Tosses**

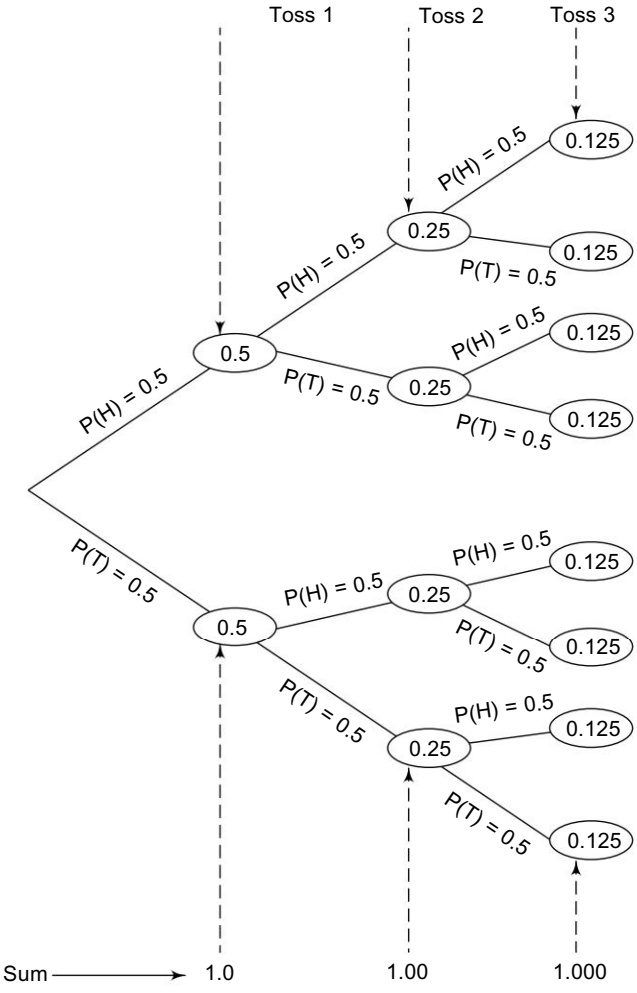


Fig. 9.9 **Probability Tree of Three Tosses**

Conditional Probabilities

The discussion so far was confined to two types of probabilities, marginal or unconditional probability and joint probability. Under statistical independence, only one type of probability remains to be discussed. This is known as the conditional probability.

Symbolically, conditional probability is written as P(A/B) which means that probability of event A, given that event B has occurred.

This appears to be contradictory. It may be recalled that independent events are those events whose probabilities are not affected by the occurrence of each other. This means that P(A/B) = P(A). Let us take an example to explain this.

Example 9.14 Suppose we are asked: what is the probability that the second toss of a fair coin will result in tail, given that tail resulted on the first toss?

Solution This can be written as $P(T_2/T_1)$. It should be noted that the outcome of the first event has no influence whatsoever on the outcome of the second event since the two events are independent. The probability of a tail on the second toss is 0.5. Thus, we can write $P(T_2/T_1) = 0.5$.

We may now summarise the three types of probabilities under statistical independence as follows:

Types of Probability	*Symbol*	*Formula*
Marginal	P(A)	P(A)
Joint	P(AB)	P(A) × P(B)
Conditional	P(B/A)	P(B)

9.7 PROBABILITY UNDER CONDITIONS OF STATISTICAL DEPENDENCE

Statistical dependence exists when the probability of an event is dependent on or affected by the occurrence of another event. As in the case of independent events, the types of probabilities under statistical dependence are:

1. Conditional **2.** Joint **3.** Marginal

Conditional Probabilities

In order to understand conditional probability under statistical dependence, let us take an example.

Example 9.15 Suppose we have an urn containing ten balls of different colours such that

2 balls are red and dotted
1 ball is green and dotted
4 balls are red and striped
3 balls are green and striped

What is the probability of drawing any particular ball from this urn?

Solution In all there are ten balls, each with equal probability of being drawn. The probability of drawing any particular ball from this urn is 0.1. To facilitate our discussion further, the above information is shown in the following table:

Table 9.1 Colour and Pattern of Ten Balls

Event	*Probability of Event*	
1	0.1	
2	0.1	Red and dotted
3	0.1	Green and dotted
4	0.1	
5	0.1	
6	0.1	
7	0.1	Red and striped
8	0.1	
9	0.1	
10	0.1	Green and striped

Example 9.16 Suppose we draw a ball from the urn and find it is red, what is the probability that it is striped?

Solution Since our problem relates to red balls, we ignore the green balls completely. In all, there are six red balls of which two are dotted and four are striped. Our problem now boils down to finding the simple probabilities of dotted and striped balls. These are shown as follows:

$$P(S/R) = 4/6 = 2/3$$

$$P(D/R) = 2/6 = \frac{1/3}{1.0}$$

It will be seen that each category of red ball has been divided by the total number of red balls. Since our problem is regarding the striped red balls, the answer is 2/3. This can be shown symbolically:

$$P(S/R) = \frac{P(SR)}{P(R)}$$

Thus, to calculate the probability of striped red balls, we divide the probability of red and striped balls by the probability of red balls. The same can be written in a generalised form as

$$P(A/B) = \frac{P(AB)}{P(B)}$$

This is the formula for calculating conditional probability under statistical dependence.

Example 9.17 What is the probability of getting a dotted ball given that it is green?

Solution We know that the total probability of green balls is 0.4 because there are four green balls out of total ten balls. To find the probability of the ball being dotted given that it is green, we have to divide the probability of green and dotted by the probability of green. Thus,

$$P(D/G) = \frac{P(DG)}{P(G)} = \frac{0.1}{0.4} = \frac{1}{4}$$

Similarly, we can determine the probability of drawing a green and striped ball given that it is green:

$$P(S/G) = \frac{P(SG)}{P(G)} = \frac{0.3}{0.4} = \frac{3}{4}$$

It may be noted again that the two probabilities ¼ + ¾ taken together add to 1.

Joint Probabilities

The formula that we used to determine conditional probability under statistical dependence is

$$P(A/B) = \frac{P(AB)}{P(B)}$$

We know that it contains one term P(AB) which, in fact, denotes joint probability. We may now rewrite this formula to determine joint probability. This can be easily done by cross multiplication.

Thus,

$$P(AB) = P(A/B) \times P(B)$$

This can be expressed as: the joint probability of events A and B is equal to the probability of event A, given that event B has already occurred, multiplied by the probability of event B.

Example 9.18 We now use this formula in our previous examples of green and red balls. Suppose we have to find the probability of red and striped ball.

Solution

$$P(SR) = P(S/R) \times P(R) = 2/3 \times 6/10 = 0.4$$

Similarly, we can calculate the joint probabilities of other events as well.

$$P(DR) = P(D/R) \times P(R) = 1/3 \times 6/10 = 0.2$$

This shows that the joint probability of dotted and red balls is equal to the product of the probability of dotted balls, given a red ball and probability of red balls. This comes to 0.2.

$$P(DG) = P(D/G) \times P(G) = 1/4 \times 4/10 = 0.1$$

This is the joint probability of dotted and green ball.

$$P(SG) = P(S/G) \times P(G) = 3/4 \times 4/10 = 0.3$$

This is the joint probability of striped and green ball.

Marginal Probabilities

The marginal probability of the event green ball can be determined by adding the probabilities of the joint events in which green ball is contained.

Symbolically, $P(G) = P(GD) + P(GS) = 0.1 + 0.3 = 0.4$

In the same manner, we can determine the marginal probability of the event red ball by adding the probabilities of the joint events in which red ball is contained.

Symbolically, $P(R) = P(RD) + P(RS) = 0.2 + 0.4 = 0.6$

So far, we have determined marginal probabilities of red balls and green balls. Likewise, we can determine the marginal probability of dotted balls and striped balls regardless of their colours. This has been attempted below:

$$P(D) = P(RD) + P(GD) = 0.2 + 0.1 = 0.3$$

$$P(S) = P(RS) + P(GS) = 0.4 + 0.3 = 0.7$$

It should be noted that these two probabilities add to 1.0 as was also in the case of the earlier two calculations. The following table summarises the probabilities under statistical dependence.

Type of Probability	*Symbol*	*Formula*
1. Marginal or unconditional	P(A)	P(A)
2. Joint	P(AB)	P(A/B) × P(B)
3. Conditional	P(A/B)	P(AB)/P(B)

9.8 REVISING PRIOR ESTIMATES OF PROBABILITIES: BAYES' THEOREM

In business, at times one finds that estimates of probabilities were made on a limited information that was available at that time. However, subsequently, some additional information becomes available. This additional information necessitates revision of the prior estimate of probability. The new probabilities are known as revised or *posterior* probabilities.

The origin of the concept of obtaining posterior probabilities with limited information is attributed to Reverend Thomas Bayes, and the basic formula for conditional probability under dependence

$$P(A/B) = P(AB)/P(B)$$

is called Bayes' Theorem

Bayes' Theorem

Bayes' theorem is an important statistical method, which is used in evaluating new information as well as in revising prior estimates of the probability in the light of that information. Bayes' theorem may be viewed as a means of transforming our prior probability of an event into a posterior probability of that event. Bayes' theorem, if properly used, makes it unnecessary to collect huge data over a long period in order to make good decisions on the basis of probabilities. We shall discuss it in greater detail in Chapter 22 in the context of decision-making.

Example 9.19 Suppose we have two machines, I and II, which are used in the manufacture of shoes. Let E_1 be the event of shoes produced by machine I and E_2 be the event that they are produced by machine II. Machine I produces 60 per cent of the shoes and machine II 40 per cent. It is also reported that 10 percent of the shoes produced by machine I are defective as against the 20 per cent by machine II. What is the probability that a non-defective shoe was manufactured by machine I?

Solution If E_1 be the event of the shoe being produced by machine I and A be the event of a non-defective shoe, our problem in symbolic terms is: $P(E_1/A)$. That is, given a non-defective shoe, what is the probability that it was produced by machine I?

From our conditional probability formulas, the probability $P(E_1/A)$ is

$$P(E_1/A) = P(E_1A)/P(A) \qquad (i)$$

But from the theorem on total probabilities, P(A) becomes

$$P(A) = P(AE_1) + P(AE_2) = P(A/E_1)\,P(E_1) + P(A/E_2)\,P(E_2)$$
$$= \Sigma P(A/E_i)\,P(E_i)$$

Substituting this result in (i) above, we get

$$P(E_1/A) = \frac{P(E_1A)}{\sum P(A/E_i)P(E_i)}$$

which may also be written as

$$P(E_1/A) = \frac{P(AE_1)P(E_1)}{\sum P(A/E_i)P(E_i)}$$

This is called Bayes' theorem.

It may be noted that $P(E_1)$ is the probability of a shoe being manufactured by machine I, whereas $P(E_1/A)$ is the probability of a shoe being produced by machine I, given that it is a non-defective shoe. The probability $P(E_1)$ is called *prior probability* and $P(E_1/A)$ is called *posterior probability*.

Let us set up a table to calculate the probability that a non-defective shoe was produced by machine I.

Table 9.2 Computation of Posterior Probabilities

Event	*Prior* $P(E_i)$	*Conditional* $P(A/E_i)$	*Joint* $P(E_iA)$	*Posterior* $P(E_i/A)$
(1)	(2)	(3)	(4)	(5) = (4)/P(A)
Machine I (E_1)	0.6	0.9	0.54	0.54/0.86 = 0.63
Machine II (E_2)	0.4	0.8	0.32	0.32/0.86 = 0.37
Total	**1.0**		**P(A) = 0.86**	**1.00**

On the basis of the above table we can say that given a non-defective shoe, the probability that it was produced by machine I is 0.63 and the probability that it was produced by machine II is 0.37. We can see that there is some revision in the prior probabilities when we apply Bayes' theorem.

A Problem with more than Two Elementary Events The foregoing problem related to two elementary events. Let us take a problem having three elementary events.

Example 9.20 A manufacturing firm is engaged in the production of steel pipes in its three plants with a daily production of 1,000, 1,500 and 2,500 units respectively. According to the past experience, it is known that the fractions of defective pipes produced by the three plants are respectively 0.04, 0.09 and 0.07. If a pipe is selected from a day's total production and found to be defective, find out **(a)** from which plant the defective pipe has come, and **(b)** what is the probability that it has come from the second plant?

Solution Let the probabilities of the possible events be

$P(E_1) = 1{,}000/(1{,}000 + 1{,}500 + 2{,}500) = 0.2$—probability that a pipe is manufactured in plant A.

$P(E_2) = 1{,}500/(1{,}000 + 1{,}500 + 2{,}500) = 0.3$—probability that a pipe is manufactured in plant B.

$P(E_3) = 2{,}500/(1{,}000 + 1{,}500 + 2{,}500) = 0.5$—probability that a pipe is manufactured in plant C.

Let P(D) be the probability that a defective pipe is drawn. Given that the proportions of the defective pipes coming from the three plants are 0.04, 0.09 and 0.07 respectively, these are, in fact, the conditional probabilities: $P(D/E_1) = 0.04$; $P(D/E_2) = 0.09$; and $P(D/E_3) = 0.07$.

Now we can multiply prior probabilities and conditional probabilities in order to obtain the joint probabilities.

Joint probabilities are

Plant A $0.04 \times 0.2 = 0.008$

Plant B $0.09 \times 0.3 = 0.027$

Plant C $0.07 \times 0.5 = 0.035$

Now we can obtain posterior probabilities by the following calculations:

Plant A $$\frac{0.008}{0.008 + 0.027 + 0.035} = 0.114$$

Plant B $\quad \dfrac{0.027}{0.008 + 0.027 + 0.035} = 0.386$

Plant C $\quad \dfrac{0.035}{0.008 + 0.027 + 0.035} = 0.500$

The above information resulting into posterior probabilities is summarized in Table 9.3.

Table 9.3 Computation of Posterior Probabilities

Event (1)	*Prior* $P(E_i)$ (2)	*Conditional* $P(E_1/E_i)$ (3)	*Joint* $P(E_iE)$ (4)	*Posterior* $P(E_i/E)$ (5) = (4)/P(E)
E_1	0.2	0.04	$0.04 \times 0.2 = 0.008$	0.008/0.07 = 0.11
E_2	0.3	0.09	$0.09 \times 0.3 = 0.027$	0.027/0.07 = 0.39
E_3	0.5	0.07	$0.07 \times 0.5 = 0.035$	0.035/0.07 = 0.50
Total	**1.0**		**P(E) = 0.07**	**1.00**

On the basis of these calculations, we can say that **(a)** most probably the defective pipe has come from plant C, and **(b)** the probability that the defective pipe has come from the second plant is 0.39.

Prior Probability vs Posterior We have seen in the foregoing Tables 9.2 and 9.3 that as any additional information becomes available, it can be used to revise the prior probability. The revised probability is called the posterior probability. Management should know-how to use the additional information to revise its prior probabilities. However, before collecting any additional information, it should also assess the utility or worth of the additional information. It may, at times, find that the cost of obtaining the additional information is more than its actual worth. In such cases, obviously it is not advisable to go in for any additional information, and management should be satisfied with the prior probabilities.

Additional Examples

Example 9.21 Assume that a card is randomly selected from a deck of 52 playing cards. Find the probability in each of the following cases:

(a) Card drawn is the king.
(b) Either a heart or the queen of spades.
(c) Card drawn is a "diamond".

Solution

(a) In a playing card, there are four kings. Hence, the probability of getting a king is 4/52 or 1/13.
(b) There are 13 cards of "heart" and the queen of spades is 1. Hence, the required probability is (13 + 1)/52 or 7/26.
(c) Here, the probability of drawing a card with "diamond" is 13/52 or 1/4.

Example 9.22 Determine the probability *P* for each of the following events:

(a) At least one head appears in two tosses of a fair coin.
(b) The sum 8 appears in a single toss of a pair of fair dice.
(c) An ace, a king, a queen or ten of 'hearts' appears in drawing a single card from a deck of 52 playing cards.

Solution

(a) In two tosses of a fair coin, there can be four possibilities: HH, HT, TH and TT. It will be seen that H comes 3 times out of 4. Hence, $P = 3/4$.

(b) When a pair of fair dice is tossed, we can get 8 as follows:

(2, 6), (3, 5), (4, 4), (5, 3) and (6, 2)

As each of the six faces of one die can be associated with each of the six faces of the second die, resulting $6 \times 6 = 36$ cases. Hence, $P = 5/36$.

(c) An ace, a king, a queen and ten of "hearts" add up to $4 + 4 + 4 + 1 = 13$. Hence, the probability is $13/52 = 1/4$.

Example 9.23 A letter is chosen at random from the word 'PROFESSOR':

(a) What is the probability that it is a vowel?

(b) What is the probability that it is a 'S'?

Solution

(a) The word PROFESSOR contains in all 9 letters of which 3 are vowels: O, E and O. Hence, the probability that a letter is a vowel is $3/9 = \frac{1}{3}$.

(b) The letter 'S' appears twice in PROFESSOR. Hence, the probability that the letter is S is 2/9.

Example 9.24 Suppose a fair die has its even numbered faces painted red, and the odd number faces painted white. Consider the experiment of rolling the die once and the events

A = (2 or 3 shows up)

B = (A red face shows up)

Find the following probabilities:

(a) P(A) **(b)** P(B) **(c)** P(AB) **(d)** P(A/B) **(e)** P(A or B)

Solution Since a fair die has six number—1 to 6.

As such each number has $\frac{1}{6}$ probability of occurrence.

(a) Hence P(A), being 2 or 3 showing up is $\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$, i.e., $\frac{1}{3}$.

(b) The total number of faces painted red is 3 as the die has 6 numbers.

Hence, P(B), i.e., where a red face shows up is $\frac{3}{6} = \frac{1}{2}$.

(c) P(AB) is the joint probability of A and B.

Hence, $P(AB) = P(A) \times P(B)$

$$= \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$$

(d) $$P(A/B) = \frac{P(AB)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} \quad \text{or} \quad \frac{1}{6} \times \frac{2}{1} \quad \text{or} \quad \frac{1}{3}$$

(e) $$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
$$= \frac{1}{3} + \frac{1}{2} - \frac{1}{6}$$
$$= \frac{2+3-1}{6} = \frac{4}{6} = \frac{2}{3}$$

Example 9.25 A, B and C bidding for a contract. It is believed that A has exactly half a chance that B has; B, in turn, has 4/5th as likely as C has to gain the contract. What is the probability for each to win the contract?

Solution Assuming that the probability of C to gain the contract is x.

Then,

Probability of B to win is 4/5 of $x = 4x/5$

Probability of A to win is 1/2 of $4x/5 = 4x/10$

Now $4x/10 + 4x/5 + x = 1$ (Since the total of three probabilities should be 1.)

or $(20x + 40x + 50x)/50 = 1$

or $110x = 50$

$\therefore$ $x = 50/110$ or $5/11$

Hence, the probabilities to win for C, B and A are

$C(x) = 5/11$

$B(4x/5) = 4 \times 5/11)/5$

$= (20/11)/5 = 4/11$

$A = (4x/10)\ (4 \times 5/11)/10$

$= (20/11)/10 = 2/11$

These probabilities can also be written as C = 0.454 B = 0.364 A = 0.182.

Example 9.26 Three salesmen, A, B and C have been given a target of selling 10,000 units of a particular product, the probabilities of their achieving their targets being respectively 0.25, 0.30 and 0.50. If these three salesmen try to sell the product, find the probability of success of only one salesman and failure of the other two.

Solution Probabilities are:

A	B	C
0.25	0.30	0.50

When A succeeds and B and C do not succeed, then

$P = 0.25 \times (1 - 0.30)\ (1 - 0.50)$

$= 0.25 \times 0.70 \times 0.50$

$= 0.0875$

When B succeeds and A and C do not succeed, then

$P = 0.30 \times (1 - 0.25)\ (1 - 0.50)$

$= 0.1125$

When C succeeds and A and B do not succeed, then

$P = 0.50 \times (1 - 0.25)\ (1 - 0.30)$

$= 0.2625$

Hence, the required probability that one of them succeeds and the other two do not succeed is

$$P = 0.0875 + 0.1125 + 0.2625$$
$$= 0.4625$$

Example 9.27 A sub-committee of 6 members is to be formed out of a group consisting of 7 men and 4 women. Calculate the probability that the sub-committee will consist of

(i) exactly 2 women; and **(ii)** at least 2 women.

Solution

(i) Out of 11 persons (7 men and 4 women) a sub-committee of 6 persons can be formed in

$$^{11}c_6 = \frac{11!}{(11-6)!6!} = \frac{11\times10\times9\times8\times7}{5\times4\times3\times2\times1} = 462 \text{ ways}$$

This is the exhaustive number of ways a sub-committee can be formed.

Number of ways for the sub-committee to consist of 4 men and 2 women is

$$^7c_4 \times {}^4c_2 = \frac{7!}{(7-4)!4!} \times \frac{4!}{(4-2)!2!}$$

$$= \frac{7\times6\times5}{3\times2\times1} \times \frac{4\times3\times2}{2\times2} = 35\times6$$

Hence, the probability is $\frac{35\times6}{462} = \frac{5}{11}$.

(ii) A sub-committee consisting of at least 2 women means 4 men and 2 women, 3 men and 3 women, and 2 men and 4 women.

For 4 men and 2 women probability is 5/11 as in (i)

For 3 men and 3 women

$$^7c_3 \times {}^4c_3 = \frac{7!}{(7-3)!3!} \times \frac{4!}{(4-3)!3!} = \frac{7\times6\times5}{3\times2} \times \frac{4\times3}{3\times2} = 140$$

Hence, probability is 140/462 or 10/33

For 2 men and 4 women

$$^7c_2 \times {}^4c_4 = \frac{7!}{(7-2)!2!} \times \frac{4!}{4!} = \frac{7\times6}{2\times1} \times \frac{4!}{4!} = 21$$

Hence, probability is 21/462 = 1/22.

Probability of having at least 2 women is

$$5/11 + 10/33 + 1/22 = \frac{30+20+3}{66} = 53/66$$

Example 9.28 Two computers A and B are to be marketed. A salesman who is assigned a job of finding customers for them has 60 per cent and 40 per cent chances respectively of succeeding in case of computers 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 the computer A has been sold?

Solution

Let A be the event that the salesman is able to sell computer A.

Let B be the event that the salesman is able to sell computer B.

Given P(A) = 0.60 and P(B) = 0.40 and that the two events A and B are independent.

$$P(AB) = P(A).\ P(B)$$
$$= 0.60 \times 0.40 = 0.24$$

Now, probability of selling at least one computer is given by

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$
$$= 0.60 + 0.40 - 0.24 = 0.76$$

We have to find out P(A) given P(A or B).

$$P[A/P(A \text{ or } B)] = \frac{P(A)}{P(A \text{ or } B)}$$

$$= \frac{0.60}{0.76}$$

$$= 0.7895$$

Example 9.29 A manufacturing firm receives shipments of machine parts from two suppliers A and B. Currently, 65 per cent of parts are purchased from supplier A and the remaining from supplier B. The past record shows that 2 per cent of the parts supplied by A are found defective, whereas 5 per cent of the parts supplied by B are found defective. On a particular day the machine breaks down because a defective part is fitted to it.

Given the information that the part was bad, using Bayes' theorem find the probability that it was supplied by supplier B.

Solution We have to use Bayes' theorem to work out the required probability. The necessary calculations are shown in the following table.

Calculation of Probability

Supplier (1)	*Prior Probability* (2)	*Conditional Probability* (3)	*Joint Probability* (4) = (2) × (3)	*Posterior (Revised) Probability* (5) = (4)/0.035
A	0.65	0.02	0.0130	$\frac{0.0130}{0.0305} = 0.43$
B	0.35	0.05	0.0175	$\frac{0.0175}{0.0305} = 0.57$
Total			0.0305	1.00

On this basis, we come to the conclusion that the probability of getting a part of shipment that was bad and from supplier B was 0.57.

Example 9.30 If a machine is set up correctly it produces 90 per cent good items; if it is incorrectly set up then it produces 10 per cent good items. Chances for a setting to be correct and incorrect are in the ratio of 7:3. After a setting is made, the first two items produced are found to be good items. What is the chance that the setting was correct?

Solution

Let E_1 be the event that the machine is set up correctly.
Let E_2 be the event that the machine is set up wrongly.
Let E be the event that the first two items are good.
We know as given in the problem

$$P(E_1) = 0.7$$
$$P(E_2) = 0.3$$
$$P(E/E_1) = 0.9$$
$$P(E/E_2) = 0.1$$

We have to find out the conditional probability E_1/E.
Required probability will be

$$P(E_1/E) = \frac{P(E/E_1) \cdot P(E_1)}{P(E/E_1) \cdot P(E_1) + P(E/E_2) \cdot P(E_2)}$$

$$= \frac{(0.9) \cdot (0.7)}{(0.9 \times 0.7) + (0.1 \times 0.3)}$$

$$= \frac{0.63}{0.63 + 0.03} = \frac{0.63}{0.66} = 0.95$$

Probability that the setting of the machine was correct is 0.95.

Example 9.31 A box contains 8 white balls, 6 red balls and 4 green balls. Find the probability that they are drawn in the order white, green and red.

(a) When the ball is replaced after the draw.

(b) When the ball is not replaced after the draw.

Solution

(a) When the ball is replaced after the draw, this is the case of independent events.

$$P(WGR) = P(W) \cdot P(G) \cdot P(R)$$
$$= 8/18 \times 4/18 \times 6/18$$
$$= 192/5832$$
$$= 0.0329$$

(b) When the ball is not replaced, then it is the case of dependent events.

$$P(WGR) = P(W) \cdot P(G/W) \cdot P(R/WG)$$
$$= 8/18 \times 4/17 \times 6/16$$
$$= 192/4896$$
$$= 0.0392$$

Where P(R/WG) is the conditional probability of getting a red ball given that white and green balls have already been drawn.

Example 9.32 An urn contains 5 white, 3 black and 2 red balls. If 3 balls are drawn at random, find the probability that (a) all 3 are white (b) all 3 are black (c) 2 are white and 1 is red (d) 2 are black and 1 is red, and (e) one ball of each colour is drawn.

Solution It may be noted that in the first draw, the total balls would be 10, while in the second and the third draws, total balls would be 9 and 8, respectively.

(a) $P(W_1)\ P(W_2)\ P(W_3)$

$$= \frac{5}{10} \times \frac{4}{9} \times \frac{3}{8}$$

$$= \frac{60}{720} = \frac{1}{12} = 0.083$$

(b) $P(B_1)\ P(B_2)\ P(B_3)$

$$= \frac{3}{10} \times \frac{2}{9} \times \frac{1}{8}$$

$$= \frac{6}{720} = \frac{1}{120} = 0.0083$$

(c) $P(W_1)\ P(W_2)\ P(R_1)$

$$= \frac{5}{10} \times \frac{4}{9} \times \frac{2}{8}$$

$$= \frac{40}{720} = \frac{1}{18} = 0.056$$

(d) $P(B_1)\ P(B_2)\ P(R_1)$

$$= \frac{3}{10} \times \frac{2}{9} \times \frac{2}{8}$$

$$= \frac{12}{720} = \frac{1}{60} = 0.0167$$

(e) $P(W_1)\ P(B_1)\ P(R_1)$

$$= \frac{5}{10} \times \frac{3}{9} \times \frac{2}{8}$$

$$= \frac{30}{720} = \frac{1}{24} = 0.0417$$

Example 9.33 One bag contains 4 white and 2 black balls. Another bag contains 3 white and 5 black balls. If one ball is drawn from each bag, find the probability that (a) both are white, (b) both are black, (c) one is white and one is black.

Solution

(a) $P(W_1)\ P(W_2)$

$$= \frac{4}{6} \times \frac{3}{8}$$

$$= \frac{12}{48} = \frac{1}{4} = 0.25$$

(b) $P(B_1)\ P(B_2)$

$$= \frac{2}{6} \times \frac{5}{8}$$

$$= \frac{5}{24} = 0.208$$

(c) As regards 'one is white and one is black' it may be noted that the first ball drawn may be white and the second draw may give a black ball. It can also be just the opposite: the first draw may give a black ball and the second draw may give a white ball. Hence, we may put this as follows:

$$= P(W)\,P(B) + P(B)\,P(W)$$
$$= \frac{4}{6} \times \frac{5}{8} + \frac{2}{6} \times \frac{3}{8}$$
$$= \frac{20}{48} + \frac{6}{48} = \frac{26}{48} = \frac{13}{24}$$

Example 9.34 There are three brands, X, Y and Z, of an item available in the market. A consumer chooses only one of them at a time. The probabilities that he buys brands X, Y and Z are 0.20, 0.16 and 0.45, respectively.

(a) What is the probability that he does not buy any of the brands?

(b) Given that a customer buys some brand, what is the probability that he buys Z brand?

Solution

(a) As the customer does not buy any brand, the total of three probabilities is to be subtracted from 1:

$$1 - (0.20 + 0.16 + 0.45) = 1 - 0.81$$
$$= 0.19$$

(b)
$$\frac{P(Z)}{P(X) + P(Y) + P(Z)} = \frac{0.45}{0.20 + 0.16 + 0.45}$$
$$= \frac{0.45}{0.81}$$
$$= 0.555$$

Example 9.35 An MBA applies for a job in the two firms, X and Y. The probability of his being selected in firm X is 0.7 and of being rejected in firm 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 The probability of his selection is:

in Firm X = 0.7

in Firm Y = 1 – 0.5 = 0.5

Let A denote that the MBA will be selected by firm X.

Let B denote that the MBA will be rejected by firm Y.

Then,
$$P(A) = 0.7 \quad \text{and}$$
$$P(B) = 0.5$$
$$P(\bar{A}) = 1 - 0.7 = 0.3$$
$$P(\bar{B}) = 1 - 0.5 = 0.5$$
$$P(\bar{A} \text{ or } \bar{B}) = 0.6$$
$$P(A \text{ or } B) = P(A) + P(B) - (1 - 0.6)$$
$$= 0.7 + 0.5 - 0.4$$
$$= 0.8$$

The probability that the MBA being selected in one of the firms is 0.8.

Example 9.36 A candidate is selected for interviews for Management trainees for three 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 a job in at least one of the companies?

Solution The probabilities of his getting a job in each of the companies are:

$$1^{st} \text{ company}, \quad P(1^{st}) = \frac{1}{12}$$

$$2^{nd} \text{ company}, \quad P(2^{nd}) = \frac{1}{15}$$

$$3^{rd} \text{ company}, \quad P(3^{rd}) = \frac{1}{10}$$

The probability that the candidate does not get a job in any of the three companies is:

$$1^{st} \text{ company}, \quad 1 - \frac{1}{12} = \frac{11}{12}$$

$$2^{nd} \text{ company}, \quad 1 - \frac{1}{15} = \frac{14}{15}$$

$$3^{rd} \text{ company}, \quad 1 - \frac{1}{10} = \frac{9}{10}$$

As these events are independent, the probability that the candidate does not get any job in the three companies is

$$\frac{11}{12} \times \frac{14}{15} \times \frac{9}{10} = \frac{1386}{1800} = 0.77$$

Hence, the probability that he gets a job in at least one of the companies is 1 – 0.77 = 0.23.

Example 9.37 Assume that a factory has two machines. Past records show that Machine 1 produces 30% of the items of output and Machine 2 produces 70% of the items. Further, 5% of the items produced by Machine 1 are defective, and only 1% produced by Machine 2 are defective. If a defective item is drawn at random, what is the probability that it was produced by (i) Machine 1 or (ii) Machine 2?

Solution

Machine 1 P(A) 0.3

Machine 2 P(B) 0.7

$$P(A) \times P(D/A) = 0.3 \times 0.05$$

$$P(B) \times P(D/B) = 0.7 \times 0.01 \qquad = 0.007$$

Sum of Jt. probabilities = 0.015 + 0.007

$$P(D) = 0.022$$

$$P(A/D) = 0.015/0.022 = 0.68$$

$$P(B/D) = 0.007/0.022 = 0.32$$

On the basis of the above calculations, the probability that the defective item was produced by Machine 1 comes to 0.68 and by Machine 2, 0.32.

Example 9.38 Five men, in a company of 20, are graduates. If three men are picked, out of 20, at random, what is the probability that they are all graduates? What is the probability that none is a graduate? What is the probability that at least one is a graduate?

Solution

(i) All graduates

$$P = \frac{{}^{5}c_3 \times {}^{15}c_0}{{}^{20}c_3}$$

Here, $${}^{5}c_3 = \frac{5!}{(5-3)!3!} = \frac{5\times4\times3\times2\times1}{2\times1\times3\times2\times1} = 10$$

$${}^{15}c_0 = \frac{15!}{(1500)!} = 1$$

$${}^{20}c_3 = \frac{20!}{(20-3)!(3!)}$$

$$= \frac{20\times19\times18}{3\times2} = \frac{6840}{6} = 1140$$

Hence, $$P = \frac{10\times1}{1140} = \frac{1}{114} = 0.008$$

(ii) None is a graduate

$$P = \frac{{}^{15}c_3 \times {}^{5}c_0}{{}^{20}c_3}$$

$${}^{15}c_3 = \frac{15!}{(15-3)!3!} = \frac{15\times14\times13}{3\times2} = \frac{2730}{6} = 455$$

Hence, $$P = \frac{455\times1}{1140} = 0.399$$

(iii) At least one is a graduate

P = 1 – Probability of no garaduate
= 1 – 0.399
= 0.601

Example 9.39 A problem in business statistics is given to 5 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 $\frac{1}{6}$, respectively. What is the probability that the problem will be solved?

Solution Since all the events are independent, the problem can be worked out as shown below:

1 – P (problem is not solved)

When the problem is not solved by each of the five students, it can be shown as

$$1 - P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C}) \cdot P(\bar{D}) \cdot P(\bar{E})$$

$$= 1 - \left(1-\frac{1}{2}\right)\left(1-\frac{1}{3}\right)\left(1-\frac{1}{4}\right)\left(1-\frac{1}{5}\right)\left(1-\frac{1}{6}\right)$$

$$= 1 - \left(\frac{1}{\cancel{2}} \times \frac{\cancel{2}}{\cancel{3}} \times \frac{\cancel{3}}{\cancel{4}} \times \frac{\cancel{4}}{\cancel{5}} \times \frac{\cancel{5}}{6}\right)$$

$$= 1 - \frac{1}{6}$$

$$= \frac{5}{6} \text{ or } 0.83$$

Example 9.40 A product is assembled from two components A and B. The probability of component A being defective is 0.03 and the probability of component B being defective is 0.02. What is the probability that the assembled product will not be defective?

Solution

P (component A being defective) = 0.03
P (component B being defective) = 0.02
P (any of the parts is defective)

$$= P(A) + P(B) - P(AB)$$

$$= \frac{3}{100} + \frac{2}{100} - \left(\frac{3}{100} \times \frac{2}{100}\right)$$

$$= \frac{5}{100} - \frac{6}{10000}$$

$$= \frac{500 - 6}{10000}$$

$$= \frac{494}{10000}$$

$\therefore$ Probability that the assembled part is not defective is

$$1 - \frac{494}{10000}$$

$$= \frac{9506}{10000}$$

$$= 0.95$$

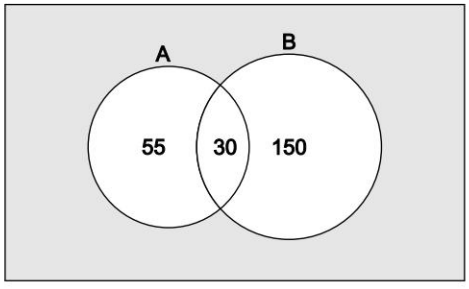
Example 9.41 In an organization, out of 250 employees, monthly salary of 55 is more than Rs 20000, and 150 of them are regular takers of Beta Brand Tea. Out of the 55 with monthly salary more than Rs 20000, 30 are regular takers of Beta Brand Tea. If a particular employee is selected, what is the probability that his monthly salary is more than Rs 20000, or he is a regular taker of Beta Brand Tea?

Solution

Total employees = 250
A Monthly salary > Rs 20000 = 55 (30 are BBT)
B Regular takers of BBT = 150

$$P(\text{Monthly salary} > \text{Rs } 20000) = \frac{55}{250} = 0.22$$

$$P(\text{Regular takers of BBT}) = \frac{150}{250} = 0.6$$



As this is a case of non-mutually exclusive events, the following formula is used.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \frac{55}{250} + \frac{150}{250} - \frac{30}{250}$$

$$= \frac{55 + 150 - 30}{250} = \frac{175}{250} = 0.7$$

Example 9.42 Five out of 100 items produced in machine A and one out of 100 items produced in machine B are found to be defective. An item drawn at random from the items produced by A and B is found to be defective. What is the probability that this item has been made in machine A, assuming that both machines produced equal number of items.

Solution This problem can be worked out using Baye's theorem. The necessary calculations are shown in the following table.

Event	*Prior $P(E_i)$*	*Conditional $P(A/E_i)$*	*Joint $P(E_iA)$*	*Posterior $P(E_i/A)$*
Machine A	0.05	0.5	0.25	0.83
Machine B	0.01	0.5	0.05	0.17
			0.30	1.00

On the basis of above table, the probability that the defective item has been made in machine A is 0.83.

GLOSSARY

A priori probability	Probability estimate made prior to receiving information.
Bayes' theorem	A rule that is used while revising probabilities of events after having obtained more information.
Classical probability rule	The method of assigning probabilities to outcomes or events of an experiment with equally likely outcomes.
Collectively exhaustive events	A list of events that represents all the possible outcomes of an experiment.

Compound (or composite) event	An event that contains more than one outcome of an experiment.
Conditional probability	The probability of event B occurring, given that event A has occurred.
Dependent events	When the occurrence of one event affects the probability of the occurrence of the other, then the two events are said to be dependent events.
Event	One of the possible outcomes of an experiment.
Experiment	A process that results in an event.
Independent events	Two events for which the occurrence of one does not change the probability of the occurrence of the other.
Joint probability	The probability of two or more events occurring together or in succession.
Marginal probability	The probability of one event without consideration of any other event.
Mutually exclusive events	Two or more events that cannot occur together.
Outcome	The result of the performance of an experiment.
Posterior probability	A probability that has been revised on the basis of new information that has become available.
Probability	A numerical measure of the likelihood that a specific event will occur.
Relative frequency of occurrence	The proportion of times that an event occurs in a very large number of trials.
Sample space	The set of all sample points or outcomes of an experiment.
Statistical dependence	The condition when the probability of a certain event is dependent on the occurrence of some other event.
Statistical independence	The condition when the occurrence of one event does not have any effect on the occurrence of another event.
Subjective probability	The probability assigned to an event by a person on the basis of his judgment as well as the information available with him.
Tree diagram	A diagram in which each outcome of an experiment is represented by a branch of a tree.
Venn diagram	A diagram showing sample space in the form of a rectangle and events as portions of that rectangle.

LIST OF FORMULAE

1. Probability of event A = P(A). A single probability refers to the probability of one particular event and is called the marginal probability.
2. $P(A) \geq 0$. Probability of any event (in this case event A) cannot be negative.

3. P(A) + P(B) + P(C) + . . . P(N) = 1, where A, B, C, . . . , N are mutually exclusive events. The sum of all possible exclusive events is unity.
4. Relative Probability $P(A) = \frac{m}{n}$, where A is an event of getting head, m is the number of times the event occurs and n is the number of times the experiment is performed.
5. P(A or B) = P(A) + P(B). Probability of either of two mutually exclusive events is the sum of their probabilities.
6. P(A or B) = P(A) + P(B) – P(AB). When the events are not mutually exclusive, probability of either A or B is the sum of the two probabilities minus the probability of A and B happening together.
7. P(AB) = P(A) × P(B), where P(AB) is the joint probability of events A and B, P(A) is the marginal probability of event A and P(B) is the marginal probability of event B. Thus, the joint probability of two events occurring together or in succession is the product of their marginal probabilities.
8. P(B/A) = P(B). In case of independent events, the conditional probability of event B, given the occurrence of event A, is simply the probability of event B.
9. P(B/A) = P(BA)/P(A) and P(A/B) = P(AB)/P(B). In case of statistically dependent events, the conditional probability of event B, given the occurrence of event A, is equal to the joint probability of events A and B divided by the marginal probability of event A. The same rule applies to the second term P(A/B).
10. P(AB) = P(A/B) × P(B) and P(BA) = P(B/A) × P(A). In case of statistically dependent events, the joint probability of events A and B is equal to the probability of event A, given probability of event B, multiplied by the probability of event B. This rule is equally applicable to the second term P(BA).

QUESTIONS

9.1 Given below are ten statements. Indicate in each case whether it is true or false.

(a) If one event is not affected by the outcome of another event, the two events are said to be mutually exclusive.

(b) An unconditional probability is also known as a marginal probability.

(c) The sample space is a set of all possible outcomes of an experiment.

(d) In classical approach to probability, one can state the outcome of an event in advance.

(e) The relative frequency of occurrence approach offers greatest flexibility in calculating probability of an event.

(f) If P(A/B) = P(B), then A and B are said to be independent events.

(g) Symbolically, P(AB) is used to denote a marginal probability.

(h) A subjective probability is just an intelligent guess regarding the occurrence of an event.

(i) When two events are not mutually exclusive, P(A or B) is the summation of P(A) and P(B).

(j) The probability of two or more independent events occurring together is the product of their marginal probabilities.

Multiple Choice Questions (9.2 to 9.17)

9.2 The probability of an occurrence of an event is known as
(a) Bayesian probability (b) conditional probability
(c) joint probability (d) marginal probability

9.3 If the outcome of one event does not influence another event, then the two events are
(a) mutually exclusive (b) dependent
(c) independent (d) both (a) and (c)

9.4 The events of tossing a coin are mutually exclusive because
(a) On any one toss it is not possible to get a head and a tail
(b) The outcome of one toss is not affected by the outcome of an earlier toss
(c) The probability of getting a head and the probability of getting a tail are the same
(d) All of these

9.5 What is the probability of a value chosen from a population is less than its median?
(a) 1.0 (b) 0.25 (c) 0.75 (d) 0.5

9.6 Symbolically, a marginal probability is
(a) P(B/A) (b) P(AB) (c) P(BA) (d) None of these

9.7 Assuming that a box contains in all ten balls of which six are green and four are red, what is the probability of drawing a red ball?
(a) 0.1 (b) 0.5 (c) 1.0 (d) 0.4

9.8 What is the probability of getting an even number when a die is tossed?
(a) $\frac{1}{3}$ (b) $\frac{1}{2}$ (c) $\frac{1}{6}$ (d) 1

9.9 What is the probability of getting more than 2 when a die is tossed?
(a) $\frac{1}{3}$ (b) $\frac{1}{2}$ (c) $\frac{2}{3}$ (d) 1

9.10 What is the probability of getting a total of 5 when a pair of die is tossed simultaneously?
(a) $\frac{1}{36}$ (b) $\frac{1}{12}$ (c) $\frac{4}{9}$ (d) $\frac{1}{9}$

9.11 It P(AB) = 0, then the two events A and B are said to be
(a) dependent (b) independent
(c) equally likely (d) none of these

9.12 Baye's theorem is useful in
(a) computing conditional probabilities (b) computing sequential probabilities
(c) revising prior probabilities (d) none of these

9.13 What is the probability of drawing an ace from a bunch of playing cards?
(a) $\frac{1}{10}$ (b) $\frac{1}{13}$ (c) $\frac{1}{52}$ (d) none of these

9.14 On the assumption that the two events A and B are mutually exclusive, P(A or B) = P(A) + P(B). How does P(A or B) change if the two events are not mutually exclusive?
(a) [P(A) + P(B)] must be multiplied by P(AB)
(b) [P(A) + P(B)] must be divided by P(AB)
(c) P(AB) must be subtracted from P(A) + P(B)
(d) P(AB) must be added to P(A) + P(B)

9.15 The probability of getting head, while tossing a coin three times is
(a) 1 (b) 0.75 (c) 0.25 (d) 0.125

9.16 What is the probability of getting three heads or three tails on three successive tosses?
(a) 0.25 (b) 0.125 (c) 0.175 (d) none of these

9.17 In the context of posterior probability, which of the following statements is *not* true?
(a) Posterior probability is a revised probability.
(b) Prior probability and conditional probability are multiplied.
(c) Conditional probability and joint probability are multiplied.
(d) The sum of all posterior probabilities is equal to the sum of all prior probabilities.

9.18 What do you understand by the term probability? Discuss its importance in business decision-making.

9.19 Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously? Support you answer with the help of an example.

9.20 Describe briefly the various schools of thought on probability. How does the concept of probability help the decision-maker to improve his decisions?

9.21 State and prove addition theorem of probability for two events.

9.22 What is classical approach to probability? What are its limitations?

9.23 What do you mean by empirical approach to probability? Describe the importance of probability in Statistics.

9.24 Distinguish, with the help of an example, between conditional probability and joint probability under conditions of statistical dependence.

9.25 What is subjective approach to probability? Give an example of its application.

9.26 State the multiplicative theorem of probability. How is the result modified when the events are independent?

9.27 Explain the concept of posterior probability.

9.28 Write a short note on mutually exclusive events.

9.29 Distinguish between prior probability and posterior probability.

9.30 State the addition and multiplication rules of probability, giving one example of each rule.

9.31 Express the following statements in the notation of the event operations:
(a) A occurs but B does not
(b) Neither A nor B occurs
(c) Exactly one of the events A or B occurs

9.32 Examine the probability assigned in each case and state what makes it improper:
(a) Concerning tomorrow's weather,
P(rain) = 0.2 P(cloudy but no rain) = 0.4 P(sunny) = 0.6
(b) Concerning your passing the Statistics course,
P(pass) = 2 P(fail) = 0.2
(c) Concerning your grades in Statistics and Economics courses,
P(A in Statistics) = 0.5 P(A in Economics) = 0.8 P(A in both Statistics and Economics) = 0.6

9.33 A letter is chosen at random from the word 'STATISTICIAN'.
(a) What is the probability that it is a vowel?
(b) What is the probability that it is a T?

9.34 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 the shop, what is the probability that the salesman will make a sale to A or B?

9.35 A husband and wife appear in an interview for two vacancies for 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

(i) both of them will be selected?

(ii) only one of them will be selected?

(iii) none of them will be selected?

9.36 According to a survey, the probability that a family owns two cars if its monthly income is greater than Rs 15,000 is 0.7. Of the households surveyed, 50 per cent had incomes over Rs 15,000 and 40 per cent had two cars. What is the probability that a family has two cars and an income over Rs 15,000 a month?

9.37 A bag contains 8 balls of which 5 are red and 3 are white. If a man selects 2 balls at random from the bag, what is the probability that he will get one ball of each colour?

9.38 If the probability that an individual suffers a bad reaction from an injection of a given serum is 0.001, determine the probability that out of 2,000 individuals, **(i)** exactly 3; and **(ii)** more than 2 individuals will suffer a bad reaction.

9.39 In the play of two dice, the thrower loses if his first throw is 2, 4, or 12. He wins if his first throw is 5 or 11. Find the ratio between his probability of losing and probability of winning in the first throw.

9.40 From a box containing 3 white and 5 black balls, 4 balls are transferred into an empty box. From this box, a ball is drawn and is found to be white. What is the probability that out of the 4 balls transferred, 3 are white and 1 black?

9.41 An unbiased coin is tossed. If the result is a head, a pair of unbiased dice is rolled and the number obtained by adding the numbers on the two faces is noted. If the result is a tail, a card from a well-shuffled pack of 11 cards 2, 3, 4, …12 is picked and the number on the card is noted. What is the probability that the noted number is either 7 or 8?

9.42 Within the union membership of a local unit of the United Automobile Workers, 90 per cent of workers are employed and 10 per cent are unemployed. What is the probability that in a random sample of 10 members, **(i)** 4 unemployed workers will be included? **(ii)** all are unemployed workers?

9.43 A publishing company has 34 employees. Of these, 20 are men (5 of whom are in senior positions) and 14 are women (3 of whom are in senior positions).

(a) Calculate the probability that an employee selected at random is in a senior position given that the employee is **(i)** a woman; **(ii)** a man.

(b) Two are to be selected at random for participation in a conference. What is the probability that **(i)** both will be men; **(ii)** both will be women; **(iii)** they will be of different sex; **(iv)** at least one will be a woman; **(v)** at least one will be a non-senior employee?

9.44 Two shipments of machine parts are received. The first shipment contains 1000 parts with 10 per cent defectives and the second shipment contains 2000 parts with 5 per cent defectives. One shipment is selected at random. Two machine parts are tested and found good. Find the probability (a posterior) that the tested parts were selected from the first shipment.

9.45 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 option	5	5	5	5

(i) What is the probability that a consumer, at random, prefers brand A?
(ii) What is the probability that a consumer prefers brand A and from Chennai?
(iii) What is the probability that a consumer prefers brand A given that he was from Chennai?
(iv) Given that a consumer preferred brand A, what is the probability that he was from Mumbai?

9.46 A company is planning to host a conference to learn how people are using the Internet. The conference planners want to ensure a good representation across age groups. According to Media Research, if 45 per cent of individuals in the age group 30–35 and 35 per cent of individuals in the age group 35–40 use Internet, then

(i) how many individuals from age group 30–35 be invited to have average attendance of 60 from this group?
(ii) how many individuals from age group 35–40 be invited to have average attendance of 50 from this group?
(iii) If 100 individuals are invited from each of the age groups given above, what is the probability that **(a)** 20 would attend the conference in the age group of 30–35? **(b)** 15 would attend the conference in the age group of 35–40?

9.47 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 conditional probability that it was produced by manufacturer A?
(ii) A part is known to have been produced by manufacturer A. What is the conditional probability that the part has no defects?
(iii) A part is known to have two or more defects. What is the conditional probability that it was manufactured by A?
(iv) A part is known to have one or more defects. What is the conditional probability that it was manufactured by B?

9.48 A doctor has decided to prescribe two new drugs to 200 heart patients as follows: 50 get drug A, 50 get drug B and 100 get both. Drug A reduces the probability of a heart attack by 35 per cent, drug B reduces the probability by 20 per cent, and the two drugs when taken together work independently. Two hundred patients were chosen so that each has 80 per cent chance of

having a heart attack. If a randomly selected patient has a heart attack, what is the probability that the patient was given both the drugs?

9.49 A company has three plants to manufacture 8,000 scooters in a month. Out of 8,000 scooters, plant I manufactures 4,000, plant II manufactures 3,000 and plant III manufactures 1,000 scooters. At plant I, 85 out of 100 scooters are rated of standard quality or better, at plant II only 65 out of 100 scooters are rated of standard quality or better and at plant III 60 out of 100 scooters are rated of standard quality or better. What is the probability that the scooter selected at random comes from **(i)** plant I, **(ii)** plant II, and **(iii)** plant III if it is known that the scooter is of a standard quality?

9.50 A personnel director has found that he can fill a certain type of position within one week 70 per cent of the time. But he finds that 60 per cent of the time, all applicants for the position are college dropout. The 40 per cent of the time, none of them is dropout. When all applicants are college dropout, the position is filled within one week only 56 per cent of the time.

(i) What is the probability that the position is filled within one week, given that the applicants are not college dropout?

(ii) Is filling the position within a week independent of whether the applicants are college drop-outs?

9.51 The market size for a new product to be launched may be in any of the states: Low (S_1), Moderate (S_2) or High (S_3) with respective prior probabilities of 0.2, 0.5 and 0.3. It is possible to conduct pre-launch test marketing, which will indicate whether the response to the product is good (G) or bad (B). The conditional probabilities of the test market response under the states of market are shown below. Find the probability:

(i) Test marketing will give a good response (G)

(ii) Given that test result was good the product will realise high market size

(iii) Given that the test result was bad the product will realise a low market size.

Test Result

State of Market	*Good*	*Bad*
Low (S_1)	0.1	0.9
Moderate (S_2)	0.4	0.6
High (S_3)	0.8	0.2

9.52 In a large business organisation, 80 per cent of the technical employees have met the training requirements for promotion, 70 per cent have met the experience requirements, and 60 per cent have met both the requirements. If an employee is selected at random from the population, find the probability that the employee will

(i) not meet either of the requirements

(ii) meet the training requirements given that he or she meets the experience requirement.

9.53 The manager of a department store collects data on the number of employees who have left each month in the last one year. The data are given below:

Employee	0	1	3	4	2	5
Month	2	3	2	1	3	1

(a) What is the probability that there will be at least one employee leaving next month?
(b) What is the probability that there will be more than two employees leaving next month?

9.54 Suppose the probability of a light in a classroom to be burnt out is 1/3. The classroom has, in all, 5 lights, and it becomes unusable if the number of working lights is less than two. What is the probability that the classroom is unusable on a random occasion?

9.55 Suppose, on an average, one house in 1000 in a certain town has fire during the year. If there are 2000 houses, what is the probability that exactly 5 houses will have fire during the year?

9.56 It is observed that a person going to petrol pump for refuelling checks air pressure of the tyres 12% times, and checks the levels of engine oil and brake oil 29% of the times. It is also observed that 7% of the persons check both air pressure and the level of oil.

(i) Calculate the probability that the person going to the petrol pump neither checks air pressure nor the level of oil.
(ii) Calculate the probability that the person checks air pressure but not the level of oil.

9.57 Consider that in your locality, out of the 5000 people residing, 1200 are above 30 years of age and 3000 are female. Out of the 1200 above 30 years, 200 are female. Suppose a person is chosen and you are told that the person is a female. What is the probability that she is above 30 years of age?

9.58 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 the bolts are expected to be defective.

9.59 A factory uses two machines, A and B, to produce a commodity. 9 per cent of the items produced in machine A and 5 per cent in machine B are found to be defective. The machines produce 500 and 700 items, respectively, per day. An item drawn, at random, from the day's production is found to be defective. What is the probability that it was produced in machine A?

9.60 In a bolt factory, machines A, B and C manufacture, respectively, 25%, 35% and 40% of the total production. Of their output, 5%, 4%, and 2% are defective bolts. If a bolt drawn at random from the product is found to be defective, what are the probabilities of it being manufactured by (i) machine A, (ii) machine B and (iii) machine C?

9.61 The chance that a salesman will convince a customer and make a sale on call is 0.7. If this salesman calls on 5 customers in a day, find the probability that he will be successful with all the 5 customers.

9.62 A bag contains 5 white and 8 red balls. Two drawings of 3 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 3 white and the second 3 red balls in each case.

9.63 Three machines turn out non-ferrous castings. Machine A produces 1%, machine B 2% and machine C 5% defective castings. Each machine produces 1/3 of the output. An inspector examines a single casting, which he determines as non-defective. Estimate the probabilities of its having been produced by each of the machines.

9.64 A portfolio consultant firm has three advisers, A, B and C, to advise its clients regarding investments in secondary market. In a particular week, the number of clients who take the advice of A, B and C and invest are 200, 180 and 120, respectively. 'A' being higher experienced, has the reputation that 90% of his clients are benefitted. The corresponding figures for 'B' and 'C' are 80% and 75%, respectively. At the end of the week, a client was selected at random, and it was found that he had not benefitted from the advice. Find the probability that he was advised by 'B'.

9.65 A problem in mathematics is given to two students X and Y. The probability that X will solve the problem is 1/3, and that Y will solve the problem is 1/4. What is the probability that the problem will be solved by at least one of them?

9.66 A bag contains 30 balls, numbered 1 to 30. One ball is drawn at random. Find the probability that the number of the ball drawn is a multiple of 5 or 7.

9.67 On an average, six people per hour use an electronic teller machine during prime shopping hours in a department store. What is the probability that

(i) Exactly six people will use the machine during a randomly selected hour.

(ii) Fewer than five people will use the machine during a randomly selected hour.

(iii) No one will use the facility in a 10 minute interval.

(iv) No one will use the facility in a 5 minute interval.

9.68 There are two boxes B_1 and B_2. B_1 contains two red balls and one green ball. B_2 contains one red ball and two green balls.

(i) A ball is randomly drawn from one of the boxes. It is found to be red. What is the probability that it was drawn from B_1?

(ii) Two balls are drawn randomly from one of the boxes without replacement. One is red and the other is green. What is the probability that they came from B_1?

(iii) A ball drawn from one of the boxes is green. What is the probability that it came from B_2?

The McGraw-Hill Companies

CHAPTER 10 PROBABILITY DISTRIBUTIONS

Learning Objectives

By the end of your work on this chapter, you should be able to

- recognise problems that can be modelled by the binomial, Poisson and normal distributions
- solve such problems with the use of the appropriate tables
- recognise when use of these distributions involves approximations in the original problem
- outline the connections between these three distributions.

Chapter Prerequisites

Before starting work on this chapter, make sure you are fully conversant with

1. the definitions of probability and the rules for calculating it
2. the ideas of frequency distributions and histograms
3. the concepts of mean and standard deviation.

10.1 INTRODUCTION

In Chapter 3, we had seen that a frequency distribution is a listing of the values, which a variable takes together with a count (i.e. the frequency of its occurrence). In Chapter 9, we encountered some experiments where the outcomes were categorical. We found that an experiment results in a number of possible outcomes and discussed how the probability of the occurrence of an outcome can be determined.

In this chapter, we shall extend our discussion of probability theory. Our focus is on the probability distribution, which describes how probability is spread over the possible numerical values associated with the outcomes.

Since the concept of a *random variable* is fundamental to the idea of a probability distribution, we first discuss it before proceeding to probability distributions.

10.2 RANDOM VARIABLES

In Chapter 3, it was mentioned that a variable is something that can take a number of different values, for example, the number of spots on a die on each of the past seven throws. A random variable is one,

that can take a number of different values, but it is not possible to know which value it does take until some experiment is performed. Usually, the experiment takes the form of drawing a sample from a population. Flipping a coin, rolling a die, and taking a sample of 600 households are some of the examples. In other words, a random variable is a description of the numeric values that the outcomes from an experiment can take. Like the 'ordinary' variables, random variables can be classified into two categories—discrete and continuous.

A random variable is one that is said to be a discrete random variable if its possible values proceed in steps with either a finite or infinite number of steps. It can take a countable or finite number of possible values. In contrast, if a random variable represents a measurement on a continuous scale so that all values in an interval are possible, it is called a continuous random variable. Examples of a continuous random variable are price of a car and daily consumption of milk.

Probability Distribution of a Discrete Random Variable

The probability distribution of a discrete random variable, say x, is a list of the distinct numerical values of x along with their associated probabilities. Let us take an example.

Example 10.1 Let us take x as the number of heads obtained in three tosses of a fair coin. We are required to list the numerical values of x along with the corresponding outcomes.

Solution These values along with the corresponding outcomes are shown in Table 10.1.

Table 10.1 List of Outcomes

Outcome	*Value of x*
T T T	0
T T H	1
T H T	1
T H H	2
H T T	1
H T H	2
H H T	2
H H H	3

x is a variable since in three tosses of the coin, it can take any value—0,1, 2, or 3. Further, x is the random variable in the sense that we could not have predicted that value of the outcome before tossing the coin. It may be noted that for each elementary outcome, there is only one value of x. However, as we can see, two or more elementary outcomes may give the same value.

The Expected Value of a Random Variable

The mean of a discrete variable x is, in fact, the mean of its probability distribution. The mean of a discrete random variable is also called its expected value. It is denoted by E(x). When we perform an experiment a number of times, then what is our expectation from that experiment? The mean is the value that we expect to observe per repetition.

Example 10.2 Suppose we are given the following data relating to breakdown of a machine in a certain company during a given week, wherein x represents the number of breakdowns of a machine and P(x) represents the probability value of x.

Table 10.2 Number and Probability of Breakdowns

x	*0*	*1*	*2*	*3*	*4*
P(x)	0.12	0.20	0.25	0.30	0.13

Find out the mean number of breakdowns per week for this machine.

Solution In order to attempt this problem, we have to multiply each value of x by its probability and then add all these products. This has been shown in table given below:

Table 10.3 Calculation of Mean for the Probability Distribution of Breakdowns

x	$P(x)$	$x \cdot P(x)$
0	0.12	0
1	0.20	0.20
2	0.25	0.50
3	0.30	0.90
4	0.13	0.52
	$\Sigma \times P(x) \rightarrow$	2.12

Concept of Expected Value Thus, we find that the sum of these products gives $\Sigma[x \cdot P(x)] = 2.12$, which is the mean. This can be written as $\mu = \Sigma[x \cdot P(x)] = 2.12$. On the basis of this calculation, we can say that, on an average, this particular machine is expected to breakdown 2.12 times per week over a period of time. In other words, if this machine is used for several weeks, then there may not be any breakdown, for some other weeks there may be only one breakdown per week and so on. The mean number of breakdowns is expected to be 2.12 per week for the entire period. This is the concept of *expected value*.

Symbolically, E(x) = Σ[x.Prob.(x)], where E(x) = Expected value of a discrete variable x and x. Prob.(x) = Product of value of variable x with its probability.

It may be noted that expected value can be derived subjectively as well. On the basis of the experimenter's own experience and judgment, one may assign probability that the random variable will take on certain values.

Let us take another example.

Example 10.3 An accountant of a company is hoping to receive payment from two outstanding accounts during the current month. He estimates that there is 0.6 probability of receiving Rs 15,000 due from A and 0.75 probability of receiving Rs 40,000 due from B. What is the expected cash flow from these two accounts?

Solution

Table 10.4 Calculation of Expected Cash Flow

Account	*Amount (Rs)*	*Probability* (p_i)	*Amount* (x_i) *(Rs)*
A	15,000	0.60	9,000
B	40,000	0.75	30,000
		Total expected value $\rightarrow$	**39,000**

Importance of Expected Value The concept of expected value is of considerable importance to management in decision-making. This is because the criteria in decision problems involving uncer-

tainties are usually the maximisation of expected profits, or utility, and the minimisation of expected costs. In Chapter 22 on Decision Theory, we shall discuss these criteria in detail giving suitable examples.

With this introduction we now turn to the binomial distribution.

10.3 THE BINOMIAL DISTRIBUTION

The binomial distribution is also known as the *Bernoulli distribution* in honour of the Swiss mathematician Jacob Bernoulli (1654–1705) who derived it.

To begin with, we go back to our frequently used example of a fair coin in the last chapter. Assuming that the coin is tossed once, there can be two possibilities—either head (or success) or tail (or failure). The sum of the probabilities is $p + q$, where p is the probability of success and q of failure. Instead of success and failure we may also say 1 and 0.

Now, assume two coins are tossed together. Then, we can have four possibilities:

1. Both coins falling heads
2. The first coin falling head and the second falling tail
3. The first coin falling tail and the second falling head
4. Both coins falling tails

Thus, the probabilities of 2 heads (or 2 successes) = $p \times p = p^2$.

Probabilities of one head and one tail = $(p \times q) = pq$

Probabilities of one tail and one head = $(q \times p) = qp$

Probabilities of 2 tails (or 2 failures) = $q \times q = q^2$

Thus, the probabilities of 0, 1 and 2 successes are given by q^2, $2qp$, p^2, respectively, that is, by the successive terms of the expansion of the binomial $(q + p)^2$. In the same manner, if three coins are tossed simultaneously, probabilities of 0, 1, 2 and 3 successes will respectively be given by the terms q^3, $3q^2p$, $3qp^2$, p^3, being the successive terms of binomial $(q + p)^3$.

Let us put these results in the following form:

For one coin or event $(q + p)^1$ *that is,* $q + p$

For two coins or events $(q + p)^2$ *that is,* $q^2 + 2qp + p^2$

For three coins or events $(q + p)^3$ *that is,* $q^3 + 3q^2p + 3qp^2 + p^3$

Hence, for n coins or events $(q + p)^n$

$$(q + p)^n = q^n + nq^{n-1}p + \frac{n(n-1)}{2!} q^{n-2} p^2 + \ldots p^n$$

This is known as the binomial distribution.

10.4 CONDITIONS NECESSARY FOR BINOMIAL DISTRIBUTION

At this stage, we should know that there are certain conditions that must be fulfilled by a distribution if it is to be a binomial distribution. These conditions are:

1. It is necessary that each observation is classified in two categories such as success and failure. For example, if raw material is obtained by a firm from its suppliers, it may be classified as defective or non-defective on the basis of its normal quality. Similarly, if a die is thrown, we may call 4, 5 or 6 a success and getting 1, 2 or 3 a failure.

2. It is necessary that the probability of success (or failure) remains the same for each observation in each trial. Thus the probability of getting head (or tail) must remain the same in each toss of the experiment. In other words, if the probability of success (or failure) changes from trial to trial or if the results of each trial are classified in more than two categories, then it is not possible to use the binomial distribution.
3. The trials or individual observations must be independent of each other. In other words, no trial should influence the outcome of another trial.

Let us take an example. The binomial distribution $(q + p)^n$ in general terms ${}^nC_r\, q^{n-r}p^r$, where ${}^nC_r = n!/\{r!(n-r)!\}$, where r is the number of ways in which we can get r successes and $n - r$ failures out of n trials.

Example 10.4 Find the chance of getting 3 successes in 5 trials when the chance of getting a success in one trial is 2/3.

Solution Here, $n = 5$, $p = 2/3$, $q = 1 - p = 1 - 2/3 = 1/3$ and $r = 3$.

Substituting these values in general terms, the required chance is

$$
\begin{aligned}
& {}^nC_r\, q^{n-r}p^r \\
= \quad & {}^5C_3(1/3)^{5-3}(2/3)^3 \\
= \quad & \frac{5!}{3!(5-3)!} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \\
= \quad & \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \\
= \quad & 0.33 \text{ approx.}
\end{aligned}
$$

Using the Binomial Tables

From the preceding discussion on binomial distribution, we can make out that the calculation of probabilities using the binomial formula becomes tedious when n is a large number. In order to overcome this difficulty, we use binomial tables. Appendix Table 3 gives binomial probabilities. We will understand the use of Appendix Table 3 by taking an example.

Example 10.5 Seven coins are thrown simultaneously. Using binomial distribution, find out the probability of obtaining at least five heads.

Solution First of all, we represent the information given in this problem in binomial distribution notation:

$x = 7$
$p = 0.5$
$r = 5$

Now, as the problem involues 7 trials, we have to refer to the relevant part of the table corresponding to $n = 7$. Having located 7, we have to see the column $p = 0.5$,which happens to be the last column. We have now to see $r = 5$, in the table we have to look $k = 5$. Thus against $k = 5$ and $p = 0.5$, the corresponding value is 0.1641. Against $k = 6$ and $p = 0.5$, the corresponding value is 0.0547 and $k = 7$ and $p = 0.5$, it is 0.0078.We have to ascertain these three values, as we have to determine the probability of at least 5 heads. These three values are now added together:

5	0.1641
6	0.0547
7	0.0078
	0.2266

Thus, our answer is that the probability of getting at least 5 successes is 0.2266. Let us take another example.

Example 10.6 For a binomial distribution the mean is 4 and variance is 2. Find probability of getting **(i)** at least 2 successes, **(ii)** at the most 2 successes.

Solution

Given, the mean $np = 4$ and $\sigma^2 = 2$

$\therefore$ σ^2 is npq, which is 2

Hence, $q = npq/np = 2/4 = 0.5$

$\therefore$ $p = 1 - q = 1 - 0.5 = 0.5$

$\therefore$ $np = 4$ or $n \times 0.5 = 4$

$\therefore$ $n = 8$

Having obtained $n = 8$, we can now use the Appendix Table 3 on binomial probabilities.

(i) Probability of getting at least 2 successes means we have to get probabilities of 2, 3, 4, 5, 6, 7 and 8 successes. Adopting the same procedure as described in the previous example, we find from the Appendix Table 3, the following probabilities:

2	0.1094
3	0.2188
4	0.2734
5	0.2188
6	0.1094
7	0.0312
8	0.0039
	0.9649

Thus, the probability of getting at least 2 successes is 0.9649.

(ii) Now, we take up the second part of the problem. Here, we are required to find at the most 2 successes. Using the same Appendix Table 3, we find the following probabilities:

0	success	0.0039
1	success	0.0312
2	success	0.1094
		0.1445

Thus, the probability of getting at the most 2 successes is 0.1445.

It may be noted that if we add up the answers in **(i)** and **(ii)**, the resultant figure is more than 1. This is because the probability of 2 successes comes in both the parts **(i)** and **(ii)**. If we omit the probability of 2 successes from one part, say **(ii),** then the total will be

0.9649	
0.0039	
0.0312	
1.0000	i.e. 1

Fitting a Binomial Distribution On the basis of some given information, if a binomial distribution is to be fitted, then the following procedure needs to be adopted.

1. Find the values of p and q. When one value is given to us, the other value can be easily obtained by subtracting the first value from 1.
2. Expand the binomial $(p+q)^n$. It may be noted that the power of n will be one less than the number of terms in the expanded binomial. For example, when $n = 5$, there will be 6 terms.
3. Multiply each of the expanded binomial terms by the total frequency (N) so that the expected frequency in each category can be obtained. Let us take an example.

Example 10.7 Fit a binomial distribution to the following data:

x	0	1	2	3	4
f	28	62	46	10	4

Solution

x	f	fx
0	28	0
1	62	62
2	46	92
3	10	30
4	4	16
	150	200

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{200}{150} = np$$

$$\therefore \quad p = \frac{200}{150 \times n} = \frac{200}{150 \times 4} = \frac{200}{600} = \frac{1}{3} \quad (\because n = 4)$$

The expected binomial frequencies can be obtained.

$$f(r) = N \cdot p(r) = N \times {}^nC_r p^r - q^{n-r}$$

$$= 150 \times {}^4C_r \left(\frac{1}{3}\right)^r \left(\frac{2}{3}\right)^{4-r}$$

Now, to get the binomial frequencies, we have to put $r = 0, 1, 2, 3$ and 4 in the above equation. These calculations are shown in the following table:

Table 10.5 Calculation of Binomial Frequencies

r	$f(r) = 150 \times {}^4C_r \left(\frac{1}{3}\right)^r \left(\frac{2}{3}\right)^{4-r}$
0	$f(0) = 150 \times {}^4C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^{4-0} = 150 \times \frac{16}{81} = 30$
1	$f(1) = 150 \times {}^4C_1 \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^{4-1} = 150 \times \frac{32}{81} = 59$
2	$f(2) = 150 \times {}^4C_2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^{4-2} = \frac{150 \times 24}{81} = 44$
3	$f(3) = 150 \times {}^4C_3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{4-3} = \frac{150 \times 8}{81} = 15$
4	$f(4) = 150 \times {}^4C_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^{4-4} = \frac{150 \times 1}{81} = 2$

The frequencies of the binomial distribution are shown in the extreme right of the above table.

10.5 MEAN AND STANDARD DEVIATION OF BINOMIAL DISTRIBUTION

The mean and standard deviation of such theoretical frequency distributions where we know the number of independent events and the probability of the happening of the event in question, can be very easily calculated. If M stands for the mean of such distribution, n for the number of independent events and p for the probability of the happening of the event in a single trial, then $M = np$. The value of the standard deviation of the expected frequencies in such cases is

$$\sigma = \sqrt{npq}$$

Example 10.8 Let us take an example to calculate mean and standard deviation of a binomial distribution. Assume the following binomial distribution:

x	0	1	2	3	4	5
f	4	20	40	40	20	4

Table 10.6 Worksheet

x	f	xf	d from 2.5	$d'(d/0.5)$	d'^2	fd'^2
0	4	0	–2.5	–5	25	100
1	20	20	–1.5	–3	9	180
2	40	80	–0.5	–1	1	40
3	40	120	0.5	1	1	40
4	20	80	1.5	3	9	180
5	4	20	2.5	5	25	100
Total	**128**	**320**			**Total**	**640**

Solution

$$\text{Mean} = \Sigma xf / \Sigma f = 320/128 = 2.5$$

$$d_i = x - \text{Mean}$$

$$\sigma = \left(\sqrt{\Sigma fd'^2 \div n}\right) \times (\text{C}) = \left(\sqrt{640 \div 128}\right) \times 0.5 = \sqrt{5} \times 0.5 = 1.12$$

Since there are 6 terms, $n = 6 - 1 = 5$

Mean $= np = 5 \times 0.5 = 2.5$ $\quad p = 2.5/5 = 0.5$ and $\quad q = 1 - 0.5 = 0.5$

The calculation of standard deviation by the following formula

$$\sigma = \sqrt{npq} = \sqrt{5 \times 0.5 \times 0.5} = \sqrt{5} \times 0.5 = 2.24 \times 0.5 = 1.12$$

The formula $\sigma = \sqrt{npq}$ gives the same result as obtained by the detailed calculations done earlier.

Example 10.9 A perfect die is thrown a large number of times in sets of 8. The occurrence of 5 or 6 is called a success. In what proportion of the sets would you expect three successes?

Solution Since a die has 1 to 6 numbers, the probability of getting a 5 is 1/6. Again, the probability of getting a 6 is 1/6.

Hence, the probability of getting a 5 or 6 (i.e. success) is $p = 1/6 + 1/6 = 1/3$. Therefore,

$$q = 1 - 1/3 = 2/3$$

Since the die is in sets of 8, the binomial distribution is

$$N\,(q + p)^n = N\,(2/3 + 1/3)^3$$

$\therefore$ The expected frequency of 3 successes is

$$N \times {}^8C_3\,(2/3)^5\,(1/3)^3$$

$$= \frac{N \times 8 \times 7 \times 6}{3 \times 2 \times 1} \times \frac{2^5}{3^8} = \frac{N \times 336 \times 32}{6 \times 6561}$$

$$= \frac{N \times 1792}{6561}$$

Hence, the proportion of the sets in which three successes are expected

$$= \frac{N \times 1792}{6561} \times \frac{1}{N} = 0.27$$

Example 10.10 If on average 8 ships out of 10 arrive safely at ports, obtain mean and standard deviation of number of ships returning safely out of 150 ships.

Solution Probability of safe returning (p) is 8/10 = 0.8 and not safe returning (q) = 2/10 = 0.2.

The probability of 0, 1, 2, . . . 150 ships safely returning out of a total of 150 will be given by the various terms of the expansion $(0.8 + 0.2)^{150}$.

The mean of the distribution found by putting probabilities against the number of ships returned safely will be np, that is, $150 \times 0.8 = 120$. The standard deviation will be

$$\sqrt{npq} = \sqrt{150 \times 0.8 \times 0.2} = \sqrt{24} = 4.9$$

Mean = up = 120 Standard deviation = 4.9

Example 10.11 A marksman can hit a target 2 out of 3 times. In 4 shots, what are his chances of hitting it 0, 1, 2, 3 or 4 times?

Solution The probability that he will miss the target is 1 – 2/3, that is, 1/3.

Hence, the required chances of hitting the target are given by the expression $(q + p)^n$

$$= (1/3 + 2/3)^4$$

$$= (1/3)^4 + 4(1/3)^3\,2/3 + 6(1/3)^2\,(2/3)^2 + 4(1/3)\,(2/3)^3 + (2/3)^4$$

$$= (1/3 \times 1/3 \times 1/3 \times 1/3) + (4 \times 1/3 \times 1/3 \times 1/3 \times 2/3) + (6 \times 1/3 \times 1/3 \times 2/3 \times 2/3) + (4 \times 1/3 \times 2/3 \times 2/3 \times 2/3) + (2/3 \times 2/3 \times 2/3 \times 2/3)$$

$$= 1/81 + 8/81 + 24/81 + 32/81 + 16/81 = 1$$

Hence, his chances of hitting the target are as given on next page:

Table 10.7 Probability of Hitting the Target

Times	*Chance or Probability*
0	1/81
1	8/81
2	24/81
3	32/81
4	16/81

Meeting the Conditions for using the Bernoulli Process

Before closing our discussion on the binomial distribution, it must be emphasised that one should be careful in using the binomial probability. It is necessary to ensure that conditions specified earlier for binomial distribution are satisfied, particularly conditions 2 and 3. Condition 2 requires that the probability of the outcome of any trial should remain unchanged for each trial. While this condition is fully met in experiments involving tossing a coin or rolling a die, in real life it may be difficult to ensure the compliance of this condition.

Condition 3 requires that the trials of a Bernoulli process must be independent of each other. This means that the outcome of one trial must not influence in any way the outcome of any other trial. This condition, too, may not be satisfied in real-life situation. For example, take the case of interviewing candidates for a certain post in a company. The expert, who is interviewing the candidates, may find that the first three candidates are far below the standard expected. In view of this, he may not remain impartial (as he was earlier) while interviewing the fourth candidate. This means violation of condition 3. One can find several situations of this type in everyday life where compliance of condition 3 becomes extremely difficult.

10.6 THE POISSON DISTRIBUTION

Having discussed the binomial distribution in the preceding section, we now turn to Poisson distribution, which is also a discrete probability distribution. It was developed by a French mathematician SD Poisson (1781–1840) and hence named after him.

Along with the normal and binomial distributions, the Poisson distribution is one of the most widely used distributions. It is used in quality control statistics (as we shall see in Chapter 23 on Quality Control) to count the number of defective items or in insurance problems to count the number of casualties or in waiting-time problems to count the number of incoming telephone calls or incoming customers or the number of patients arriving to consult a doctor in a given time period, and so forth. All these examples have a common feature: they can be described by a discrete random variable, which takes on integer values (0, 1, 2, 3, and so on).

The characteristics of the Poisson distribution are:

1. *The events occur independently. This means that the occurrence of a subsequent event is not at all influenced by the occurrence of an earlier event.*
2. *Theoretically, there is no upper limit with the number of occurrences of an event during a specified time period.*

3. *The probability of a single occurrence of an event within a specified time period is proportional to the length of the time period or interval.*
4. *In an extremely small portion of the time period, the probability of two or more occurrences of an event is negligible.*

10.7 CALCULATING POISSON PROBABILITIES

Let us take an example to show how Poisson probabilities can be calculated.

Example 10.12 Suppose, we have a production process of some item that is manufactured in large quantities. We find that, in general, the proportion of defective items is $p = 0.01$. A random sample of 100 items is selected. What is the probability that there are 2 defective items in this sample?

Solution

The Poisson formula is

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

where $P(x)$ = probability of x occurrences

λ^x = Lambda (i.e. the mean number of occurrences per interval of time) raised to the x power

e^{λ} = 2.71828 (being the base of the natural logarithm system), raised to the negative lambda power

$x!$ = x factorial

Here $\lambda^x = np = 100 \times 0.01 = 1.0$

Applying the above formula to the data given

$$P(2) = \frac{(1)^2 \times (2.71828)^{-1}}{2 \times 1}$$

(The value of $e^{-\lambda}$ can be obtained from the Appendix Table 4(a) given at the end of the book.)

$$P(2) = \frac{(1)^2 \times 0.36788}{2} = 0.18394$$

Suppose, we want to know what is the probability of having up to 2 defective items in that sample of 100 items. We simply add the 3 figures:

P(0)	0.368
P(1)	0.368
P(2)	0.184
Total	**0.92**

The answer is 0.92.

Again, if we are interested in knowing the probability of having more than 2 defective items, then the answer will be

$$1 - 0.92 = 0.08$$

Example 10.13 Suppose the probability of dialing a wrong number is 0.05. Then, what is the probability of dialing exactly 3 wrong numbers in 100 dials?

Solution

$p = 0.05$

$n = 100$

$\lambda = np$

$= 100 \times 0.05 = 5$

Applying the Poisson formula,

$$P(x) = \frac{(5)^3 \times (2.71828)^{-5}}{3!}$$

$$= \frac{125 \times 0.0067^*}{6} = 0.14$$

Example 10.14 Fit a Poisson distribution to the following data, which relate to the number of deaths due to the kick of a horse in 10 corps per army per annum over 20 years

Deaths	0	1	2	3	4	**Total (f)**
Frequency	109	65	22	3	1	200

Solution Calculate the theoretical frequencies.

The theoretical expected frequencies are given by the formula

$$N \times \frac{\lambda^x \times e^{-\lambda}}{x!}$$

where $x = 0, 1, 2, 3$ and 4

N = total frequency

λ = mean

$e = 2.71828$

In order to find the value of λ, we have to calculate the arithmetic mean.

Table 10.8 Worksheet for Data in Example 10.14

Deaths (x)	*Frequency (f)*	*fx*
0	109	0
1	65	65
2	22	44
3	3	9
4	1	4
Total	**200**	**122**

Mean = $\Sigma fx/n = 122/200 = 0.61$

$$N \times \frac{\lambda^x \times e^{-\lambda}}{x!}$$

* From the Appendix Table 4(a).

$$= \frac{200 \times (0.61)^x \times (2.71828)^{-0.61}}{x!}$$

$$e^{-0.61} = 0.5435$$

Now for each value of x from 0 to 4, we have to calculate the frequency. This is shown below:

x	f
0	$200 \times 0.5435 = 108.7$
1	$200 \times 0.61 \times 0.5435 = 66.3$
2	$\frac{200 \times (0.61)^2 \times 0.5435}{2} = 20.2$
3	$\frac{200 \times (0.61)^3 \times 0.5435}{3 \times 2} = 4.1$
4	$\frac{200 \times (0.61)^4 \times 0.5435}{4 \times 3 \times 2} = 0.6$

Thus, the theoretical frequencies are:

Table 10.9 Theoretical Frequencies of Data in Example 10.14

x	Tf	f
0	109	109
1	66	65
2	20	22
3	4	3
4	1	1
Total	**200**	**200**

Tf = theoretical frequency

f = frequencies given earlier

Example 10.15 Let us take another example. The following table shows the number of days in a 50–day period during which x number of automobile accidents occurred in a city. Fit a Poisson distribution to the data.

No. of accidents (x)	0	1	2	3	4
No. of days (f)	21	18	7	3	1

Solution

We have to first calculate the arithmetic mean for which the following worksheet is set up.

Table 10.10 Worksheet for Data in Example 10.15

X	f	fx
0	21	0
1	18	18
2	7	14
3	3	9
4	1	4
Total	**50**	**45**

Mean = $\Sigma fx/\Sigma f = 45/50 = 0.9$

According to the Poisson distribution

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

Since the total number of frequency is 50, we may write the above formula

$$P(x) = \frac{N \cdot \lambda^x \times e^{-\lambda}}{x!} = \frac{50 \times (0.9)^x \times (2.71828)^{-0.9}}{x!}$$

From the Appendix Table 4(a), we find

$$e^{-0.9} = 0.40657$$

Now, we may calculate the Poisson frequencies for each value of x.

Table 10.11 Calculation of Poisson Frequencies

x	f	f *(in integers)*
0	$50 \times 0.40657 = 20.3$	20
1	$50 \times 0.9 \times 0.40657 = 18.3$	18
2	$\frac{50 \times (0.9)^2 \times 0.40657}{2} = 8.2$	8
3	$\frac{50 \times (0.9)^3 \times 0.40657}{3 \times 2} = 2.5$	3
4	$\frac{50 \times (0.9)^4 \times 0.40657}{4 \times 3 \times 2} = 0.6$	1

10.8 USE OF POISSON PROBABILITIES' TABLES

At this stage, we may point out that manual calculations of Poisson probability generally become tedious and time-consuming. We may avoid manual calculations by using the Poisson probabilities as given in the Appendix Table 4(b). In order to use this table, we need only 2 values, i.e., the values for x and λ.

For the preceding example of automobile accidents, we find the table value of Poisson probability for $x = 0$ and $\lambda = 0.9$ as 0.4066. Since total number of frequencies, that is $N = 50$, we multiply this figure by 50: $0.4066 \times 50 = 20.33$ approximated to 20.3. This is the same as obtained earlier, manually. In the same way, we can proceed further and determine the Poisson probabilities for other values of x for this problem.

It should be remembered that one has to use the Appendix Table 4(b) properly. The values of x are given vertically while those of λ are given horizontally. For most of the problems on Poisson distribution, the table can be used with advantage. Appendix Table 4(a) giving the values of $e^{-\lambda}$ for different negative values of λ is also very useful, as we have seen earlier in our examples. Thus, these two tables together greatly facilitate us in the calculation work.

Let us take an example to show how Appendix Tables 4(a) and 4(b) can be used to find the expected frequencies of a Poisson distribution.

Example 10.16 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

Solution First, we have to calculate the mean. This is calculated below.

Mistakes Per Page x	*No. of Pages* f	fx
0	142	0
1	156	156
2	69	138
3	27	81
4	5	20
5	1	5
	400	**400**

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{400}{400} = 1$$

This means $\lambda = 1$

Now, if we refer to Table 4(b) in an appropriate column where λ is shown as 1.0. Below this column are listed values of $K = 0$ to $K = 7$. We are concerned with $K = 0$ to $K = 5$. Against $K = 0$, we find the Poisson probability as 0.3679. Now to find the expected frequency, we have to multiply this figure by the total frequencies, i.e., 400. Thus, $0.3679 \times 400 = 147.16$ or 147. All these calculations are shown in the following table.

Table 10.12

Variable	*Poisson Probability*	*Total No. of Frequencies*	*Expected Frequencies*
(1)	*(2)*	*(3)*	*(4) = (2) × (3)*
0	0.3679	400	147
1	0.3679	400	147
2	0.1839	400	74
3	0.0613	400	25
4	0.0153	400	6
5	0.0031	400	1
			400

If we do not use the table and calculate expected frequencies by using the formula:

$$f(r) = \mathrm{N}p(r) = 400 \times \frac{e^{-1}}{r!}$$

putting $r = 0$, 1, 2, 3, 4 and 5, we will get the same expected frequencies as shown in the above table. It thus becomes obvious that the use of Appendix Table 4(b) saves us from tedious calculations.

10.9 POISSON DISTRIBUTION AS AN APPROXIMATION OF BINOMIAL DISTRIBUTION

As the binomial distribution involves tedious calculations, the Poisson distribution can be used in its place. However, this holds good under certain conditions such as when the number of trials is large while the binomial probability of success is small. When n is ≥ 20 and $p < 0.05$, then the Poisson distribution approaches as a limiting form of the binomial distribution. The calculation of binomial probabilities becomes much simpler by using Poisson distribution.

The formula for calculating Poisson probabilities is

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

If we change λ by np in the above formula, then the Poisson distribution becomes an approximation of the Binomial:

$$P(x) = \frac{(np)^x \times e^{-np}}{x!}$$

Let us take an example.

Example 10.17 An urn contains 99 white balls and one black ball. A random sample of size $n = 100$ is taken with replacement from this urn. Calculate the probabilities of getting 0 to 10 black balls and show these probabilities by means of a diagram. Also obtain Poisson probabilities and comment on the results.

Solution This may be considered as a sequence of 100 Bernoulli trials where the probability of drawing a white ball is 0.99 and that of drawing a black ball is 0.01. From the binomial distribution, we may state that the probability of having k black balls in the sample is

$$b(k;\, n = 100,\, p = 0.01) = \binom{100}{k} p^k (1-p)^{n-k}$$

With the help of binomial probabilities tables, we can get the probability of k balls. Taking the value of k from 0 to 10, we get the corresponding probabilities of getting black balls. These are shown below:

Table 10.13 Calculation of Binomial Probabilities

k	$P(k)$
0	0.366
1	0.370
2	0.185
3	0.061
4	0.015
5	0.003
6	0.000
7	0.000
8	0.000
9	0.000
10	0.000

If we chart out these probabilities on a graph paper, then Fig. 10.1 will emerge:

It can be seen that this graph of the probabilities is highly skewed to the right. We can see that the binomial distribution differs from the normal distribution. In cases where p is very small as is the case here, it is not appropriate to use the normal distribution as an approximation of the binomial distribution. In such cases, the Poisson distribution can be taken as an approximation of the binomial distribution.

As we know, the Poisson distribution of k balls is

$$p(k;\, \lambda) = \frac{\lambda^k\, e^{-\lambda}}{k!}$$

where λ indicates the average number of occurrences of the event. In our case, $\lambda = np = 100 \times 0.01 = 1$, which implies that the average (or expected) number of black balls per 100 draws is 1. With the help of Poisson probabilities tables, we can get the probabilities of k balls from 0 to 10. These probabilities are shown in Table 10.14.

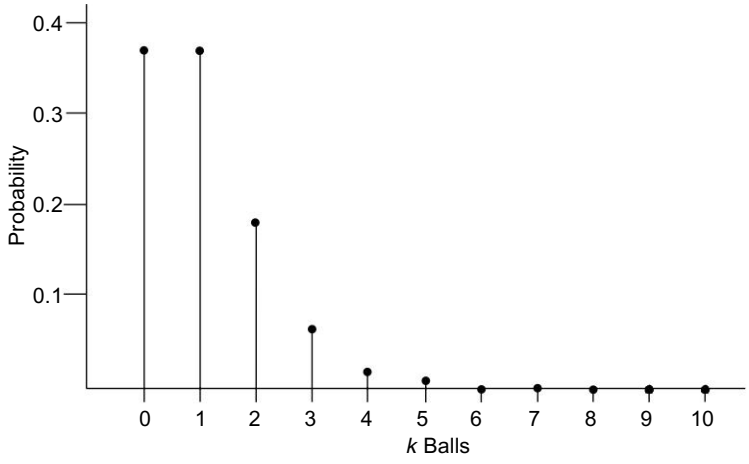


Fig. 10.1 **Probabilities of *k* Balls**

Table 10.14 **Comparison of Poisson and Binomial Probabilities**

k balls	*p(k; λ = 1)*	*Binomial Probabilities P(k; n = 100; p = 0.01)*
0	0.367	0.366
1	0.367	0.370
2	0.183	0.185
3	0.061	0.061
4	0.015	0.015
5	0.003	0.003
6	0.000	0.000
7	0.000	0.000
8	0.000	0.000
9	0.000	0.000
10	0.000	0.000

For the sake of comparison, the binomial probabilities given in Table 10.13 are also given side by side with the Poisson probabilities in Table 10.14. This comparison shows distinctly a very close resemblance between the two probability distributions.

We may reiterate that when n becomes larger and p becomes proportionately smaller where np remains unchanged, then the binomial distribution approaches the Poisson distribution. It may also be mentioned here that the calculation of probabilities in the case of the Poisson distribution is easier than in the case of binomial distribution. As such, in certain situations, the Poisson distribution can be substituted for the binomial distribution. However, in doing so, there is a loss of some accuracy. While attempting the Poisson distribution as an approximation of the binomial distribution, as mentioned earlier, it is necessary to ensure that $n \geq 20$ and $p \leq 0.5$. If we adhere to this assumption, our results will be reasonably good.

10.10 THE NORMAL DISTRIBUTION

The preceding two distributions discussed in this chapter were discrete probability distributions. We shall now take up another distribution in which the random variable can take on any value within a given range. This is the normal distribution, which is an important continuous probability distribution. This distribution is also known as the Gaussian distribution after the name of the eighteenth century mathematician-astronomer Karl Gauss, whose contribution in the development of the normal distribution was very considerable. As a vast number of phenomena have approximately normal distribution, it has wide application in Statistics. In business, there arise a number of situations where management has to make inferences by drawing samples. The normal distribution has certain characteristics, which make it applicable to such situations.

10.11 CHARACTERISTICS OF NORMAL PROBABILITY DISTRIBUTION

Figure 10.2 shows the normal probability distribution:

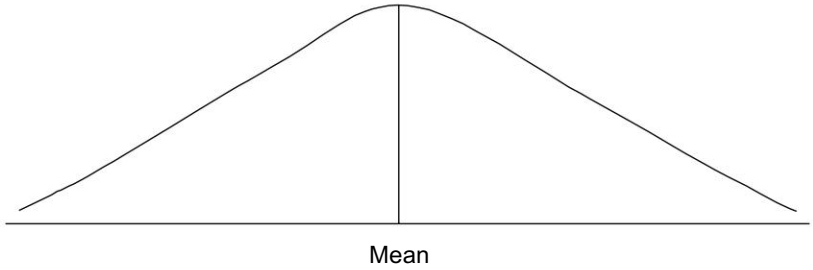


Fig. 10.2 The Normal Probability Distribution

Let us see what does this figure indicate in terms of characteristics of the normal distribution. It indicates the following *characteristics*.

1. The curve is bell-shaped, that is, it has the same shape on either side of the vertical line from mean.
2. It has a single peak. As such it is unimodal.
3. The mean is located at the centre of the distribution.
4. The distribution is symmetrical.
5. The two tails of the distribution extend indefinitely but never touch the horizontal axis.
6. Since the normal curve is symmetrical, the lower and upper quartiles are equidistant from the median, that is, Q_3 – Median = Median – Q_1.
7. The mean, median and mode have the same value, that is, mean = median = mode.
8. The percentage distribution of area under standard normal curve is broadly as follows: $\pm 1\sigma$ 68.27%; $\pm 2\sigma$ 95.44% and $\pm 3\sigma$ 99.73%. This was also shown in Fig. 7.1.

The units for the standard normal distribution curve are denoted by Z and are called *the Z values* or *Z scores*. They are also called *standard units* or *standard scores*. The Z score is known as a 'standardised' variable because it has a zero mean and a standard deviation of one.

As can be seen from Fig. 10.3, the horizontal axis is labelled *Z*. The *Z* values on the right side of the mean are positive while those on its left side are negative. The *Z* for a point on the horizontal axis gives the distance between the mean and that point in terms of the standard deviation. For example, a specific value of *Z* gives the distance between the mean and the point represented by *Z* in terms of 1 standard deviation to the right of the mean. Likewise, a point with a value of $Z = -1$ is one standard deviation to the left of the mean. It can be seen that the mean is at the centre and its value has been shown as zero. The area on either side of the mean is 0.5. Thus, the total area under the curve is 1.

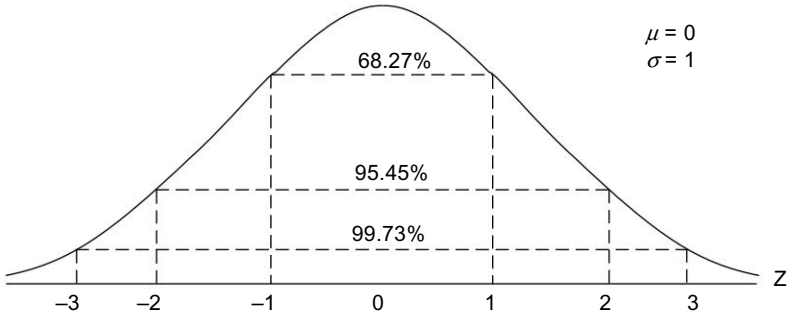


Fig. 10.3 **The Standard Normal Distribution Curve**

It may be noted that there is not a single normal curve but a whole series of normal curves. See, for example, Fig. 10.4,* which shows three different normal curves—A, B and C. Although the curves are different but the average is the same. What is different is the spread of the distributions. Curve A shows that a large number of values are closer to the mean. Curve B shows that its values are more spread out than those of curve A. However, as compared to curve C its values are less spread out. It is clear that amongst these three curves, curve C has the widest spread out.

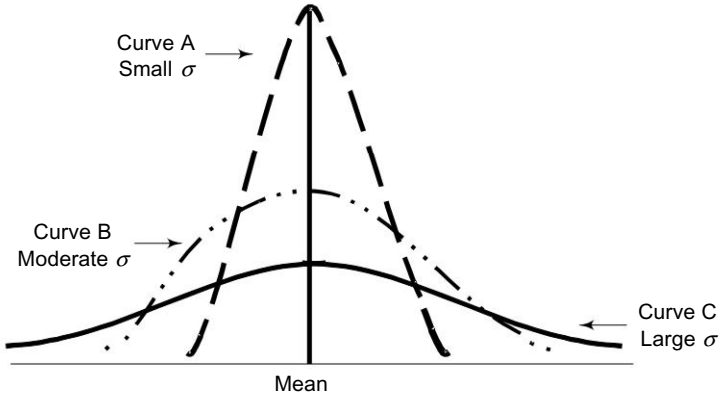


Fig. 10.4 **Three Different Normal Curves**

*This is the same as Fig. 8.1.

Figure 10.4 shows clearly that the mean alone cannot give any idea of the actual shape of the curve. Thus, any normal distribution is defined by two measures, the mean and the standard deviation. The mean shows the location of the centre while the standard deviation measures the spread around the centre. As we have discussed the steps involved in the calculation of standard deviation in Chapter 7 on Dispersion earlier, we shall not repeat them. We shall discuss here the standard normal probability distribution. Let us take two normal curves with different means.

Figure 10.5 shows two normal probability distributions each having a different mean and a different standard deviation. It should be noted that in both the distributions, the shaded areas, that is, area A and area B contain the identical proportion of the total area under the normal curve. This is because in each case the shaded area lies between the mean and 1 standard deviation to the right of the mean. Suppose, we draw another figure showing a normal probability distribution having yet another mean and standard deviation. Here too, the area between the mean and 1 standard deviation to the right of mean will have the same proportion as in the other two distributions. As a result of this phenomenon, it is possible to use one standard normal probability table. (Appendix Table 1).

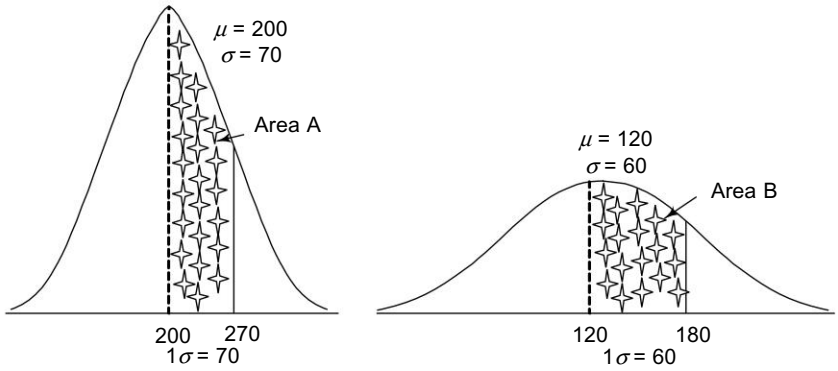


Fig. 10.5 **Two Normal Probability Distributions with Different Means**

10.12 USING THE STANDARD NORMAL PROBABILITY TABLE

Appendix Table 1 shows the area under the normal curve. In this section, we show how to use the normal curve table. There are two kinds of normal area tables: one gives the proportion of area of the tail-end as shown by the area marked 'I' in Fig. 10.6 and the second normal area table gives the area that is marked by 'II' in Fig. 10.6. If the proportion of I is 0.15 then the proportion of II, as given by the second table, is 0.5 – 0.15 = 0.35. This is because the mean μ divides the normal curve into two equal parts; the total area of the curve is 1.

We shall use the first type of normal area table in this chapter. The reason is that in our subsequent discussion of statistical inference, the area in the tail-end of the normal curve will be more relevant to us.

As the normal area table gives the value of z, we should know what it stands for. The value of z is obtained from the formula

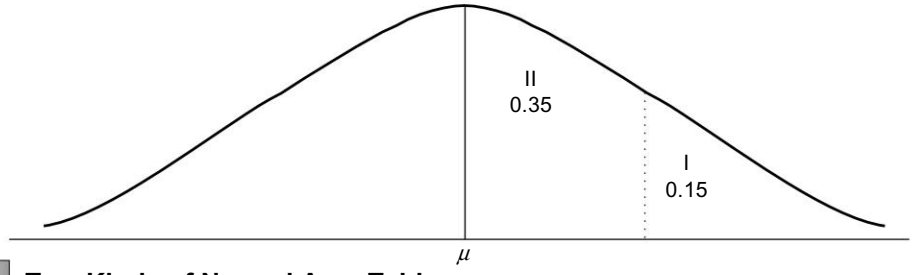


Fig. 10.6 **Two Kinds of Normal Area Table**

$$z = \frac{x - \mu}{\sigma}$$

where the symbols z, x, μ and σ stand for the same meaning as given earlier. Here, one may ask: what is the need for using z. This is because our normal distributions will have different units of measurement such as rupees, time, tonnes, and so forth. Depending on the problems on hand, we should use some standard unit, which can be applied regardless of the varying units of measurement. All this boils down to standard deviations where standard units are given a symbol of z. Let us take a couple of examples where we use the normal area table given as Appendix Table 1.

Example 10.18 Assume a variable X is distributed normally with $\mu = 5$ and $\sigma = 2$. What is the probability of obtaining a value of x as large or larger than 8.

Solution

Now, $z = \dfrac{x - \mu}{\sigma} = \dfrac{8 - 5}{2} = 1.5$

Referring to the Appendix Table 1, we find the probability figure 0.0668 against the value of $z = 1.5$. This means that 6.68% of the area in the distribution falls to the right of $z = 1.5$. In other words, the probability of obtaining a value of X as large or larger than 8 is 0.0668.

Example 10.19 Given that a random variable X is distributed normally with a mean of 70 and a variance of 36. Find the probability that X assumes a value

(a) 82 and greater;
(b) 75 or greater, and
(c) Between 75 and 85.

Solution Certain values of z have been extracted from the Appendix Table 1. These are given in Table 10.15:

Table 10.15 **Certain Values Extracted from the Normal Area Table**

z	*0.00*	. . .	*0.03*
0.0	.5000	. . .	.4880
0.1	.4602	. . .	.4483
0.8	.2119	. . .	.2033
1.5	.0668	. . .	.0630
2	.0228	. . .	.0212
2.5	.0062	. . .	.0057

(a) $z = \dfrac{X - \mu}{\sigma} = \dfrac{82 - 70}{\sqrt{36}} = 2$

It will be seen that against $z = 2$, we find from the table the figure of .0228. This means that there are just over 2 chances in 100 that X assumes a value of 82 and greater than 82. The shaded part in Fig. 10.7(a) shows this area.

(b) $z = \dfrac{X - \mu}{\sigma} = \dfrac{75 - 70}{\sqrt{36}} = 0.83$

Probability against $z = 0.83$ is 0.2033. This means that the probability of X assuming a value of 75 and more is just 0.2. Figure 10.7(b) shows it as the shaded area.

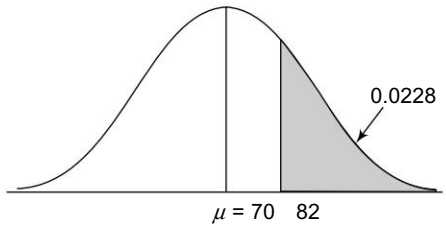


Fig. 10.7(a) Area Under the Normal Curve when $X \geq 82$

(c) $z = \dfrac{X - \mu}{\sigma} = \dfrac{85 - 70}{\sqrt{36}} = 2.5$

Probability is .0062

Now the area between $X = 75$ and $X = 85$ is

For $X = 75$ and above .2033

For $X = 85$ and above $\dfrac{.0062}{.1971}$

Our calculation shows that 19.71% of the area of the normal curve lies between 75 and 85 values of X. Figure 10.7(c) shows this as the shaded area.

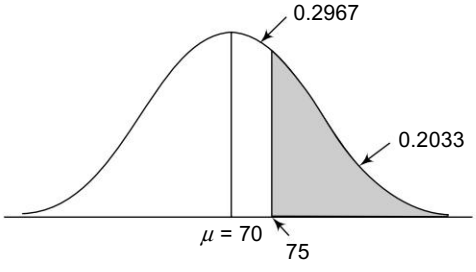


Fig. 10.7(b) Area Under the Normal Curve when $X \geq 75$

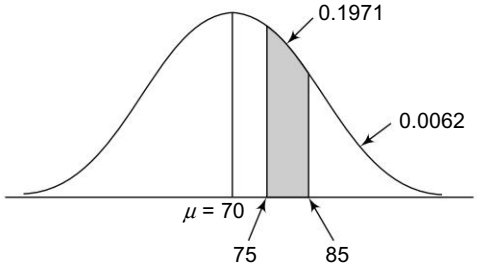


Fig. 10.7(c) Area Under the Normal Curve when X is between 75 and 85

Example 10.20 Suppose the owner of a bakery knows that the daily demand for his wholemeal bread is a random variable having the mean $\mu = 400$ loaves and the standard deviation = 20. What is the probability that the demand for its bread will exceed 450 loaves?

Solution It is better to give a normal curve diagram to understand the implications of the problem. This is shown in Fig. 10.8.

$$z = \frac{X - \mu}{\sigma} = \frac{450 - 400}{20} = 2.5$$

As can be seen from the normal area table given earlier, the probability against $z = 2.5$ is 0.0062. The shaded area at the right tail-end in Fig. 10.8 shows this probability. Thus we can say that the probability of demand exceeding 450 loaves is extremely low being 0.0062 or merely 0.62%.

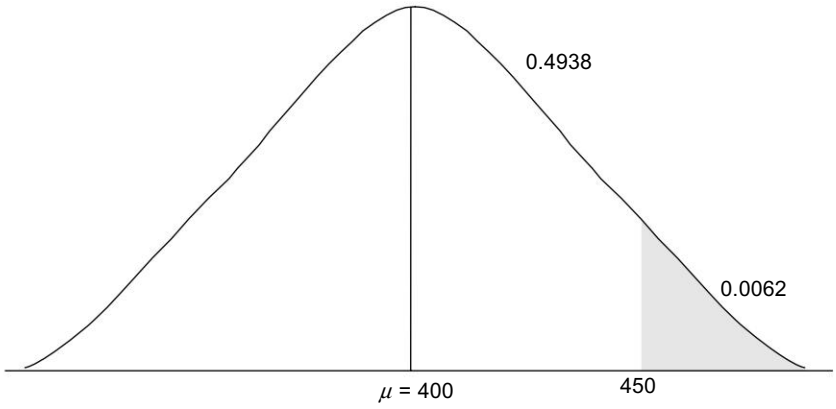


Fig. 10.8 **Probability of Demand for Bread**

Example 10.21 The average monthly sales of 5000 firms are normally distributed. Its mean and standard deviation are Rs 36,000 and Rs 10,000, respectively. Find

(i) the number of firms having sales over Rs 40,000;

(ii) the percentage of firms having sales between Rs 38,500 and Rs 41,000;

(iii) the number of firms having sales between Rs 30,000 and Rs 40,000.

The relevant extract of the Area Table[1] (under the Normal Curve) is given below.

Z	0.25	0.40	0.5	0.6
Area	0.0987	0.1554	0.1915	0.2257

Solution

Given $\mu = 36{,}000$ $\sigma = 10{,}000$ and N = 5,000

Let X denote the monthly sales of the firms.

(i) $z = \dfrac{X - \mu}{\sigma} = \dfrac{40{,}000 - 36{,}000}{10{,}000} = 0.4$

$= 0.5 - 0.1554 = 0.3446$. Since there are 5,000 firms, we multiply this value by 5,000.

Therefore, the number of firms having sales over Rs 40,000 is $0.3446 \times 5{,}000 = 1723$.

(ii) $z = \dfrac{X - \mu}{\sigma} = \dfrac{38{,}500 - 36{,}000}{10{,}000} = 0.25$

$z = \dfrac{X - \mu}{\sigma} = \dfrac{41{,}000 - 36{,}000}{10{,}000} = 0.5$

$P(0 \le z \le 0.50) = 0.1915$

$P(0 \le z \le 0.25) = 0.0987$

[1]It may be noted that the values given here are from the second kind of normal-area table.

Therefore, P $(0 \le z \le 0.50)$ – P $(0 \le z \le 0.25)$ = 0.1915 – 0.0987
= 0.0928.

Hence, 9.28% of the firms have sales between Rs 38,500 and Rs 41,000.

(iii) $z = \dfrac{X - \mu}{\sigma} = \dfrac{30,000 - 36,000}{10,000} = -0.6$

$z = \dfrac{X - \mu}{\sigma} = \dfrac{40,000 - 36,000}{10,000} = 0.4$

= P $(-0.60 \le z \le 0.40)$
= P $(-0.60 \le z \le 0)$ + P$(0 \le z \le 0.40)$ = 0.2257 + 0.1554
= 0.3811

Hence, the number of firms having sales between Rs 30,000 and Rs 40,000 is $0.3811 \times 5,000 = 1906$ approx.

10.13 NORMAL APPROXIMATION TO BINOMIAL DISTRIBUTION

When we come across a similar situation repeatedly and are interested in success (one of the two outcomes), then the total number of successes may follow a binomial distribution. Of course, as was mentioned earlier, a binomial distribution must satisfy certain conditions such as the probability of success must remain the same during the experiment, and the outcome of an event must not influence the outcome of any successive event.

When the probability of success p is not very close to 0 or 1 and the number of trials is large, then the normal distribution can be a good approximation to the binomial distribution.

The normal approximation to the binomial distribution is particularly useful in problems where the formula for the binomial distribution is to be used repeatedly to obtain the values of several different terms. Let us explain this by taking an example.

Example 10.22 Suppose we are interested in knowing the probability of getting at least 15 replies to a questionnaire sent to 100 persons, assuming the probability that any one of these persons would reply is 0.2. In other words, we would like to know the probability of getting at least 15 successes (a success is taken as a reply to the questionnaire) in 100 trials when the probability of success on an individual trial is 0.2.

Solution If we have to solve this problem by applying the formula for the binomial distribution, we will have to calculate the sum of individual probabilities corresponding to 15, 16, 17, . . . , and 100 successes (or those corresponding to 0, 1, 2, . . . , and 14 successes). As we can see, this would obviously involve considerable calculations.

As an alternative to the above approach, suppose we use the normal approximation. In such a case, we have only to find shaded area of Fig. 10.9, namely, the area to the right of 14.5. It should be noted that instead of 15 we are using 14.5. This amounts to the use of *continuity correction* according to which 15 is represented by the interval from 14.5 to 15.5; 16 is represented by the interval from 15.5 to 16.5, and so on.

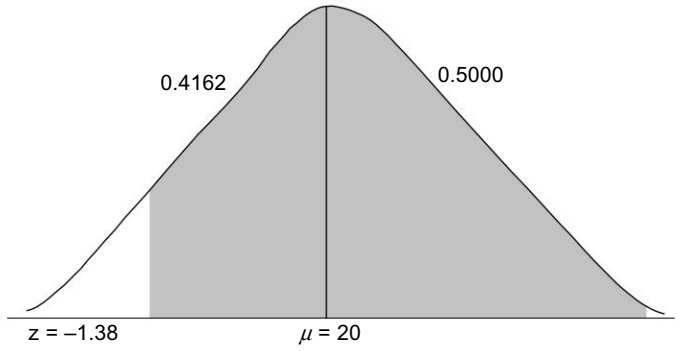


Fig. 10.9 **Distribution of Number of Replies**

Since $\mu = 100\ (0.2) = 20$ and $\sigma = \sqrt{100 \times (0.2)(0.8)} = 4$, we find that in standard units 14.5 becomes $z = \dfrac{14.5 - 20}{4} = -1.375$

or say –1.38 (rounding off to two decimal places).

Now, we have to find the probability of $z = -1.38$ from the Appendix Table 1 on standard normal distribution. Using that table, we find the value as .0838. Since this is the left tail end, we have to deduct this from 0.5 to get to the shaded area. This comes to 0.4162. As such the desired probability is 0.4162 + 0.5000 = 0.9162. This means that we can expect to get at least 15 replies from 100 persons about 92% of the time provided our assumption that the probability of 0.2 that any one person will reply is correct.

Limitations of the Normal Probability Distribution

We know that the tails of the normal distribution curve approach but never touch the horizontal axis. This means that there is some probability (however negligible it may be) that the random variable can take on extreme values. Further, we should note that the normal probability distribution is not the only continuous distribution. There are other continuous distributions that play an important role in statistical inference as has been discussed in subsequent chapters. It is true that the normal distribution has a major role in several statistical problems. All the same, as its indiscriminate use can lead to misleading results, one should ensure that it is used judiciously.

Additional Examples

Example 10.23 If $\bar{X}$ is the mean and s is the standard deviation of a set of measurements which are normally distributed, what percentage of the measurement is **(a)** within the range ($\bar{X} \pm 2s$), **(b)** outside the range ($\bar{X} \pm 1.2s$), **(c)** greater than ($\bar{X} - 1.5s$)?

Solution

(a) From standard normal table, we find the value of $z = 2$ as 0.4772. This is only one-half of the table. Since the question pertains to $\bar{X} \pm 2s$, this means we have to add two values, viz.

0.4772
0.4772
0.9544

This gives 95.4%.

(b) For $z = 1.2$, the corresponding value from Standard Normal Table is 0.3849. Hence, $Z = \pm 1.2s$ means 0.3849 + 0.3849 = 0.7698. Outside this range means we have to subtract this figure from 1, i.e., 1 – 0.7698 = 0.2302. This gives 23%.

(c) For $z = 1.5$, the corresponding value from Standard Normal Table is 0.4332, which means 0.5 – 0.4332 = 0.0668 is the area under the curve . To find the measurement greater than $\bar{X} - 1.5s$, we have to subtract this value from 1. Thus, 1 – 0.0668 = 0.9332. This gives 93.3%.

Example 10.24 Assuming that the height distribution of a group of men is normal, find the mean and standard deviation, given that 84 per cent of the men have heights less than 65.2 inches and 68 per cent have heights between 65.2 and 62.5 inches.

Solution

$z = \dfrac{\bar{X} - \mu}{\sigma}$, we have to find both μ and σ. For this there must be two simultaneous equations.

$$z = \frac{\bar{X} - \mu}{\sigma}$$

$$= \frac{65.2 - \mu}{\sigma} = 84$$

or $65.2 - \mu = 84\sigma$ (1)

$$z = \frac{\bar{X} - \mu}{\sigma}$$

$$= \frac{62.5 - \mu}{\sigma} = 16 \text{ (This is because of subtracting 68\% from 84\%)}$$

or $62.5 - \mu = 16\sigma$ (2)

Now subtracting equation (2) from (1)

$65.2 - \mu = 84\sigma$ (1)
$62.5 - \mu = 16\sigma$ (2)
$- \quad + \quad -$
$2.7 \quad = 68\sigma$

$\therefore \quad \sigma = \dfrac{2.7}{68} = 0.0397$

Substituting the value of $\sigma = 0.0397$ in (2) above, we get

$62.5 - \mu = (16 \times 0.0397)$

or $-\mu = 0.6352 - 62.5$

or $\mu = 61.86$

If we apply the value of μ and σ in the equations above, we can verify the accuracy of the results.

Example 10.25 The following table gives the distribution of height status among the management trainees in Delhi.

Height in Inches	*Number of Trainees*
61	2
62	10
63	11
64	38
65	57
66	93
67	106
68	126
69	109
70	87
71	75
72	23
73	9
74	4

Test the normality of 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 distribution and for the normal curve.

Solution

Table 10.16 Worksheet for Calculating $\bar{X}$ and σ

x	f	d from 68	fd	fd^2
61	2	–7	–14	98
62	10	–6	–60	360
63	11	–5	–55	275
64	38	–4	–152	608
65	57	–3	–171	513
66	93	–2	–186	372
67	106	–1	–106	106
68	126	0	0	0
69	109	1	109	109
70	87	2	174	348
71	75	3	225	675
72	23	4	92	368
73	9	5	45	225
74	4	6	24	144
	750		**–75**	**4201**

$$N = 750, \Sigma fd = -75, \Sigma fd^2 = 4201$$

$$\bar{X} = A + \frac{\Sigma fd}{N} = 68 + \left(\frac{-75}{750}\right) = 68 - 0.1 = 67.9$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{4201}{750} - \left(\frac{-75}{750}\right)^2}$$

$$= \sqrt{5.6 - 0.01} = \sqrt{5.59} = 2.4$$

$\bar{X} \pm 1\sigma = 67.9 \pm 2.4 = 65.5$ and 70.3

Number of cases lying within this range

$93 + 106 + 126 + 109 + 87 = 521$

Proportion $\frac{521}{750} = 0.69$ or 69%

$\bar{X} \pm 2\sigma = 67.9 \pm 4.8 = 63.1$ and 72.7

Number of cases lying within this range

$11 + 38 + 57 + 93 + 106 + 126 + 109 + 87 + 75 + 23 = 725$

Proportion $\frac{725}{750} = 0.96$ or 96%

$\bar{X} \pm 3\sigma = 67.9 \pm 7.2 = 60.7$ and 75.1

Number of cases lying within this range = 750, i.e. 100% . As in a normal distribution the proportion lying within these limits is about 68%, 95% and 99%, respectively, the given distribution is approximately normal.

Example 10.26 If z is normally distributed with mean 0 and variance 1, find:

(i) $P(z \geq -1.64)$

(ii) $P(-1.96 \leq z \leq 1.96)$

(iii) $P(z \geq 1)$

Solution

(i) $P(z \geq -1.64)$

Aganist $z = 1.64$ the standard normal table gives the corresponding value of 0.4495. Since we have to find $z > -1.64$ as well, we have to add 0.5 in this value. Hence $P(z \geq -1.64) = 0.4495 + 0.5 = 0.9495$.

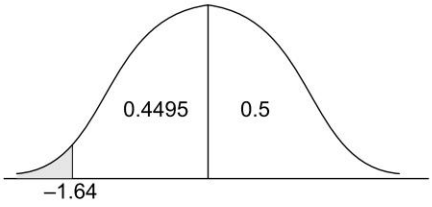


Fig. 10.10

(ii) $P(-1.96 \leqq z \leqq 1.96)$

Aganist $z = 1.96$, the corresponding value from the standard normal table is 0.4750. Since this is to be taken for both sides of the normal curve, the required probability is

$P(-1.96 \leqq z \leq 1.96) = 0.4750 + 0.4750 = 0.95$.

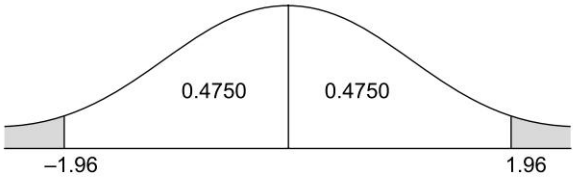


Fig. 10.11

(iii) P($|z| \geq 1$)

The term $|z|$ indicates z ignoring positive and negative signs, i.e., we have to consider both left and right sides. The corresponding value of $z = 1$ from standard normal table is 0.3413. Hence, P ($|z| \geq 1$) = 0.3413 + 0.3413 = 0.6826.

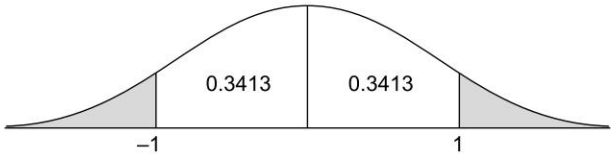


Fig. 10.12

Example 10.27 The average daily sales of 500 branch officials was Rs 1,50,000 and the standard deviation was Rs 15,000. Assuming the distribution to be normal, indicate how many branches have sales between:

(i) Rs 1,20,000 and Rs 1,45,000
(ii) Rs 1,40,000 and Rs 1,65,000
(iii) More than Rs 1,65,000

Solution

(i) Standard normal variate corresponding to 120 is (To simplify calculations, '000 omitted.)

$$z = \frac{X - \mu}{\sigma} = \frac{120 - 150}{15} = -2$$

and corresponding to 145 is

$$z = \frac{145 - 150}{15} = \frac{-5}{15} = -0.33$$

From the Appendix Table 1, we find the areas corresponding to the values of z are 0.4772 and 0.1293.

Hence, 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 1,20,000 and Rs. 1,45,000 is

$0.3479 \times 500 = 173.95$ or 174 approx.

(ii) Standard normal variate corresponding to 140 is

$$z = \frac{140 - 150}{15} = \frac{-10}{15} = -0.67$$

and corresponding to 165 is

$$z = \frac{165 - 150}{15} = \frac{15}{15} = 1$$

From the Appendix Table 1, the area corresponding to the z values are 0.2486 and 0.3413.

Hence, the desired area is 0.2486 + 0.3413 = 0.5899

Hence, the expected number of branches having sales between Rs 1,40,000 and Rs 1,65,000 is

$0.5899 \times 500 = 294.95$ or 295 approx.

(iii) More than Rs 1,65,000
$= 500 - (174 + 295)$
$= 500 - 469$
$= 31$ branches

Example 10.28 A wholesale distributor of a product finds that the annual demand for a product is normally distributed with the mean of 120 and the standard deviation of 16. 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. (Area between $z = 0$ and $z = 1.64$ is 0.45, where z is the standardised normal variate.)

Solution Let the random variable X denote the annual demand for a product. Then X is a normal variate with mean (μ) = 120 and standard deviation (σ) = 16. The standardised normal variate Z is

$$z = \frac{X - \mu}{\sigma} = \frac{X - 120}{16}$$

Let X_1 be the quantity ordered so that
$P(X > X_1) = 0.05$

When $X = X_1, z = \dfrac{X_1 - 120}{16}$, (say)

Then $P(z > z_1) = 0.05$
or $P(0 < z < z_1) = 0.45$

or $z_1 = \dfrac{X_1 - 120}{16} = 1.64$ (given)

Hence, $X_1 = 120 + (1.64)\,(16)$
$= 120 + 26.24 = 146.24$

Hence, if the wholesale distributor orders 146, there is only, a 5% chance of running short.

Example 10.29 In each case, find the probability of x successes in n Bernoulli trials with success probability p for each trial:

(a) $x = 2$ $n = 4$ $p = 1/3$
(b) $x = 2$ $n = 6$ $p = 3/4$
(c) $x = 3$ $n = 6$ $p = 1/4$

Solution

(a) $\dfrac{n!}{r!(n-r)!}\, p^r\, q^{n-r}$

$$= \frac{4!}{2!(4-2)!}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{4-2}$$

$$= \frac{\cancel{4} \times \cancel{3} \times 2 \times 1}{2 \times 1 \times \cancel{2} \times 1} \times \frac{1}{\cancel{3}} \times \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3}$$

$$= \frac{8}{27} = 0.296$$

(b) $\frac{n!}{r!(n-r)!}p^r q^{n-r}$

$$= \frac{6!}{2!(6-2)!}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^4$$

$$= \frac{{}^{3}\cancel{6}\times 5\times \cancel{4\times 3\times 2\times 1}}{\cancel{2}\times 1\times \cancel{4\times 3\times 2\times 1}}\times\frac{9}{16}\times\frac{1}{256}$$

$$= \frac{15\times 9}{16\times 256} = \frac{135}{4096} = 0.033$$

(c) $\frac{n!}{r!(n-r)!}p^r q^{n-r}$

$$= \frac{6!}{3!(6-3)!}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^3$$

$$= \frac{\cancel{6}\times 5\times 4\times \cancel{3\times 2\times 1}}{\cancel{3\times 2\times 1}\times\cancel{3\times 2\times 1}}\times\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{3}{4}$$

$$= {}^{5}\cancel{20}\times\frac{27}{\underset{1024}{\cancel{4096}}} = \frac{135}{1024} = 0.132$$

Example 10.30 If a set of measurements are normally distributed, what percentage of these differ from the mean by **(a)** more than half the standard deviation, **(b)** less than three quarters of the standard deviation?

Solution

(a) Percentage of a set of measurements which differ from the mean by more than half the standard deviation

Area between 0 and $z = 0.5$

From the standard normal table, the corresponding area against $z = 0.5$ is 0.1915

Hence the area greater than $z = 0.5$ is

$$0.5 - 0.1915 = 0.3085$$

This is for one side. As such the required figure will be

$$0.3085 + 0.3085 = 0.6170$$

Hence, 61.7%

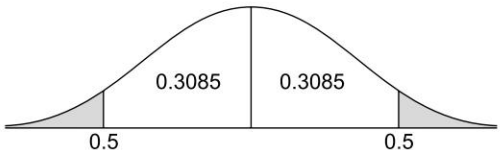


Fig. 10.13

(b) The corresponding area against $z = 0.75$ is 0.2734

This is for one side. For both sides of the normal curve it will be

$$0.2734 + 0.2734 = 0.5468$$

The required percentage is 54.7%.

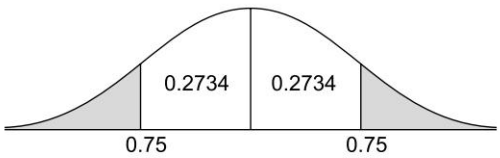


Fig. 10.14

Example 10.31 In a certain locality, half of the households is known to use a particular brand of soap. In a household survey, sample of 10 households are allotted to each investigator and 2048 investigators are appointed for the survey. How many investigators are likely to report:
(i) 3 users; **(ii)** not more than 3 users; and **(iii)** at least 4 users?

Solution $p = P$ i.e. Household using a particular brand of soap

$$p = \frac{1}{2} \text{ and } q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

On the basis of binomial probability law, the probability that there are r users in a sample of 10 households is

$$P(r) = {}^{10}C_r\, p^r q^{10-r}$$

$$= {}^{10}C_r \left(\frac{1}{2}\right)^{10} = \frac{1}{1024}\ {}^{10}C_r$$

(i) Now, probability that in a sample of 10 households, 3 to use a particular brand of soap is

$$P(3) = \frac{1}{1024} \times {}^{10}C_3 = \frac{120}{1024}$$

Hence, out of 2048 investigators, the number of investigators who will report that 3 are users of a particular brand of soap in a sample of 10 is

$$2048 \times \frac{120}{1024} = 240$$

(ii) The probability that in a sample of 10, not more than 3 household using a particular brand of soap is

$$P(0) + P(1) + P(2) + P(3)$$

$$= \frac{1}{1024}\ ({}^{10}C_0 + {}^{10}C_1 + {}^{10}C_2 + {}^{10}C_3)$$

$$= \frac{1}{1024}\ (1 + 10 + 45 + 120) = \frac{176}{1024}$$

Hence, out of 2048 investigators, the no. of investigators who will report 3 households using a particular brand of soap in a sample of 10 is

$$2048 \times \frac{176}{1024} = 352$$

(iii) Probability that in a sample of 10, at least 4 households are using a particular brand of soap is

$$P(4) + P(5) + \ldots + P(10) = 1 - [P(0) + P(1) + P(2) + P(3)]$$

$$= 1 - \frac{176}{1024} = \frac{848}{1024}$$

Hence, out of 2048 investigators, the no. of investigators, reporting at least 4 households using a particular brand in a sample of 10 is

$$2048 \times \frac{848}{1024} = 1696$$

Example 10.32 In a certain factory manufacturing razor blades, there is a small chance—1/50—for any blade to be defective. The blades are placed in packets, each containing 10 blades. Using the Poisson distribution, calculate the approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets.

Solution The formula for Poisson distribution is

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

$$\lambda = np = 10 \times \frac{1}{50} = 0.2$$

$$e^{-\lambda} = e^{-0.2} = 0.81873$$

$$\text{For 0 defect} = \frac{(0.2)^0\,(0.81873)}{0!}$$

$$= 0.81873$$

$$\text{For 1 defect} = \frac{(0.2)^1\,(0.81873)}{1!}$$

$$= 0.163746$$

$$\text{For 2 defects} = \frac{(0.2)^2\,(0.81873)}{2!}$$

$$= \frac{(0.04)\,(0.81873)}{2 \times 1}$$

$$= \frac{0.0327492}{2}$$

$$= 0.0163746$$

Hence, probability of packets having ≤ 2 defects is

$$\begin{array}{r} 0.81873 \\ 0.163746 \\ \underline{0.0163746} \\ 0.9988506 \end{array}$$

Out of 10000 packets: 0.9988506×10000

$$= 9988.506$$

or 9988 complete packets

Example 10.33 One hundred car stereos are inspected as they come off the production line and number of defects per set is recorded below:

No. of defects (X)	0	1	2	3	4
No. of sets	79	18	2	1	0

Fit a Poisson distribution to the above data.

Solution

The theoretical expected frequencies are given by the formula

$$N = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

where $x = 0, 1, 2, 3$ and 4

N = total frequency

λ = mean

$e = 2.71828$

In order to find the value of λ, we have to calculate the arithmetic mean

Table 10.17 Worksheet

No. of Defects (x)	*No. of Sets (f)*	*fx*
0	71	0
1	19	19
2	9	18
3	1	3
4	0	0
	100	**40**

$$\text{Mean} = \frac{\Sigma fx}{N} = \frac{40}{100} = 0.4$$

$$N \times \frac{\lambda^x \times e^{-\lambda}}{x!}$$

$$\frac{100 \times (0.4)^x \times (2.71828)^{-0.4}}{x!}$$

The value of $e = 2.71828^{-0.4} = 0.67032$

Now for each value of x from 0 to 4, we have to calculate the frequency. This is shown below.

Table 10.18 Calculation of Frequencies for Poisson Distribution

x	f	*Frequency*
0	$100 \times 0.67032 = 67.03$	67
1	$100 \times 0.4 \times 0.67032 = 26.81$	27
2	$\frac{100 \times (0.4)^2 \times 0.67032}{2 \times 1} = 5.36$	5
3	$\frac{100 \times (0.4)^3 \times 0.67032}{3 \times 2 \times 1} = 0.72$	1
4	$\frac{100 \times (0.4)^4 \times 0.67032}{4 \times 3 \times 2 \times 1} = 0.07$	0

Table 10.19 Theoretical Frequencies of Data

x	f	Tf
0	71	67
1	19	27
2	9	5
3	1	1
4	0	0
	100	**100**

Example 10.34 For a binomial distribution, mean is 6 and standard deviation is $\sqrt{2}$, find n, p and q.

Solution In a binomial distribution, mean is np and standard deviation is $\sqrt{npq}$.

$\because \quad \sigma = \sqrt{npq}$

Given $\quad \sigma = \sqrt{2}$

$\sqrt{2} = \sqrt{npq}$

$\therefore \quad npq = 2$

As mean (np) is 6

$\therefore \quad 6q = 2$

$\therefore \quad q = 2/6 \quad \text{or} \quad 0.33$

$\because \quad (p + q) = 1$

or $\quad p + 0.33 = 1$

$\therefore \quad p = 1 - 0.33$

$= 0.67$

Now, $\quad np = 6$

or $\quad (n)(0.67) = 6$

$\therefore \quad n = \dfrac{6}{0.67}$

$= 8.955$

Hence, $\quad n = 8.955 \quad \text{or} \quad 9 \text{ approx.}$

$p = 0.67$

$q = 0.33$

Example 10.35 A manufacturing company finds that 10 per cent of its tools turn out to be defective in the production process. Find the probability that in a sample of 20 tools chosen at random, exactly 4 will be defective.

Solution As 10 per cent tools turn out to be defective, probability is 10/100 = 0.1
We have to find the probability of defective tools in a sample of 20 tools.

$$\lambda = Np = 20 \times 0.1 = 2$$

P(4 defective tools in 20):

$$\frac{\lambda^X e^{-\lambda}}{X!} = \frac{2^4 e^{-2}}{4!}$$

$$= \frac{16 \times 0.13534^{*}}{24}$$

$$= \frac{2.16544}{24}$$

$$= 0.0902^{**}$$

---

*Taken from Appendix Table 4(a)

**Can be directly obtained from Appendix Table 4(b)

Example 10.36 Suppose, on an average, one house in 1000, in a certain town, has a fire during the year. If there are 2000 houses, what is the probability that (a) exactly 3 houses, (b) more than 2 houses will have fire during the year?

Solution

(a) As one house out of 1000 has fire, the probability is 1/1000 = 0.001

$$N = 2000$$
$$\lambda = Np = 2000 \times 0.001$$
$$= 2$$

Applying the formula,

$$\frac{\lambda^x e^{-\lambda}}{X!} = \frac{2^3 e^{-2}}{3!}$$

$$= \frac{8 \times 0.13534^*}{3 \times 2 \times 1}$$

$$= \frac{1.08272}{6}$$
$$= 0.1804^{**}$$

(b) P(more than 2 houses) means P(0 or 1 or 2 houses)

$$\text{P(0 house)} = \frac{2^0 e^{-2}}{0!} = \frac{1}{e^2}$$

$$\text{P(1 house)} = \frac{2^1 e^{-2}}{1!} = \frac{2}{e^2}, \text{P(2 houses)} = \frac{2^2 e^{-2}}{2!} = \frac{2}{e^2}$$

$$\text{P(more than 2 houses)} = 1 - \text{P(0 or 1 or 2 houses)}$$
$$= 1 - (1/e^2 + 2/e^2 + 2/e^2)$$
$$= 1 - 5/e^2$$
$$= 1 - 5/7.389$$
$$= 1 - 0.677$$
$$= 0.323^{***}$$

Example 10.37 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.

Solution

$$n = 200$$
$$p = 2\% \quad \text{i.e.} \quad 0.02$$
$$\lambda = np = 200 \times 0.02$$
$$= 4$$

---

*From Appendix Table 4(a).

**This can be directly obtained from Appendix Table 4(b).

***This can be directly obtained from Appendix Table 4(b) by adding all Poisson probabilities for $k = 3$ to 9.

$$p = \frac{\lambda^X e^{-\lambda}}{X!}$$

$$= \frac{4^5 e^{-4}}{5!}$$

$$= \frac{1024 \times 0.01832}{5 \times 4 \times 3 \times 2 \times 1}$$

$$= \frac{18.75968}{120}$$

$$= 0.1563^*$$

Example 10.38 The incidence of a certain disease is such that, on the average, 20% of the workers suffer from it. If 10 workers are selected at random, find the probability that (a) exactly 2 workers suffer from the disease, (b) not more than 2 workers suffer from the disease.

Solution

(a)

$$P = 20/100 = 0.2$$

$$Np = 10 \times 0.2 = 2$$

$$P = \frac{\lambda^X e^{-\lambda}}{X!}$$

$$= \frac{2^2 e^{-2}}{2!}$$

$$= \frac{4 \times (2.71828)^2}{2}$$

$$= \frac{4 \times 7.389046}{2}$$

$$= 0.27068 \quad \text{or} \quad 0.2707^*$$

(b)

$$\text{P(0 disease)} = \frac{1}{e^2}$$

$$\text{P(1 disease)} = \frac{2}{e^2}$$

$$\text{P(2 disease)} = \frac{2}{e^2}$$

$$\text{P(not more than 2 workers)} = \frac{1}{e^2} + \frac{2}{e^2} + \frac{2}{e^2}$$

$$= 0.1353 + 0.2707 + 0.2707$$

$$= 0.6767^*$$

---

*The answers can be checked from Appendix Table 4(b).

Example 10.39 A workshop produces 2000 units per day. The average weight of units is 130 kg, with a standard deviation of 10 kg. How many units are expected to weigh less than 142 kg?

Solution

Given $n = 2000 \quad \bar{X} = 130 \text{ kg} \quad \sigma = 10 \text{ kg}$

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

$$= \frac{142 - 130}{10}$$

$$= 12/10 = 1.2$$

From Appendix Table 1, we get $Z = 1.2$ gives an area of 0.1151. One-half of the area of the normal curve is 0.5.

Hence, $0.5 - 0.1151 = 0.3849$

Therefore, $0.3849 \times 2000 = 770$ units are expected to weigh less than 142 kg.

Example 10.40 As a result of tests on 20000 electric bulbs manufactured by a company, it was found that the lifetime of the bulbs was normally distributed, with an average life of 2040 hours and standard deviation of 60 hours. On the basis of the information, estimate the number of bulbs that are expected to burn for (a) more than 2150 hours, (b) less than 1960 hours.

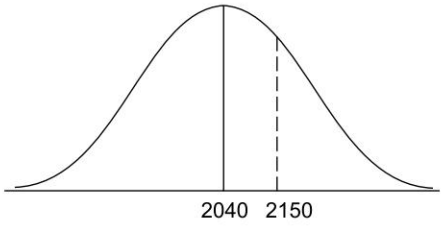
Solution

(a)

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

$$= \frac{2150 - 2040}{60}$$

$$= \frac{110}{60} = 1.83$$



When $Z = 1.83$, the area under the normal curve is 0.0336

$0.5 - 0.0336 = 0.4664 \times 20{,}000$

Number of bulbs expected to burn for more than 2150 hours = 9328

(b)

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

$$= \frac{1960 - 2040}{60}$$

$$= -\frac{80}{60} = -1.33$$

When $Z = 1.33$, the area under the normal curve is 0.0918

$0.5 - 0.0918 = 0.4082$

$0.4082 \times 20{,}000 = 8164$

Number of bulbs expected to burn for less than 1960 hours = 8164

GLOSSARY

Bernoulli process	One repetition of a binomial experiment. Also called a *trial*.
Binomial distribution	The probability distribution that gives the probability of x successes in n trials when the probability of success is p for each trial of a binomial experiment.
Continuity correction factor	The addition of 0.5 to and subtraction of 0.5 from the x value where x is the number of successes in n trials. It is a method of converting a discrete variable into a continuous variable.
Continuous probability distribution	A probability distribution in which the variable can take on any value within a given range.
Continuous random variable	A random variable that can assume any values within a given range.
Discrete probability distribution	A probability distribution in which the variable takes on only a limited number of values that can be listed.
Discrete random variable	A random variable can take a limited number of values that are countable.
Normal distribution	A symmetrical distribution with a single-peakad, bell-shaped curve and becomes sparse at the extremes. The two tails never touch the horizontal axis.
Poisson distribution	Like the binomial, but unlike the normal, it is a discrete probability distribution, that gives the probability of x (success) in an interval. It is appropriate when the probability of x is very small and n is large.
Probability distribution	A distribution of the probabilities associated with each of the values of a random variable. It is a theoretical distribution and is used to represent population.
Random variable	A variable that assumes a unique numerical value for each of the outcomes in a sample space of a probability experiment.
Standard normal probability distribution	A normal probability distribution, which has its mean as zero and the standard deviation as1.

LIST OF FORMULAE

1. *The binomial formula:*
Probability of 'r' successes in 'n' Bernoulli trials

$$= \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

where r = number of successes desired
n = number of trials
p = probability of success
q = probability of failure ($q = 1 - p$)

2. Mean of a binomial distribution $\mu = np$—Number of trials multiplied by the probability of success.

3. Standard deviation of a binomial distribution $s = \sqrt{npq}$ —Square root of the product of three terms—number of trials, probability of a success, and the probability of a failure.
4. *Poisson formula*:

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

The probability of x occurrences is equal to λ raised to the power x, multiplied by e (which is equal to 2.71828) raised to the negative λ power. The resultant numerator is to be divided by x factorial.

5. Poisson distribution as an approximation of the binomial

$$P(x) = \frac{(np)^x \times e^{-np}}{x!}$$

The mean of the Poisson distribution (λ) has been substituted by the mean of the binomial distribution (np). The approximation is good when $n \geq 20$ and $p \leq 0.05$.

6. $$z = \frac{x - \mu}{\sigma}$$

where x = value of the random variable x in which we are interested
μ = mean of the distribution of this variable x
s = standard deviation
z = number of standard deviations from x to the mean of this distribution

7. $$z = \frac{x - np}{\sqrt{npq}}$$

Normal approximation of the binomial distribution. Here, m has been substituted by np, being the mean of the binomial distribution, and s has been substituted by the standard deviation of the binomial distribution.

QUESTIONS

10.1 Given below are ten statements. Indicate in each case whether it is true or false.

(a) A distribution where the mean and the median have different values is not a normal distribution.
(b) The right and left tails of the normal curve always touch the horizontal axis.
(c) In a Bernoulli process, the probability of the outcome of any trial (toss) need not be fixed over time.
(d) In a Bernoulli process the trials must always be statistically independent.
(e) The standard deviation of a binomial distribution is $\sqrt{npq}$.
(f) The Poisson distribution is not a discrete probability distribution.
(g) The value of a random variable can be predicted in advance even before the occurrence of an event.
(h) A binomial distribution need not be symmetrical when the probability of success in a Bernoulli trial is $p = 0.5$.
(i) A normal curve is bell-shaped and has a single peak.
(j) In the formula used in Poisson distribution, the symbol λ stands for the mean.

Multiple Choice Questions (10.2 to 10.15)

10.2 If the expected profit of a business firm for January 2005 is Rs 10 lakh, then the profit for February will be
(a) Rs 10 lakh (b) Less than Rs 10 lakh
(c) More than Rs 10 lakh (d) None of the above

10.3 The standard deviation of a binomial distribution depends on
(a) probability of success (b) probability of failure
(c) number of trials (d) all the above

10.4 Which of the following conditions is *not* necessary for a distribution to be a binomial distribution?
(a) Each observation is classified in two categories.
(b) Probability of success (or failure) remains the same.
(c) Number of observations are large, i.e., more than 30.
(d) The trial of individual observations is independent of each other.

10.5 In the context of a binomial distribution, if on an average 8 ships out of 10 arrive safely at ports and 150 ships have returned safely, the mean is
(a) 80 (b) 100 (c) 120 (d) 150

10.6 A normal distribution curve
(a) is symmetrical
(b) has a single peak
(c) has its mean located at the centre
(d) has the same value of mean , median and mode
(d) all the above

10.7 Which of the following normal curves looks very similar to the curve for $\mu = 12$ and $\sigma = 4$?
(a) Curve for $\mu = 24$ and $\sigma = 8$ (b) Curve for $\mu = 12$ and $\sigma = 8$
(c) Curve for $\mu = 20$ and $\sigma = 9$ (d) Curve for $\mu = 24$ and $\sigma = 4$
(e) None of the above

10.8 A binomial distribution may be approximated by a Poisson distribution provided
(a) both n and p are large (b) both n and p are small
(c) n is small and p is large (d) none of these

10.9 Assuming a normal curve with $\mu = 40$ and $\sigma = 8$, how much area the curve will have to the right of the value 40?
(a) 0.25 (b) 1.0 (c) 0.75 (d) 0.5

10.10 In which of the following situations, the Poisson distribution can be a good approximation of the binomial distribution?
(a) $n = 300$ and $p = 0.04$ (b) $n = 60$ and $p = 0.52$
(c) $n = 60$ and $p = 0.35$ (d) all of these

10.11 A characteristic of probability distribution for any random variable is
(a) No given probability occurs again
(b) Each item in the series has a probability
(c) The probabilities of all individual item must add up to 1
(d) both (b) and (c) but not (a)

10.12 A binomial distribution can be symmetric provided
(a) $p > 0.5$ (b) $p < 0.5$
(c) p has any value (d) $p = 0.5$

10.13 The area covered by the normal curve within $\mu \pm 3\sigma$ limits is
(a) 68.27% (b) 100% (c) 99.73% (d) 95.45%

10.14 If a normal distribution has $\mu = 25$, then its mode is
(a) 15 (b) 50 (c) 25 (d) none of these

10.15 Suppose the probability of dialing a wrong number is 0.05. Then, the probability of dialling exactly 3 wrong numbers in 100 trials is
(a) 0.18 (b) 0.23 (c) 0.14 (d) 0.27

10.16 When does binomial distribution hold good? Explain with examples.

10.17 State the conditions under which the binomial distribution tends to **(i)** Poisson distribution, **(ii)** normal distribution.

10.18 The mean and the variance of a discrete variable (X) are 6 and 2, respectively. Assuming X to follow a binomial distribution, find: $P(5 \leq X \leq 7)$.

10.19 In a binomial distribution consisting of 5 independent trials, probabilities of 1 and 2 successes are 0.4096 and 0.2048, respectively. Find the parameter 'p' of the distribution.

10.20 For the binomial distribution with $n = 4$ and $p = 0.25$, find the probability of
(a) three or more successes
(b) at the most three successes
(c) two or more failures.

10.21 A company has 6 telephones, that 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' requirement of telephones is independent, the probability that 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.

10.22 The screws produced by a certain machine were checked by examining the number of deficiencies in a sample of 12. The following table shows the distribution of 128 samples according to the number of defective items they contain.

No. of defectives in a sample of 12	0	1	2	3	4	5	6	7	Total
No. of samples	7	6	19	35	30	23	7	1	128

(i) Fit a binomial distribution and find the expected frequencies if the chance of machine being defective is ½.
(ii) Find the mean and standard deviation of the fitted distribution.

10.23 A local electrical appliances shop has found from experience that the demand for tubelights is distributed as Poisson with a mean of 4 tubelights per week. If the shop keeps 6 tubes during a particular week, what is the probability that the demand will exceed supply during that week?

10.24 The number of defects per square foot of a certain manufactured fabric is Poisson distributed with $\lambda = 0.10$. What is the probability that the number of defects is **(i)** at least one; **(ii)** exactly two in two square feet fabric? (λ is the average number of defects per square foot).

10.25 A manufacturer finds that the average demand per day for the mechanics to repair new products is 2, over a period of one year and the demand per day is distributed as Poisson variate. He employs 3 mechanics. On how many days in one year: **(i)** both the mechanics will be free, and **(ii)** some demand is refused?

10.26 In the accounting department of the bank, 100 accounts are selected at random and examined for errors. Suppose the following results have been obtained:

No. of errors	0	1	2	3	4	5	6
No. of accounts	35	40	19	2	0	2	2

On the basis of above information, can it be concluded that the errors are distributed according to Poisson probability law?

10.27 Five hundred televisions 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 1, 2, 3 and 4 defects assuming Poisson distribution.
[Given $e^{-0.408} = 0.6649$]

10.28 In a Poisson frequency distribution, frequency corresponding to 3 successes is 2/3 times the frequency corresponding to 4 successes. Find the mean and the standard deviation.

10.29 The number of accidents in a year attributed to taxi drivers in a city follows Poisson distribution with mean 3. Out of 1000 taxi drivers, find approximately the number of drivers with **(i)** no accident in a year, and **(ii)** more than 3 accidents in a year.

10.30 Test at 5% level of significance whether the following data are in conformity with the assumption of a Poisson distribution.

Number of Breakdowns	*Number of Weeks*
0	19
1	58
2	59
3	40
4	32
5	12

10.31 If the number of complaints that a laundry receives per day is a random variable having the Poisson distribution $\lambda = 3.5$. Find the probabilities that on any given date the laundry will receive **(a)** no complaints, **(b)** exactly 2 complaints, **(c)** exactly 4 complaints.

10.32 A book containing 1000 pages has 0, 1, 2, 3 or 4 misprints per page as shown below:

Number of Misprints	*Number of Pages*
0	500
1	340
2	120
3	30
4	10
	1000

Fit the Poisson distribution to the above data and compare the theoretical frequencies with those given in the question.

10.33 What is a normal probability distribution? Explain the characteristic feature of a normal distribution.

10.34 Define and describe important properties of a normal distribution. Also, describe its importance in statistical analysis.

10.35 Find the probability that the value of an item drawn at random from a normal distribution with mean 20 and standard deviation 10 will be between **(i)** 10 and 15, **(ii)** –5 and 10, and **(iii)** 15 and 35.

10.36 In a large institution, 2.28 per cent of employees receive income below Rs 4,500 and 15.87% of employees receive income above Rs 7,500 pm. Assuming normal distribution, find $\bar{x}$ and σ of income.

10.37 A factory turns out an article by mass production methods. From the past experience it appears that 20 articles, on an average, are rejected out of every batch of 100. Find the variance of the number of rejects in a batch. What is the probability that the number of rejects in a batch exceeds 30?
(Given: area under a normal curve between $z = 0$ and $z = 2.5$ is 0.4938.)

10.38 The income of a group of 10,000 persons was found to be normally distributed with mean Rs 1,750 pm and standard deviation Rs 50. Show that of this group 95 per cent had income exceeding Rs 1,668 and only 5 per cent had income exceeding Rs 1832. What was the lowest income amongst the richest 100?

10.39 Two thousand students appeared in an examination. Distribution of marks is assumed to be normal with mean $\mu = 30$ and standard deviation $\sigma = 6.25$. How many students are expected to get marks **(i)** between 20 and 40, **(ii)** less than 35?

10.40 Suppose a paint manufacturer has a daily production, X, that is normally distributed with a mean of 100,000 gallons and a standard deviation of 10,000 gallons. The management wants to create an incentive bonus for the production crew when the daily production exceeds the 90^{th} percentile of the distribution, it hopes that the crew p will, in turn, become more productive. At what level of production should management pay the incentive bonus?

10.41 The customer accounts at a certain departmental store have an average balance of Rs 480 and a standard deviation of Rs 160. Assuming that the account balances are normally distributed

(i) what proportion of the accounts is over Rs 600?

(ii) what proportion of the accounts is between Rs 400 and Rs 600?

(iii) what proportion of the accounts is between Rs 240 and Rs 360?

10.42 The marks obtained by the MBA students in an examination are known to be normally distributed. If 31 per cent of the students got less than 45 marks while 8 per cent got over 64 marks, find the mean and standard deviation of the marks.

10.43 Six hundred candidates appeared for an entrance test for admission to a management course. The marks obtained by the candidates were found to be normally distributed with a mean of 152 marks and a standard deviation of 18 marks.

If the top 60 performers were given confirmed admission, what are the minimum marks (to the nearest integer) above which a candidate would be sure of being admitted?

Further, those obtaining at least 170 marks, but not qualified for confirmed admission were included in a provisional list. How many candidates were included in this list? (Answer to the nearest integer.)

10.44 For a Binomial distribution with P = 0.6 and $n = 50$, find the mean and variance. Fit a Poisson distribution to the following data and calculate the theoretical frequencies.

x	0	1	2	3	4
f	123	59	14	3	1

10.45 The following mistakes, per page, were observed in a book:

Number of mistakes per page	*Number of times the mistake occurred*
0	221
1	90
2	19
3	5
4	0
	325

Fit a Poisson distribution to the data and test the goodness of fit.

10.46 At an ATM centre, arrivals occur according to Poisson distribution, with a rate of 5 per hour. Service time per customer is exponentially distributed with mean 5 minutes.

(i) Find the expected number of customers using service.

(ii) For what percentage of time is the facility idle?

10.47 The following is the distribution of the trucks arriving hourly at a company's warehouse:

Trucks per hour:	0	1	2	3	4	5	6	7	8
Frequency:	52	151	130	102	45	12	5	1	2

Find the mean of this distribution and using it (rounded to one decimal place) as the parameter λ, fit a Poisson distribution.

Test for goodness of fit at the 0.05 level of significance.

10.48 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

10.49 The length of a machine part is known to have a normal distribution, with mean 100 mm and standard deviation 2 mm.

(i) What proportion of the parts will be above 103.3 mm?

(ii) What proportion of the parts will be between 98.5 and 102 mm?

(iii) What proportion of the parts will be shorter than 96.5 mm?

(iv) If no more than 5% of the parts should be oversized, what specification limit should be recommended?

10.50 The income of a group of 10,000 employees was found to be normally distributed, with mean Rs 1750 and standard deviation Rs 50.

(i) Show that 95% of the group had an income exceeding Rs 1668 and only 5% had an income exceeding Rs 1832. (Z at 95% is 1.64)

(ii) What was the lowest income of the richest 100 employees? (Z at 99% is 2.57).

The McGraw-Hill Companies

CHAPTER 11 SAMPLING AND SAMPLING DISTRIBUTIONS

Learning Objectives

By the end of your work on this chapter, you should be able to

- choose a simple random sample with the help of table of random numbers
- differentiate amongst some major sample designs
- understand the sampling distributions of sample mean and sample proportion
- determine appropriate sample size to estimate population mean or proportion for a given level of accuracy and with a prescribed level of confidence.

Chapter Prerequisites

Before starting work on this chapter, make sure you are fully conversant with

1. the binomial distribution
2. the normal distribution
3. the solution of equations

11.1 INTRODUCTION

In Chapter 1, it was mentioned that the data could be primary or secondary. Primary data are collected by conducting a survey. If the survey covers the entire population, then it is known as the census survey or complete enumeration. In contrast, if the survey covers only a part of a population, or a subset from a set of units, with the object of investigating the properties of the parent population or set, it is known as the sample survey. One may ask here: why is a sample survey necessary? Can we not undertake a census survey on a particular subject of enquiry? To answer these questions, let us see the relative advantages and limitations of sampling.

Advantages of Sampling

There are several advantages of sampling as given below:

1. A sample survey is cheaper than a census survey. It is obviously more economical, for instance, to cover a sample of households than all the households in a territory although cost per unit of study may be higher in a sample survey than in a census survey.

2. Since the magnitude of operations involved in a sample survey is small, both the execution of the field work and the analysis of the results can be carried out speedily.
3. Sampling results in a greater economy of effort as a relatively small staff is required to carry out the survey and to tabulate and process the survey data.
4. As compared to the census survey, more detailed information can be collected in a sample survey. In addition, information of a more specialised type can be collected, which would not be possible in a census survey on account of the availability of a small number of specialists.
5. Since the scale of operations involved in a sample survey is small, the quality of interviewing, supervision and other related activities can be better than that in a census survey.

Limitations of Sampling

1. When the information is needed on every unit in the population such as individuals, dwelling units or business establishments, a sample survey cannot be of much help for it fails to provide information on individual count.
2. Sampling gives rise to certain errors. If these errors are too large, the results of the sample survey will be of extremely limited use.
3. While in a census survey it may be easy to check the omissions of certain units in view of complete coverage, this is not so in the case of a sample survey.

As relative advantages of sample survey are decidedly more than the limitations, census surveys are seldom undertaken and the sample surveys are extremely common.

Parameter and Statistic

Samples and populations can be described by using measures such as the mean, median, mode, and standard deviation, which formed the subject-matter of Chapters 6 and 7. When any of these terms is used to describe the characteristic of a sample, it is called a *statistic*. When it is used to describe a characteristic of a population, it is called a *parameter*. In other words, a statistic is a characteristic of a sample while parameter is a characteristic of a population.

One of the objectives of sample surveys is to estimate certain population parameters. For example, we may be interested in determining the average annual expenditure on clothing in a city or the proportion of employees working overtime in a factory and so on. In the first example, parameter refers to the average annual expenditure on clothing and in the second example, the proportion of employees working overtime.

It may be noted that different notations are used to denote statistic and parameters as is shown in Table 11.1.

Table 11.1 Differences between Populations and Samples

	Population	*Sample*
Characteristics	Parameters	"Statistics"
Symbols	Population size = N	Sample size = n
	Population mean = μ	Sample mean = $\bar{x}$
	Population standard deviation = σ	Sample standard deviation = s
	Population proportion = p	Sample proportion = $\bar{p}$

A point to note is that the true value of a population parameter is an unknown constant. We can correctly determine it only by a complete study of the population. The concept of statistical inference comes into play whenever this is impossible or practically not feasible.

A statistic, which is a sample-based quantity, must serve as our source of information about the value of a parameter. In this context, there are three points that are crucial.

1. As a sample is only a part of the population, the numerical value of a statistic is normally not expected to give us the correct value of the parameter.
2. Since different samples can be drawn from a particular population, the observed value of a statistic depends on the particular sample that is chosen.
3. The value of a statistic will have some variability over different occasions of sample.

An example will make these points clear. Suppose we undertake a survey of households in a city. Here, the population comprises all the households in that city. We want to know how much expenditure is incurred annually on clothing per household. A random sample of 100 households is chosen and the survey is conducted. Suppose the survey results indicate that this amount is Rs 1,500. This means $\bar{x}$ = Rs 1,500. Evidently, the population mean μ cannot be claimed to be exactly Rs 1,500. If someone is to take another sample of 100 households, will he get the same sample mean of Rs 1,500? Obviously, it is highly unlikely that the two results would be identical. This is because the second sample would not have exactly the same households that were selected randomly in the first one. Likewise, the sample mean would also vary on different occasions of sampling. In practice, we observe only one sample such as $\bar{x}$ = Rs 1,500, which is used in our subsequent calculations.

As we will see later in this chapter that, in repeated random samples, the values of $\bar{x}$ tend to concentrate in the neighbourhood of the population mean μ. In view of this, it seems justified to use $\bar{x}$ as an estimate of μ.

11.2 RANDOM AND NON-RANDOM SAMPLES

There are two types of samples—*random* and *non-random*. Whether a sample is random or non-random depends on how it is selected. A sample is selected in such a way that every unit or element in the population has a given probability of being chosen. In other words, all samples of size n from a given population N have a given chance of being selected. Random samples are obtained by either sampling with replacement from a finite population or by sampling without replacement from infinite population.

When a random sample is drawn, it is necessary to ensure that each element has a given probability of being selected. However, many times mistakes crop up while selecting such a sample on account of the term random (which means a given chance of probability) is confused with haphazard (without pattern). As we shall see shortly, a random sample is chosen by following the proper procedure such as the use of a table of random numbers. It may be noted that the term random sample is not used to describe the data obtained through it but is used to indicate the process used in its selection. Thus, randomness is the property of the sampling process and not of a particular sample. A random sample is usually a representative sample.

A non-random sample, in contrast, does not follow the random procedure. Here, each element does not have an equal chance of being selected. A non-random sample is selected by a procedure other than probability considerations. For example, a sample may be chosen as per the convenience of the person selecting it. It may be selected by an expert on his judgment.

A major advantage of random sampling is that sampling variability can be determined whereas it is not possible to determine sampling variability in the case of a non-random sample.

We shall look into some sample designs, both within each broad category of random and non-random sampling, later in the chapter. But before we do so, we would like to know how many random samples could be drawn in a given situation.

11.3 ALL POSSIBLE RANDOM SAMPLES

In order to understand the concept of a random sample from a finite population, let us consider a finite population comprising 5 elements, which we call A, B, C, D and E. These 5 elements might be incomes of 5 managers, the prices of 5 brands of scooters and so on. To begin with, let us first find out how many different samples of size 3 can be taken from a finite population of size 5. In order to find the answer, we have to use the formula $\binom{n}{r}$, where n is the number of elements in the population and r is the proposed sample size. In case of the above example, n is 5 and r is 3. This means there are $\binom{5}{3}$ samples that can be selected.

Now,
$$\binom{n}{r} = \frac{n!}{(n-r)!\cdot r!} = \frac{5!}{(5-3)!\cdot 3!}$$
$$= \frac{5\times 4\times 3\times 2\times 1}{2\times 1\times 3\times 2\times 1} = 10$$

These 10 samples are: ABC, ABD, ABE, ACD, ADE, BCD, BDE, BCE, CED and EAC.

When we select one of the 10 samples in such a way that each has the same probability of being selected, then the sample so selected is the simple random sample, or more briefly, a random sample. The underlying idea in random sampling is that the selection of a random sample must, in some way, be left to chance. Let us take a few examples.

Example 11.1 A group of 10 students are to be divided into 2 groups of 5 each and seated at two tables. How many different ways are there of dividing the 10 students?

Solution

$$\binom{n}{r}\binom{10}{5} = \frac{10!}{(10-5)!\cdot 5!}$$
$$= \frac{10\times 9\times 8\times 7\times 6}{5\times 4\times 3\times 2\times 1} = 252$$

Example 11.2 Suppose there are 8 persons from whom we have to select samples of size 3. How many samples can be selected?

Solution

$$\binom{n}{r}\binom{8}{3} = \frac{8!}{(8-3)!\cdot 3!}$$
$$= \frac{8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}{5\times 4\times 3\times 2\times 1\times 3\times 2\times 1} = 56$$

Example 11.3 At a dinner party, 12 guests have been invited. They are to be divided into two groups of 6 each and seated at two tables. In how many different ways these guests can be seated?

Solution

$$\binom{n}{r}\binom{12}{6} = \frac{12!}{(12-6)!\cdot 6!}$$

$$= \frac{12\times 11\times 10\times 9\times 8\times 7}{6\times 5\times 4\times 3\times 2\times 1} = 924$$

Thus, there are 924 different ways of dividing the group.

Example 11.4 Suppose we are asked to find how many different random samples of size 3 can be selected from a finite population of size 50.

Solution

Here, we shall use the same approach, that is

$$\binom{n}{r}\binom{50}{3} = \frac{50!}{(50-3)!\cdot 3!}$$

$$= \frac{50\times 49\times 48}{3!} = 19{,}600$$

The process of selecting 19,600 samples would mean that we have to first list all the possible combinations of size 3 from our population of size 50. Such a procedure is, no doubt, extremely tedious, and elaborate. As such, we use an alternative procedure to get the identical results.

To obtain a random sample of size 3 from a population of size 50, we have to list each element on a slip of paper, mix the 50 slips of paper thoroughly, and then draw three slips in succession. This method, though more convenient than listing all the possible samples of the requisite size, can be further simplified in practice by using a table of random numbers. This necessitates the selection of the sample values one at a time, making sure that in each successive selection each of the remaining elements of the population has the same probability of being selected.

11.4 SAMPLING WITH AND WITHOUT REPLACEMENT

Before we proceed further, it is necessary at this stage to know that we can choose a simple random sample with and without replacement.

Sampling with Replacement

Sampling with replacement implies that an element once chosen is returned to the population and as such it is available to be chosen again. This means that an element can be selected more than once when sampling with replacement is adopted. Suppose we have a population of five elements A, B, C, D and E. We have to choose a sample of 2 elements from this population. Every element in the population has an equal chance of 1/N to be included in the sample. After having selected the first element and its value recorded, it is placed back in the population. Then we draw the second element exactly in the same manner as the first one. This element too will have 1/N probability of being selected in the sample. Whatever may be the sample size, we have to continue this process until we have selected the desired sample size of n elements.

Sampling Without Replacement

In case of sampling without replacement, before choosing the first element from the population, every element has an equal chance ($1/N$) of being included in the sample. But once the first element has been chosen, it is not replaced in the population. This means that the second element is selected from the remaining $N - 1$ elements of the population. Hence, each element has a chance of $1/(N - 1)$ to be included in the sample. When the third element is to be drawn, the probability of each of the remaining elements is $1/(N - 2)$ and so on until the n^{th} draw when the probability is $1/N - (n - 1)$, i.e., $1/(N - n + 1)$.

We now discuss the process of selecting a simple random sample.

11.5 SELECTING A SIMPLE RANDOM SAMPLE

The process of randomness does not mean that it is haphazard, as a layman may be inclined to think. What it means is that the process of selecting a sample is independent of human judgment. To ensure this, there are two methods that are followed when drawing a random sample. These are: **(i)** the lottery method and **(ii)** the use of random numbers.

Use of Lottery Method

In the lottery method, each unit of the population is numbered and shown on a chit of paper or disc. The chits are folded and put in a box from which a sample of the requisite size is to be drawn. In case discs are used, these are mixed up well before a draw is made so that no particular unit can be identified before it gets selected.

Use of Random Numbers

In this method, the tables of random numbers are used. The units of the population are numbered from 1 to N from which n units are selected. This process is explained below with the help of an example.

Example 11.5 Suppose a sample of size 10 is to be selected from a population of 200. Describe the procedure to be followed in choosing such a sample.

Solution First, number the 200 units from 1 to 200, the order being quite immaterial. While numbering the units, ensure that each unit in the population has uniform digits, in this case, three. Thus, first unit would have a three-digit number 001, second unit 002, 10^{th} unit 010, and so on. After the units have been given three-digit numbers, the table of random numbers is to be used. Appendix Table 12 gives two thousand random numbers. One way is to start from the top left-hand corner of the table of random numbers and proceed down systematically.

In fact, the table should be called as the table of random digits, which are given in two's. Any three-digit number that is more than 200 is to be discarded since our population size is 200. Another point to note is that if a certain figure occurs the second time, that too is to be ignored. Table 11.2 gives a portion of the Appendix Table 12.

Table 11.2 **A Portion of Appendix Table 12 (Random Numbers)**

	1–4	*5–8*	*9–12*
1	23 15	75 48	59 01
2	05 54	55 50	43 10
3	14 87	16 03	50 32
4	38 97	67 49	51 94
5	97 31	26 17	18 99
6	11 74	26 93	81 44
7	43 36	12 88	59 11
8	93 80	62 04	78 38
9	49 54	01 31	81 08
10	36 76	87 26	33 37
11	07 09	25 23	92 24
12	43 31	00 10	81 44
13	61 57	00 63	60 06
14	31 35	28 37	99 10
15	57 04	88 65	26 67
16	09 24	34 42	00 68
17	97 95	53 50	18 40
18	93 73	25 95	70 43
19	72 62	11 12	25 00
20	61 02	07 44	18 45
21	97 83	98 54	74 33
22	89 16	09 71	92 22
23	25 96	68 82	20 62
24	81 44	33 17	19 05
25	11 32	25 49	31 42

Using this table, we find that the first three-digits number is 231, as such it is to be discarded. In the second row, we find the number as 055. Since it does not exceed 200, this is our first number in the sample. In the third row, the number is 148, which is to be included in the sample. Thus, we proceed in this manner until we have selected 10 numbers. It may be noted that we could have proceeded horizontally from left to the right by using the table of random numbers. While selecting our sample, we found that when we cover the last row (which gave us number 113), our sample size of 10 was not complete. We, therefore, continued from the first row with random digits 4, 5 and 6. Table 11.3 gives the desired sample size selected by using random numbers as explained above.

Table 11.3 **Desired Sample of Size Ten**

055	148	117	070	092
113	126	062	100	035

11.6 OTHER SAMPLE DESIGNS

We have seen in the preceding sections, the nature of simple random sampling and how it can be selected. In this section, we discuss other sampling designs including non-random samples.

Systematic Sampling

In practice, the method followed in systematic sampling is simpler than that explained earlier. *First*, a sampling interval k is calculated. Suppose we have to select a sample of 50 out of 500 units, then we calculate the sampling interval k (N/n), where N is the total number of units in the population and n is the size of the sample. In our example, k is 500/50=11. *Second*, a number between 1 and 10 is chosen at random. Suppose the number thus selected happens to be 9, then the sample will comprise numbers 9, 19, 29, 39, 49, … 489 and 499.

It will be seen that it is extremely convenient to select a sample in this way. The main point to note is that once the first unit in the sample is selected, the selection of subsequent units in the sample becomes obvious. In view of this, it has been questioned whether the process of selection of subsequent units is random. Here, the selection of a unit is dependent on the selection of a preceding unit in contrast to simple random sampling where the selection of units is independent of each other. In view of this, *systematic sampling is sometimes called quasi-random sampling*.

Stratified Random Sampling

A stratified random sampling is one where the population is divided into mutually exclusive and mutually exhaustive strata or sub-groups and then a simple random sample is selected within each of the strata or sub-groups. Thus, human population may be divided into different strata on the basis of sex, age groups, occupations, education or regions. It may be noted that stratification does not mean absence of randomness. All that it means is that the population is first divided into certain strata and then a simple random sample is selected within each stratum of the population.

The following example will make this clear.

Table 11.4 Break-up of Population and Sample by Income Group

Income Strata (Rs '000)	*Population (No. of Households)*	*Sample (Proportionate)*	*Sampling Ratio*
Below – 5	10,000	500	0.05
5–10	12,000	600	0.05
10–15	16,000	800	0.05
15–20	14,000	700	0.05
20 +	8,000	400	0.05
	60,000	**3,000**	**0.05**

In the above example, the population consists of 60,000 households, divided into five income groups. Column 3 shows the sample, that is the number of households selected from each stratum. The sample constitutes 5 per cent of the population. A sample of this type, where each stratum has a uniform sampling fraction, is called a proportionate stratified sample. The last column of the table shows the sampling ratio, which is uniform for each stratum.

It may be noted that a stratified random sample with a uniform sample fraction results in greater precision than a simple random sample. But, this is possible only when the selection within strata is made on a random basis. Further, a stratified proportionate sample is generally convenient on account of practical considerations.

If we want to select a proportionate stratified sample of size n from a population of size N, which has been stratified into, say, 4 strata with sizes N_1, N_2, N_3 and N_4 respectively, then the sample sizes for different strata will be:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \frac{n_4}{N_4} = \frac{n}{N}$$

The strata and the samples from each stratum are shown in the form of a Venn diagram in Fig. 11.1 where S_1, S_2, S_3 and S_4 are the four strata.

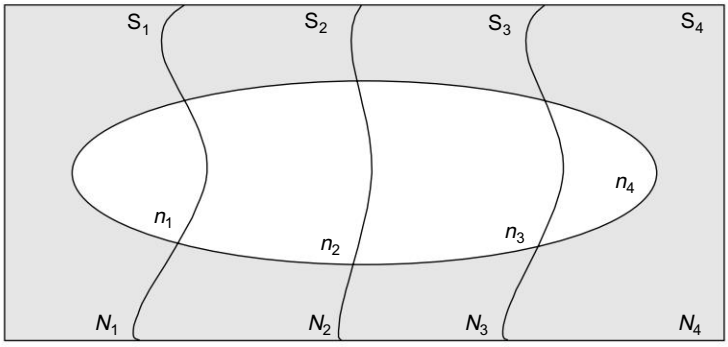


Fig. 11.1 **Stratified Sampling**

Disproportionate Stratified Sampling

In the case of a disproportionate stratified sampling, different strata will not have the same sampling ratio. Given below is Table 11.5 wherein disproportionate sampling has been chosen.

Table 11.5 **Break-up of Population and Sample by Income-group**

Strata Income Per Month ('000 Rs)	*Population No. of Households*	*Sample (Disproportionate)*	*Sampling Ratio*
Below 5	10,000	600	0.060
5–10	12,000	800	0.067
10–15	16,000	800	0.050
15–20	14,000	500	0.036
20+	8,000	300	0.037
	60,000	**3,000**	**0.050**

A point to note in the case of disproportionate sample is that the sample size should be more in a stratum where variance is more than in a stratum where variance is less. It should be obvious that in the case of a stratum, which has a good deal of homogeneity, a very small size of sample is enough.

Advantages of Disproportionate Stratified Sampling

1. At times, it may be preferable to use variable sampling fractions, resulting in disproportionate stratified sampling. When the population in some strata is more heterogeneous than in others, it may be advisable to use variable sampling fractions. The reason is that the use of a uniform sampling fraction may not lead to 'representative' samples in such strata. As such, larger sampling fractions may be used in strata with greater variability.

2. Another reason for using disproportionate stratified sampling fraction may be the higher cost per sampling unit in some strata compared to the others. In such a situation, precision can be increased by taking a smaller fraction in the costlier strata and a higher fraction in the cheaper strata. An optimum precision can be obtained for a given cost if the sampling fractions in the different strata happen to be proportional to their standard deviations and inversely proportional to the square root of the costs per unit in the strata.

Cluster Sampling

Another form of sampling is cluster sampling, which can be much cheaper than other forms of sampling allowing a larger sample size to be collected. In cluster sampling, instead of selecting individual units from the population, entire groups or clusters are selected at random. Let us illustrate this method by taking an example. Figure 11.2 shows clusters or blocks in a city.

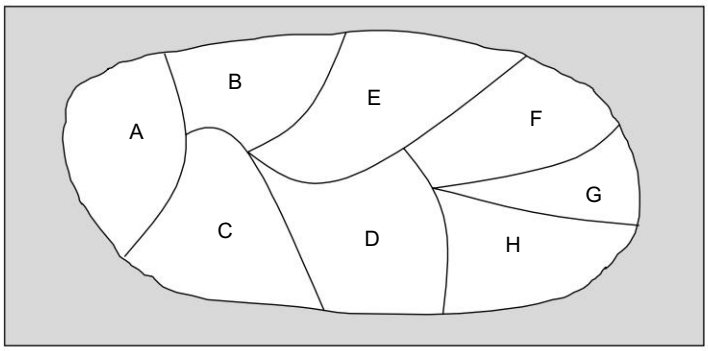


Fig. 11.2 **Clusters in a City**

Suppose a business firm, engaged in the manufacture of sports goods, is interested to know the changing patterns of family expenditures for recreation in a metropolitan city. The firm finds it difficult to use simple random sampling as suitable lists of families are not available apart from excessive cost of contacting families spread over a wide area. Faced with such a situation, the firm may use an alternative method. It may divide the total area of interest into a number of smaller areas or blocks, ensuring that they are not overlapping. A number of these blocks can then be selected by using random numbers. It can then collect the requisite information from all (or samples of) the families residing in these blocks. If these blocks or clusters are geographic subdivisions, as is the case in this example, then this type of sampling is also called *area sampling.*

Features of Cluster Sampling A few points regarding cluster sampling may be noted here.

1. Whether or not a particular aggregate of units should be called a cluster will depend on circumstances of each case. In the foregoing example, blocks were taken as clusters and families as individual units. In another case, the families may be taken as a cluster and the members of the families as individual units.
2. It is not necessary that clusters should always be natural aggregates such as blocks, polling constituencies, schools or colleges. Artificial clusters may be formed, as is generally done in area sampling, where grids may be determined on the maps.
3. Several levels of clusters may be used in any one sample design. Thus, in a city survey, localities or wards, streets and families may be selected in which case localities or wards are the clusters at the first level and streets at the second level.

Limitations of Cluster Sampling Estimates based on cluster samples are generally not as reliable as those based on simple random samples of the same size. However, in terms of per unit cost, they are more reliable. It is easy to see that at the same amount of expenditure, it is possible to select a cluster sample several times the size of a simple random sample. A major limitation of cluster sampling is the high degree of intra-cluster homogeneity. On account of the similarity of one unit in the cluster with the other units, selection of a few clusters may not give a really representative sample. As against this, when clusters have a high degree of intra-cluster heterogeneity, cluster sampling may be more representative. Finally, we should know that when each group has small variation within itself but there is a wide variation between the groups, a stratified sample should be preferred. In contrast, when there is a wide variation within each group but groups are very similar to each other, then a cluster sample should be preferred.

Since both stratified sampling and cluster sampling have some subgroups, a pertinent question may be asked: How do they differ from each other? The following three considerations clearly bring out the difference between the two types of sample designs:

Stratified Sampling	*Cluster Sampling*
1. Population is divided into a few subgroups, each having many elements. The selection of subgroups is made on the basis of some criterion, which is related to the variables under study.	**1.** Population is divided into many subgroups, each having a few elements. Here, the selection of subgroups is made on the basis of some criterion of ease or availability of requisite data.
2. In selecting the sample, homogeneity within subgroups and heterogeneity between subgroups are aimed at.	**2.** Here, the approach is just the opposite as in selecting the sample heterogeneity within subgroups and homogeneity between subgroups are aimed at.
3. Elements are randomly selected from within each subgroup.	**3.** Only a few subgroups are randomly selected and then all elements within those subgroups are covered in the study.

Multi-stage Sampling

Multi-stage sampling, as the name implies, involves the selection of units in more than one stage. In such a sampling, the population consists of number of first-stage units, called primary sampling units (PSUs). Each of these PSUs consists of a number of second-stage units. First, a sample is taken of the PSUs, then a sample is taken of the second-stage units. This process continues until the selection of the final sampling units. It may be noted that at each stage of sampling, a sample can be selected with or without stratification.

An illustration would make the concept of multi-stage sampling clear. Suppose a sample of 5000 urban households from all over the country is to be selected. In such a case, the first-stage sample may involve the selection of districts. Suppose 25 districts out of say 500 districts are selected. The second stage may involve the selection of cities, say four from each district. Finally, 50 households from each selected city may be chosen. Thus, one would have a sample of 5000 urban households, arrived at in three stages. It is obvious that the final sampling unit is the household.

In the absence of multi-stage sampling of this type, the process of the selection of 5000 urban households from all over the country would be extremely difficult. Besides, such a sample would be very thinly spread over the entire country, and if personal interviews are to be conducted for collecting information, it will be an extremely costly affair. In view of these considerations, a sampling from a widely spread population is generally based on multi-stage.

The number of stages in a multi-stage sampling varies depending on convenience and the availability of suitable sampling frames at different stages. Often, one or more stages can be further included in order to reduce cost. Thus, in our earlier example, the final stage of sampling comprised 50 households from each of the four selected cities. Since this would involve the selection of households all over the city it would turn out to be quite expensive and time consuming if personal interviews are to be conducted. In such a case, it may be advisable to select two wards or localities in each of the four selected cities and then to select 25 households from each of the 2 selected wards or localities. Thus, the cost of interviewing as also the time in carrying out the survey could be reduced considerably. It will be seen that an additional stage comprising wards or localities has been introduced here. Thus, this sample has become a four stage sample—

1st stage—districts
2nd stage—cities
3rd stage—wards or localities
4th and final stage—households

From the preceding discussion, it should be clear that a multi-stage sample results in the concentration of field work. This in turn, leads to saving of time, labour and money. There is another advantage in its use. Where a suitable sampling frame covering the entire population is not available, a multi-stage sample can be used.

Area Sampling

Area sampling is a form of multi-stage sampling in which maps, rather than lists or registers, are used as the sampling frame. This method is more frequently used in those countries which do not have a satisfactory sampling frame such as population lists.

In area sampling, the overall area to be covered in a survey is divided into several smaller areas within which a random sample is selected. Thus, for example, a city map can be used for area sampling. Various blocks can be identified on the map and this can provide a suitable frame. The entire city area can be divided into these blocks which are then numbered and from which a random sample is finally drawn.

In sampling the blocks, stratification and sampling with probability proportional to a measure of size are commonly employed. However, stratification in area sampling is based on geographical considerations. Thus, when blocks are identified and numbered on the map, they can be grouped into some meaningful strata representing the different neighbourhoods of the town. The point to emphasise is that these blocks must be identifiable without any difficulty.

On the basis of the blocks thus identified, numbered and assigned to strata, a stratified sample of dwellings can be selected. This can be done in either of two ways. First, a sample of dwellings may be drawn from all the dwellings included in a selected block. Second, blocks may be divided into segments of a more or less equal size, and a sample of these segments can be chosen and finally all the dwellings from the selected segments may be taken in the sample. It will thus be seen that the second method introduces another stage of sample, namely, segments.

Non-probability Sample Designs

A non-probability sample is also known as a non-random sample. Here, each element in the population does not have an equal chance of being selected. We briefly discuss below three types of non-probability samples, viz., *quota sampling*, *judgement sampling* and *convenience sampling*.

Quota Sampling Quota sampling involves the fixation of certain quotas, which are to be fulfilled by the interviewers. Suppose that in a certain territory we want to conduct a survey of households. Their total number is 2,00,000. It is required that a sample of one per cent, that is, 2,000 households is to be covered. We may fix certain controls.

A sample of this size can be selected subject to the condition that 1,200 households should be from rural areas and 800 from the urban areas of the territory. Likewise, another quota can be fixed. Of the 2,000 sample households, rich households should number 150, the middle class ones 650 and the remaining 1,200 should be from the poor class.

In view of the non-random element of quota sampling, it has been severely criticized especially as it is theoretically weak and unsound. But in marketing research, it is frequently used on account of its being convenient and economical.

There are points both in favour of and against quota sampling. These are given below.

Advantages of Quota Sampling

1. It is economical as travelling costs can be reduced. An interviewer need not travel all over a town to track down pre-selected respondents. However, if numerous controls are employed in a quota sample, it will become more expensive though it will have less selection bias.
2. It is administratively convenient. The labour of selecting a random sample can be avoided by using quota sampling. Also, the problem of non-contacts and call-backs can be dispensed with altogether.
3. When the field work is to be done quickly, perhaps in order to minimise memory errors, quota sampling is most appropriate and feasible.
4. It is independent of the existence of sampling frames. Wherever a suitable sampling frame is not available, quota sampling is perhaps the only choice available.

Limitations of Quota Sampling

1. Since quota sampling is not based on random selection, it is not possible to calculate estimates of standard errors for the sample results.
2. It may not be possible to get a 'representative' sample within the quota as the selection depends entirely on the mood and convenience of the interviewers.
3. Since too much latitude is given to the interviewers, the quality of work suffers if they are not competent.
4. It may be extremely difficult to supervise and control the field investigation under quota sampling.

Judgment Sampling The main characteristic of judgment sampling is that units or elements in the population are purposively selected. It is because of this that judgment samples are also called *purposive samples*. Since the process of selection is not based on the random method, a judgment sample is considered to be non–probability sampling.

Occasionally, it may be desirable to use judgment sampling. Thus, an expert may be asked to select a sample of 'representative' business firms. The reliability of such a sample would depend upon the judgment of the expert. The quota sample, discussed earlier, is in a way judgment sample where the actual selection of units within the earlier fixed quota depends on the interviewer.

It may be noted that when a small sample of a few units is to be selected, a judgment sample may be more suitable as the errors of judgment are likely to be less than the random errors of a probability sample. However, when a large sample is to be selected, the element of bias in the selection could be quite large in the case of a judgment sample. Further, it may be costlier than the random sample.

Convenience Sampling Convenience sampling, as the name implies, is based on the convenience of the statistician who is to select a sample. This type of sampling is also called *accidental sampling*, as the respondents in the sample are included in it merely on account of their being available on the spot where the survey is in progress. Thus, a person may stand at a certain prominent point and interview all those or selected people who pass through that place. A survey based on such a sample of respondents may not be useful if the respondents are not representative of the population. It is not possible in convenience sampling to know the 'representativeness' of the selected sample. As such, it

introduces an unknown degree of bias in the estimate. In view of this major limitation, convenience sampling is generally avoided in serious studies.

11.7 SAMPLING DISTRIBUTION OF A STATISTIC

In this chapter, we have so far discussed how samples can be selected from populations as well as some major sample designs. If we take several samples from a population, the statistic for each sample need not be the same. In all probability, such statistic would vary from sample to sample. The fact that the value of the sample mean, or any other statistic, will vary as the sample process is repeated, is a basic concept. Any statistic, the sample mean in particular, varies from sample to sample, is a random variable and has its own probability distribution. It may be noted that the use of the word 'sampling' indicates that the distribution is conceived in the context of repeated sampling from a population though in practice the word 'sampling' is dropped and it is called simply the distribution of a statistic.

11.8 SAMPLING DISTRIBUTION OF MEAN

It was mentioned earlier in this chapter that the population parameter is constant. For example, for any given series, there can be only one value of μ. But in case of a sampling distribution, we cannot say that the sampling mean is constant. As different samples of the same size can be selected from the same population, they will give different values of the sample mean $\bar{x}$. The value of sample mean for any one sample will depend on the elements included in that sample. This means that the sample mean $\bar{x}$ is a random variable and it possesses a probability distribution, which is usually known as the sampling distribution of $\bar{x}$. Other sample statistics such as the median, mode, and standard deviation also possess sampling distributions.

Let us take an example. Suppose there are 5 workers—A, B, C, D and E—comprising the population. Their monthly wages in thousand rupees are:

A	*B*	*C*	*D*	*E*
3	5	7	7	8

Further, suppose we have to select a sample of size 3 out of 5 workers. As we have seen earlier, there can be in all 10 samples of size 3 out of a population size of 5. Table 11.6 shows all possible samples along with their means.

Table 11.6 All Possible Samples and their Means

(*Sample size being 3*)

Sample	*Wages in the Sample ('000 Rs)*	*Sample Mean* $\bar{x}$
ABC	3,5,7	5.00
ABD	3,5,7	5.00
ABE	3,5,8	5.33
ACD	3,7,7	5.67
ACE	3,7,8	6.00
ADE	3,7,8	6.00
BCD	5,7,7	6.33
BCE	5,7,8	6.67
BDE	5,7,8	6.67
CDE	7,7,8	7.33

(The values of $\bar{x}$ *are rounded to two decimal places.)*

On the basis of values of $\bar{x}$ given in Table 11.6, we can form a frequency distribution of $\bar{x}$ as shown in Table 11.7.

Table 11.7 Frequency Distribution of $\bar{x}$ when the Sample Size is 3

$\bar{x}$	*Frequency*
5.00	2
5.33	1
5.67	1
6.00	2
6.33	1
6.67	2
7.33	1
	10

Now, by dividing the frequencies of different $\bar{x}$ values by the total number of frequencies, we obtain the relative frequencies of these classes. The relative frequencies are used as probabilities of classes, as shown in Table 11.8.

Table 11.8 Sampling Distribution of $\bar{x}$ when the Sample Size is 3

$\bar{x}$	$P(\bar{x})$
5.00	2/10 = 0.2
5.33	1/10 = 0.1
5.67	1/10 = 0.1
6.00	2/10 = 0.2
6.33	1/10 = 0.1
6.67	2/10 = 0.2
7.33	1/10 = 0.1
	$\Sigma P(\bar{x}) = 1.0$

It will be clear that if we draw just one sample of size 3 from the population of size 5, any of the 10 possible samples can be drawn. This means that the sample mean $\bar{x}$ can have any of the values shown in Table 11.8 with the corresponding probability. For example, if we draw a sample with $\bar{x} = 6.67$, then the probability of drawing such a sample is 0.2. This can be written as $P(\bar{x} = 6.67) = 0.2$.

11.9 SAMPLING AND NON-SAMPLING ERRORS

Sampling Errors

Whenever we undertake a sample survey, there may arise two types of error, viz., sampling error and non-sampling error. As we have seen earlier, a number of samples of the same size can be selected from a population. Different samples selected from the same population will give different results as the elements included in the sample will be different. When there is a difference between a sample

statistic, say, sample mean and the population mean, then that difference is known as the sampling error provided there is no non-sampling error. It should be obvious here that if we undertake a census survey, which implies that all the elements in a population are covered, the question of sampling error does not arise at all. Ordinarily, when a sample is chosen, we find that the difference between a sampling statistic and the population parameter consists of both sampling and non-sampling errors.

Types of Sampling Errors Sampling errors can be of two types—*biased* and *unbiased*. Biased errors arise on account of any bias in selection, estimation, etc. One may adopt a wrong process of, selection or during the survey, one may record wrong data. Further, after the survey is over, wrong methods of analysis may be used. As long as there is some bias, the survey results cannot be regarded as objective conclusions. These may turn out to be misleading. In view of this, the statistician should endeavour to make the survey as objective as possible. One way to do so is to draw a sample of adequate size on completely random basis. However, in some cases some restrictions may have to be imposed while choosing a random sample. In such cases, one should ensure that such restrictions do not introduce bias in the results. A point worth emphasizing here is that sampling errors can be minimised by increasing the sample size. Of course, there is substantial reduction initially when the sample size is increased, subsequently the reduction in sampling error is not much. As such one may choose sample size in such a way that the survey results are within the permissible limits of error.

Non-sampling Errors

Even when a census survey is undertaken, there may arise several types of errors, which are known as non-sampling errors. Such errors can occur both in the census and sample survey though they are likely to be more in the census survey. These errors can occur at any stage of the survey—right from the planning stage to the execution of the survey. For example, the survey may use a defective method of data collection, or on account of entry of wrong data in the tabulation an error may occur.

Causes of Non-sampling Errors There can be numerous causes which contribute to non-sampling errors. Some of these are:

1. Using imprecise definition or wrong concept while launching the survey.
2. Entrusting the survey work to untrained and inexperienced investigators.
3. Despatching a defective mail questionnaire to respondents who may not clearly understand certain questions.
4. Errors that may arise on account of non-response from respondents.
5. Poor supervision of the field staff.
6. Faulty tabulation while transferring the questionnaire data to tabulation sheets.
7. Calculation mistakes in the processing and analysis of data.
8. Committing mistakes while oral or written presentation of the survey results.

It will be seen that these are several causes responsible for non-sampling errors. Those responsible for organising and conducting a survey, whether census or sample, must be quite vigilant in carrying out their jobs. Adequate instructions must be given pointing out the possible areas where errors may occur, to the concerned staff. In this way, one can minimise the occurrence of both sampling and non-sampling errors.

11.10 SAMPLING FROM NORMAL POPULATIONS

The sampling distribution of a mean of a sample taken from a normal population shows two important properties. *First*, the sampling distribution has a mean that is equal to the population mean. In symbols, $\mu_{\bar{x}} = \mu$. *Second*, the sampling distribution has a standard deviation (a standard error) that is equal to the population standard deviation divided by the square root of the sample size. This can be written in symbolic form

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ = *standard error of the mean*
σ = *population standard deviation*
n = *sample size*

Let us take an example to explain further these properties of the sampling distribution of a mean.

Example 11.6 Suppose in a normally distributed population, average income per household is Rs 10,000 pm with the standard deviation of Rs 800. A survey based on a random sample of 100 households is undertaken. What is the probability that the sample mean will be between Rs 9,800 and Rs 10,100?

Solution As this is a question relating to the sampling distribution of the mean, we have to first calculate the standard error of the mean. In order to calculate it, we use the formula given earlier:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Substituting the two values in the above formula, we get

$$\sigma_{\bar{x}} = \frac{800}{\sqrt{100}}$$

= Rs 80—Standard error of the mean.

Having obtained the standard error of the mean, we now use the following formula

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

This formula transforms any normal random variable to a standard normal random variable. Since we have been given the range within which the sample mean will lie, we have to use the formula for the two values.

For $\bar{x} = 9{,}800$, $Z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}}$

$$= \frac{\text{Rs } 9{,}800 - 10{,}000}{\text{Rs } 80}$$

$$= -2.5$$

For $\bar{x} = 10{,}100$, $Z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}}$

$$= \frac{\text{Rs } 10{,}100 - \text{Rs } 10{,}000}{\text{Rs } 80}$$

$$= 1.25$$

Now, we have to use the table on the Standard Normal Probability Distribution (Appendix Table 1). We find that the table gives an area of 0.0062 corresponding to a Z value of –2.5, and an area of 0.1056 corresponding to a Z value of 1.25. It may be noted that the former figure relates to the left-tail of the normal curve while the latter figure relates to its right-tail. As such, each figure is to be subtracted from 0.5 which is the total area of one-half of the curve. Thus, we get 0.5 – 0.0062 = 0.4938 and 0.5 – 0.1056 = 0.3944. Adding these two values, we get

$$0.4938 + 0.3944 = 0.8882$$

This is the probability that the sample mean will lie between Rs 9,800 and Rs 10,100. This is shown in Fig. 11.3.

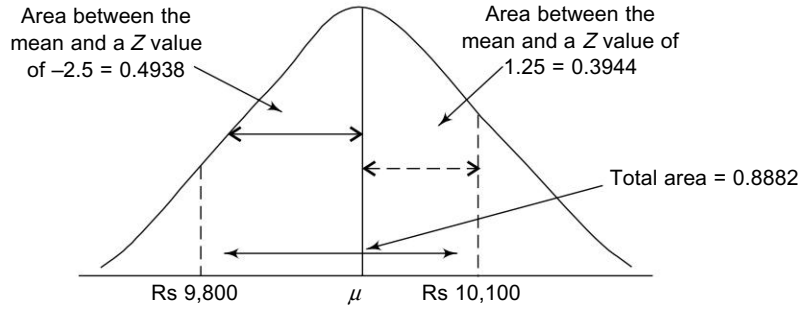


Fig. 11.3 **Probability of Sample Mean**

11.11 SAMPLING FROM NON-NORMAL POPULATIONS

The foregoing example was related to a normal population. It showed that in case of a normal population, the sampling distribution of the mean is also normal. However, in most of the cases the population from which a sample is taken is not normally distributed. In such cases, we use an important theorem to infer the shape of the sampling distribution of the mean.

The Central Limit Theorem

The central limit theorem states that as sample size gets large enough, the sampling distribution of the mean can be approximated by the normal distribution. This is true regardless of the distribution of the population from which the random sample is drawn.

Symbolically, the mean of the sampling distribution $\bar{x}$ *is* $\mu_{\bar{x}} = \mu$ *and the standard deviation is*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

On the basis of this theorem we can make probability statements about the possible range of values the sample mean may take.

The question here is: which sample size is considered large enough? As we do not know the shape of the population distribution, it is necessary for us to apply some general rule which may indicate as to when a sample is large enough so that the central limit theorem can be applied. In general, when a sample comprises 30 or more elements, it is considered large enough for the application of the central limit theorem.

A point worth noting is that when the population distribution is very different from a normal distribution, it is necessary to have a large minimum sample size for a good normal approximation. By the same token, when the population distribution is close to a normal distribution, even a smaller minimum sample size is considered sufficient. In short, the larger the sample size, the better is the approximation to the normal distribution and vice versa.

According to the central limit theorem, for a large sample size, the sampling distribution of the sample mean $\bar{x}$ is approximately normal, regardless of the shape of the population distribution. Symbolically, the mean of the sampling distribution $\bar{x}$ is $\mu_{\bar{x}} = \mu$ and the standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

By 'large sample size', we mean that n is ≥ 30.

Figure 11.4 shows a number of situations with varying sample sizes. It will be seen from Fig. 11.4(a) that the population is not normally distributed and it is positively skewed. Figure 11.4(b), (c) and (d) show the shape of sampling distributions with different sample sizes. The figures show that as the sample size increases, the shape of the sampling distribution undergoes change and it tends to be more and more approximate to normal distribution. Another point to note is that the central limit theorem is applicable to large samples only, that is, when the sample size is 30 or more.

Let us take an example.

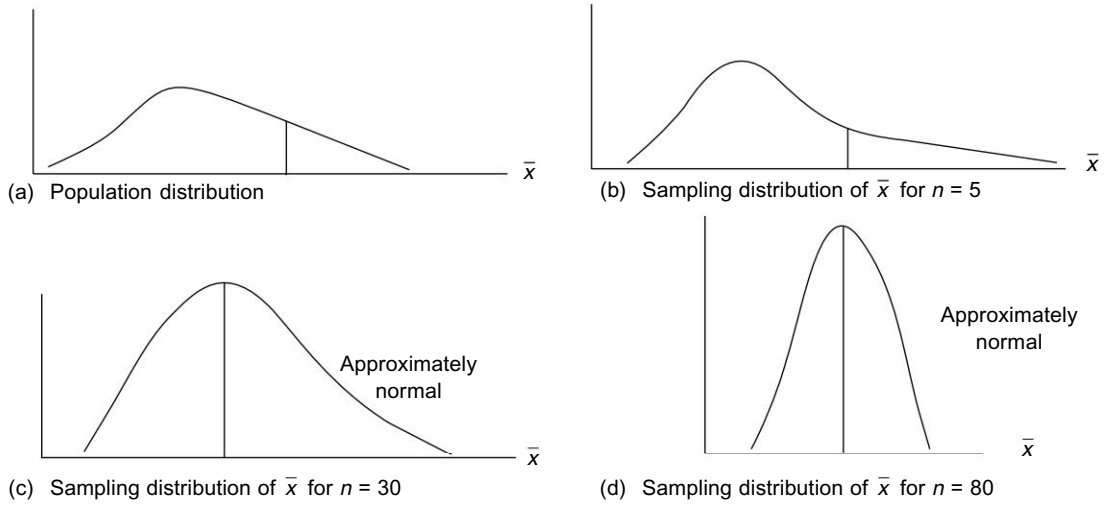


Fig. 11.4 Population Distribution and Sampling Distributions of $\bar{x}$

Example 11.7 In a certain locality, the average rent paid by all tenants amounts to Rs 1,500 pm with a standard deviation of Rs 450. However, the population distribution of rents pertaining to all tenants in that city is positively skewed. Find out the mean and standard deviation of $\bar{x}$ when the sample size is **(a)** 30 and **(b)** 100. Also describe the shape of its sampling distribution in both the cases.

Solution

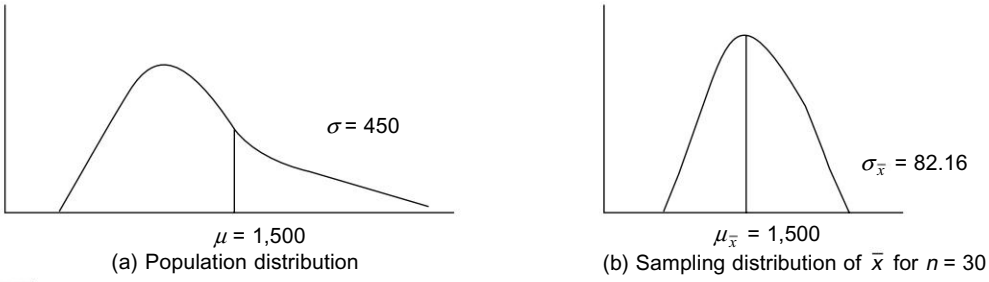
Given that the population distribution is not normal, but the sample size in both the cases is large as $n \geq 30$, the central limit theorem can be applied to infer the shape of the sampling distribution of $\bar{x}$.

(a) Let $\bar{x}$ be the average rent paid by a sample of 30 tenants. Then, the sampling distribution of $\bar{x}$ is approximately normal with the values of the mean and standard deviation as

$\mu_{\bar{x}} = \mu =$ Rs 1,500 and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

$= 450/\sqrt{30} = 82.16$

Figure 11.5 shows the population distribution and the sampling distribution of $\bar{x}$.



(a) Population distribution (b) Sampling distribution of $\bar{x}$ for $n = 30$

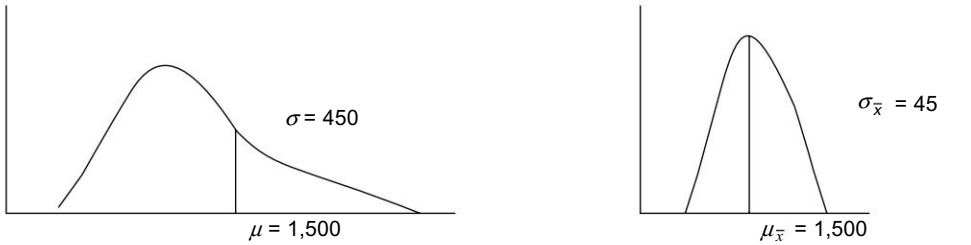
Fig. 11.5 Population Distribution and Sampling Distribution

(b) Let $\bar{x}$ be the average rent paid by a sample of 100 tenants. Then, the sampling distribution of $\bar{x}$ is approximately normal with the values of the mean and standard deviation as

$\mu_{\bar{x}} = \mu =$ Rs 1,500 and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

$= 450/\sqrt{100} = 45$

Figure 11.6 shows the population distribution and the sampling distribution of $\bar{x}$.



(a) Population distribution (b) Sampling distribution of $\bar{x}$ for $n = 100$

Fig. 11.6 Population Distribution and Sampling Distribution

Let us take another example.

In respect of the central limit theorem, the following three aspects need to be noted:

1. When the sample size is large enough, the sampling distribution of $\bar{x}$ is normal.
2. The expected value of $\bar{x}$ is μ.
3. The standard deviation of $\bar{x}$ is $\dfrac{\sigma}{\sqrt{n}}$.

The last statement is very important as it shows that when the sample size increases, the variation between $\bar{x}$ and its mean μ decreases. This will be evident from Example 11.7 and Figs. 11.6 and 11.7. When the sample size is 30, $\sigma_{\bar{x}}$ is 82.16. In contrast, when the sample size is increased to 100, $\sigma_{\bar{x}}$ is much lower, being only 45.

Example 11.8 Suppose a random sample of $n = 25$ observations is selected from a population with mean $\mu = 15$ and $\sigma = 11$. What is the probability that the sample mean $\bar{x}$ will be **(a)** less than 14 **(b)** more than 14 and **(c)** ±1 of the population mean $\mu = 15$?

Solution

(a) Regardless of the shape of the population, the sampling distribution of $\bar{x}$ will possess a mean $\mu_{\bar{x}} = \mu = 15$ and a standard deviation

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{25} = 2$$

For a sample as large as $n = 25$, it is likely that the sampling distribution of $\bar{x}$ is approximately normally distributed on the basis of the central limit theorem. It is assumed that it is so in this case. As such, the probability P that $\bar{x}$ will be less than 14 is approximated by the shaded area as shown in Fig. 11.7.

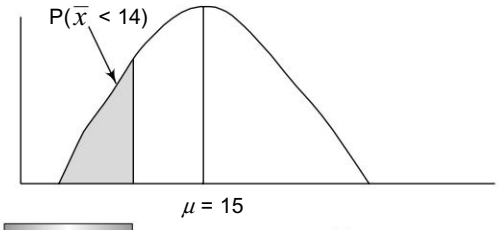


Fig. 11.7 Probability of $\bar{x}$ less than 14

Now, we calculate the value of Z which is the distance between $\bar{x} = 14$ and $\mu_{\bar{x}} = \mu = 15$ expressed in standard deviations of the sampling distribution.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{14 - 15}{2} = -0.5$$

From the Appendix Table 1, we find that the area corresponding to $Z = 0.5$ is 0.3085.

(b) The event that $\bar{x}$ is more than 14 is the complement event of $\bar{x}$ is less than 14. Symbolically, this can be written as

$$P(\bar{x} > 14) = 1 - P(\bar{x} < 14) = 1 - 0.3085 = 0.6915$$

This area is the blank area of Fig. 11.7. Thus, the two areas together add to 1.

(c) The probability that $\bar{x}$ lies within ±1 of $\mu = 15$ is the shaded area of Fig. 11.8. In Part (a) of this problem, we found that the area between $\bar{x} = 14$ and $\mu = 15$ was 0.3085. Since the area under the normal curve between $\bar{x} = 16$ and $\mu = 15$ is identical to the area between $\bar{x} = 14$ and $\mu = 15$, we can calculate the probability as follows:

$$P(14 < \bar{x} < 16) = 2(0.3085) = 0.6170$$

Figure 11.8 shows this value in the shaded area.

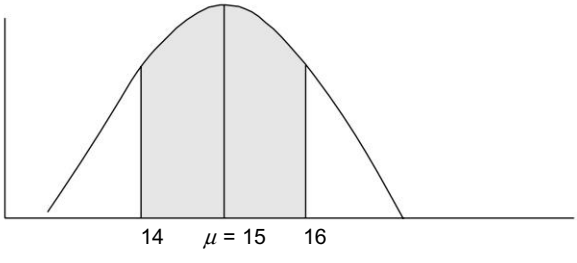


Fig. 11.8 Probability of $\bar{x}$ Being Within ±1 of Population Mean

At this stage, we may summarise the steps for calculating the probability that the statistic $\bar{x}$ falls in some interval, as given below:

1. Calculate the mean and the standard deviation of the sampling distribution of $\bar{x}$.
2. Draw a chart of the sampling distribution. Show the location of the mean μ and use the value of $\sigma_{\bar{x}}$ to determine the approximate location of the tails of the distribution.
3. Locate the interval on the chart drawn earlier in (2) above, and shade the area corresponding to the probability that we would like to calculate.
4. Find the *Z*–score(s) associated with the value(s) of interest in our problem. Use Appendix Table 1 to find the probability.
5. When we have obtained our answer, we should look the chart of the sampling distribution again to ascertain whether our calculated answer agrees with the shaded area.

11.12 RELATIONSHIP BETWEEN SAMPLE SIZE AND STANDARD ERROR

It may be recalled that we use the term $\sigma_{\bar{x}}$ (the standard error) as a measure of dispersion of the sample means around the population mean μ. If the standard error is small, then it shows that the dispersion of the sample means is small and the values taken by the sample mean tend to concentrate more closely around μ. Conversely, if the standard error increases, then it shows that the values taken by the sample mean tend to drift from the population mean μ. This phenomenon can be put in this form: As the standard error decreases, the precision increases, that is, the difference between the sample mean and the population mean narrows down.

The relationship between standard error and sample size can be expressed by the equation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This expression shows that as n increases, $\sigma_{\bar{x}}$ decreases and vice versa.

Let us show this by two numerical examples.

Example 11.9 Assume that the standard deviation σ is 300 and *n* is 25. Calculate the standard error of the sample mean.

Solution Applying these values in the above equation, we get

$$\sigma_{\bar{x}} = \frac{300}{\sqrt{25}} = 60$$

Thus, the standard error of the sample mean comes to 60.

Example 11.10 Let us now increase the sample size to find out how it affects the standard error. Suppose the sample size *n* is increased four-times from 25 to 100 and the standard deviation remains the same. Calculate the standard error and interpret the result.

Solution Applying the formula given above, the standard error would be $300/\sqrt{100} = 30$.

We find that as the sample size has increased, the standard error has declined. Further, when the sample size increases four-times, the standard error is reduced to half. It is worth noting from the above examples that there is an inverse relationship between *n* and $\sigma_{\bar{x}}$. At the same time, we find that the reduction in $\sigma_{\bar{x}}$ is relatively much less than the increase in *n*.

The Finite Population Correction Factor The equation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ depicting the relationship between the standard error and the sample size is applicable when the population is infinite, or in which a sample is selected from a finite population with replacement. Sampling with replacement implies that when an item is sampled, it is put back in the population before choosing another item. This means that the same item can possibly be chosen again.

As in many of the situations population tends to be finite, some adjustment is called for in the above formula. Thus, the adjusted equation now becomes

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

where N = size of the population and n, as earlier, is size of the sample. The additional term given on the right-hand side of the equation is called as the finite population correction factor (fpc). It is also known as the finite population multiplier.

It will be seen from the finite population correction factor that when sample size is extremely small in relation to population size, then it is very close to 1. In that case, it is unnecessary to use it as it would hardly have any impact on the standard error. *In the case of sampling with replacement*, there is an infinite population and as such, the finite population correction factor need not be used. In other cases too, if the sample is relatively too small vis-à-vis the population, the fpc factor need not be used as it will approach 1. The question is how to decide that N is relatively larger than n? Different people may take different values but the general practice is to use this formula (which excludes the correction factor) where n is less than 5 per cent of N.

Let us take a few examples to understand the application and interpretation of the fpc factor.

Example 11.11 Suppose we are interested in 15 electronic companies of the same size. All these companies are confronting a serious problem in the form of excessive turnover of their employees. It has been found that the standard deviation of the distribution of annual turnover is 60 employees. Take a sample of 3 electronic companies, without replacement, and determine the standard error of the mean.

Solution Our calculations would be as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{60}{\sqrt{3}} \times \sqrt{\frac{15-3}{15-1}}$$
$$= \frac{60}{1.7321} \times \sqrt{0.8571} = \frac{60}{1.7321} \times 0.9258$$
$$= 34.64 \times 0.9258 = 32.07$$

Thus, the standard error is 32.07. It will be seen that in this case the fpc factor of 0.9258 reduced the standard error from 34.64 to 32.07.

Example 11.12 Suppose in the foregoing example, if we have a large number of electronic companies, say, 800 and our sample continues to be of 3 companies, then what will be the fpc factor? Do you think that the fpc factor should be used in this case?

Solution

The fpc factor will be

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{800-3}{800-1}} = \sqrt{\frac{797}{799}} = 0.9987$$

This figure is extremely close to 1. In fact, if we round it off, it becomes 1. As such, it would hardly have any effect on the standard error. In such a case, there is no need to use the fpc factor.

It may be pointed out that fraction *n/N* is known as the sampling fraction because it is the proportion or fraction of the population that the sample contains. When the sampling fraction is less than 0.05 (i.e. 5 per cent of the population), we need not use the fpc factor.

Reverting to our earlier formula of standard error $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, we find that the magnitude of standard error depends on the absolute size of a sample, *n*. The degree of precision, which implies how close is the standard error from the population standard deviation, depends on the absolute sample size, *n* and not on the sampling fraction, *n/N*.

11.13 SAMPLING DISTRIBUTION OF SAMPLING PROPORTION

Our discussion so far was confined to the sampling distribution of the sample mean $\bar{x}$. We shall now discuss the distribution of the sample proportion $\bar{p}$. In Chapter 10, we discussed the binomial distribution where *p* is taken as the success while $q = 1 - p$ is taken as the failure. Suppose we take a random sample of *n* persons from a population and if *x* of these persons are smokers then the sample proportion, $\bar{p} = x/n$ is used to estimate the population proportion *p*. It may be noted that each value of *x* has a distinct value of $\bar{p} = x/n$. This would mean that the probabilities associated with $\bar{p}$ are equal to the probabilities associated with the corresponding values of *x*. When the sample size *n* is large, this sampling distribution of the sample proportion can be approximated by a normal distribution as shown in Fig. 11.9. The mean of the sampling distribution of $\bar{p}$ is $\mu_{\bar{p}} = p$ and its standard deviation is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

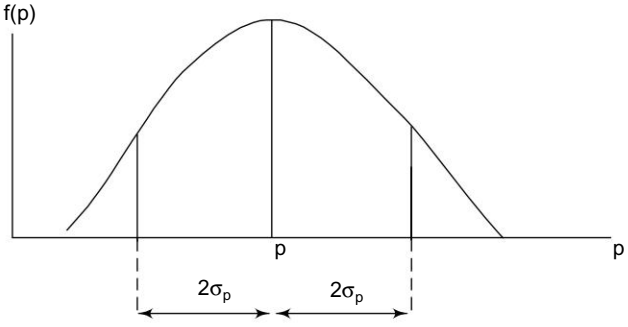


Fig. 11.9 The Sampling Distribution of the Sample Proportion *p*

Let us take another example.

Example 11.13 It has been found that 7 per cent of the tools manufactured by a factory are defective. What is the probability that in a shipment of 625 such tools **(a)** 8 per cent or more, and **(b)** 7 per cent or less will be defective?

Solution

(a) $\mu_{\bar{p}} = p = 0.07$ and $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

$$= \sqrt{\frac{0.07\,(0.93)}{625}} = \sqrt{\frac{0.0651}{625}} = \frac{0.255}{25} = 0.0102$$

Using the correction factor for discrete variables,

$1/(2N) = 1/1250 = 0.0008$, we have $(0.07 - 0.0008)$ in standard units

$= (0.07 - 0.0008 - 0.06) / 0.0102 = 0.902$ say 0.9.

For $Z = 0.9$ the Appendix Table 1 gives the area under normal curve as 0.1841. This is the answer when we have used the correction factor for discrete variables. If we had not used the correction factor, we would have obtained 0.1635, which is the area under normal curve to the right of Z as compared to the earlier answer of 0.1841.

(b) $(0.07 + 0.0008)$ in standard units

$= (0.07 + 0.0008 - 0.07) / 0.0102 = 0.078$, say 0.08

Required probability = (area under normal curve to left of $Z = 0.08$)

$= 0.5000 + 0.0319 = 0.5319$

These two results are shown as shaded portions in Fig. 11.10(a) and (b).

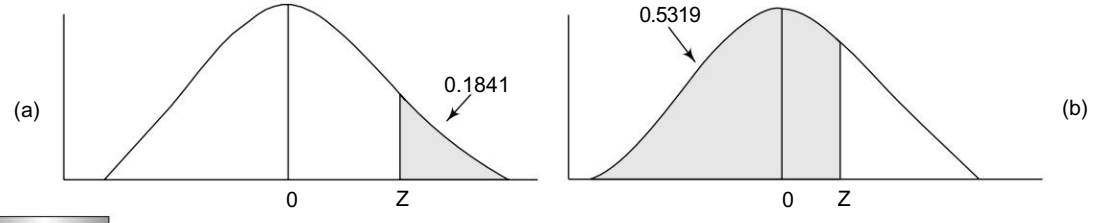


Fig. 11.10 **Sampling Distribution of Sample Proportions**

Example 11.14 We have been told that in a particular city, 20 per cent of the households subscribe to "*Outlook*", a fortnightly magazine. What is the probability of selecting a random sample of size $n = 400$ with a sample proportion $\bar{p} = 0.16$ or less?

Solution In solving such problems, we should first interpret them in terms of charts. Figure 11.11 shows that we are interested in the shaded area in its left tail.

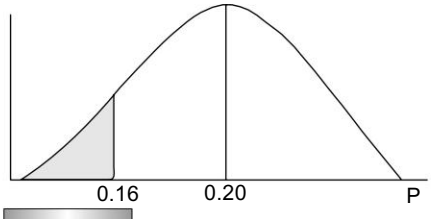


Fig. 11.11 **Sampling Distribution of Sample Proportion**

By using the normal approximation, we can easily find the required probability.

$$Z = (\bar{p} - p) / \sigma_{\bar{p}}$$

We have to first calculate $\sigma_{\bar{p}}$ as follows:

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.20)\,(0.80)}{400}} = \sqrt{\frac{0.16}{400}}$$

$$= \frac{0.4}{20} = 0.02$$

Applying this value in the above formula and using the correction factor for discrete variable,

$$Z = \left[0.16 + \left(\frac{1}{2} \times \frac{1}{400}\right) - 0.20\right] \Big/ 0.02$$

$$= -1.9375$$

It may be noted that we have assumed finite population correction factor = 1 and the continuity correction as $1/2 \times 1/N = 1/2 \times 1/400 = 0.00125$. As $Z = -1.9375$ or say 1.94, we find from the normal area table that $P[\bar{p} \leq 0.16 \mid p = 0.20] = 0.0262$.

This result shows that there are 2.62 chances in 100 (26.2 chances in 1000) of selecting samples of size $n = 400$ with a sample proportion ≤ 0.16 from a population where $p = 0.20$.

Let us take another example where we may have to use the right tail of the normal curve.

Example 11.15 A firm is engaged in the production of pocket calculators and has a large network of 1000 dealers all over the country. Of these 1000 dealers, 40 per cent informed the firm that they planned to increase their orders for pocket calculators. What is the probability of drawing a simple random sample of 200 dealers with a sample proportion of $\bar{p} = 0.45$ or more?

Solution We first interpret the information in terms of a normal curve figure. Figure 11.12 shows that we want the shaded area in the right tail in the normal curve.

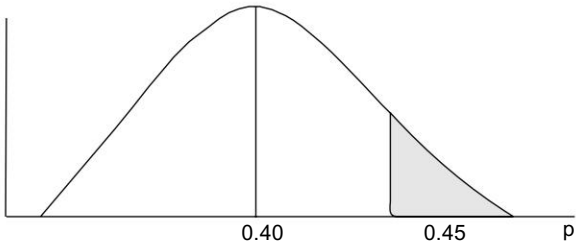


Fig. 11.12 Sampling Distribution of Sample Proportion

As in the earlier example, we have to use the formula

$$Z = (\bar{p} - p)/\sigma_{\bar{p}}$$

But this can be done when we know the value of $\sigma_{\bar{p}}$. Hence, we first calculate it.

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{(0.4)(0.6)}{200}}\sqrt{\frac{1000-200}{1000-1}}$$

$$= 0.0346 \times 0.8949 = 0.03096$$

Applying this value in the above formula and using the correction factor for discrete variable

$$Z = \left[0.45 - \left(\frac{1}{2} \times \frac{1}{200}\right) - 0.40\right] \Big/ 0.03096$$

$$= 1.5342 \text{ say } 1.53$$

Since $Z = 1.53$, we find from the normal area table the figure of 0.0630. We can now write $P[\bar{p} \geq 0.45 \mid p = 0.40] = 0.0630$.

This means that there are 6.3 chances in 100, of selecting samples of size $n = 200$ with sample proportion greater than or equal to 0.45 from a population where $p = 0.40$.

Additional Examples

Example 11.16 A finite population consists of 5 elements: A, B, C, D and E. Enumerate all the possible samples of size 3 that can be drawn from this population. If each of these samples is assigned the probability 1/10, find

(a) the probability that any specific element (say, the element C) will be contained in such a sample.
(b) the probability that any specific pair of elements (say, the elements D and E) will be contained in such a sample.

Solution There are 5 elements: A, B, C, D and E.
Samples of size 3 are to be specified.

First, we have to decide how many samples can be drawn. This means there are $\binom{5}{3}$ samples can be selected

$$\frac{5!}{(5-3)!\cdot 3!}$$

$$=\frac{5\times 4\times 3\times 2\times 1}{2\times 1\times 3\times 2\times 1}=10$$

In all, 10 samples of size 3 can be selected out of 5 elements. These samples are:
ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE and CDE

(a) Probability of C being contained in the sample ABC, ACD, ACE, BCD, BCE and CDE
C occurs 6 times.

Hence, probability is $\frac{6}{10}=\frac{3}{5}$

(b) Probability of a pair of elements, i.e., D and E is: ADE, BDE and CDE. D and E appear 3 times out of 11.

Hence, the probability is $\frac{3}{10}$.

Example 11.17 A population has a normal distribution with an unknown mean and a standard deviation $\sigma = 5$.

Find the probability that the sample mean will be within ±1 unit of the population mean in each of the following cases:

(a) $n = 25$ **(b)** $n = 100$ **(c)** $n = 225$

Solution

(a) $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = \frac{5}{5} = 1$$

$$Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{1}{1} = 1$$

The corresponding area for Z = 1 is 0.1587. We have to multiply it by 2 as $\bar{x}$ is to be ±1 of the population mean and then the product is to be subtracted from 1.

$$P = 1 - (0.1587 \times 2) = 1 - 0.3174 = 0.6826$$

(b) $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = \frac{5}{10} = 0.5$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{1}{0.5} = 2$$

Here, the area corresponding to Z = 2 is 0.0228. We have to carry out the same process as in (a) above

$$P = 1 - (0.0228 \times 2) = 1 - 0.0456 = 0.9544$$

(c) $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{225}} = \frac{5}{15} = \frac{1}{3} = 0.33$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{1}{0.33} = 3.03$$

Here, the area corresponding to Z = 3.03 is 0.0012.
Hence, P = 1 – (0.0012 × 2) = 1 – 0.0024 = 0.9976.

Example 11.18 A random sample of $n = 25$ was selected from a normally distributed population. The population mean is 106 and standard deviation is 12.

(a) Find the standard error of the sampling distribution.
(b) What is the probability that the sample mean $\bar{x}$ is greater than 100?
(c) What is the probability that the sample mean $\bar{x}$ will be within ±4 from population mean?

Solution

(a) $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{25}} = \frac{12}{5} = 2.4$$

(b) $$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{100 - 106}{2.4} = \frac{-6}{2.4} = -2.5$$

From the Appendix Table 1, we find that the area corresponding to $Z = -2.5$ is 0.0062. The event that $\bar{x}$ is greater than 100 can be written as

$$P(\bar{x} > 100) = 1 - P(\bar{x} < 100)$$
$$= 1 - 0.0062 = 0.9938$$

(c) $\bar{x} \pm 4$ from the population mean implies that $\bar{x}$ should lie between (106 – 4) and (106 + 4), i.e., 102 and 110.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{102 - 106}{2.4} = \frac{-4}{2.4} = -1.67$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{110 - 106}{2.4} = \frac{4}{2.4} = 1.67$$

From the Appendix Table 1, we find that the area corresponding to $Z = 1.67$ is 0.0475. It may be noted that the former figure relates to the left-tail of the normal curve while the latter figure to the right-tail. As such each figure is to be subtracted from 0.5, which is the total area of one-half of the curve. Thus, we get 0.5 – 0.0475 = 0.4525. This is to be multiplied by 2 as the Z value in both the cases is the same. Hence, the required probability is 0.452 × 2 = 0.905.

Example 11.19 The standard deviation of a certain population is about 5 units. How large a simple random sample should be if the standard error of the sample mean is to be 1? How large it should be for the standard error of 0.1?

Solution The relationship between the standard deviation of the population, the standard error of the sample mean and the sample size is given by the following formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Substituting the given values in the formula,

$$1 = \frac{5}{\sqrt{n}}$$

$$= 1\sqrt{n} = 5$$

or $n = 25$

$$0.1 = \frac{5}{\sqrt{n}}$$

or $0.1\sqrt{n} = 5$

or $\sqrt{n} = 5/0.1$

or $\sqrt{n} = 50$

or $n = (50)^2 = 2500$

Example 11.20 You have been asked to select a simple random sample from a population of 20,000 sales invoices to estimate the average amount per invoice.

(i) Suppose $\sigma = 500$, determine the sample size required if the allowable error is Rs 100 and the confidence coefficient at 95 per cent.

(ii) The standard deviation of the population is found to be 1000, that is, σ = Rs 1,000 and the researcher accepts the allowable error of Rs 200, determine the sample size and comment on comparison with **(i)**.

Solution

(i) The formula for determining sample size n is

$$n = \frac{Z^2\sigma^2}{E^2}$$

where E = the maximum error allowed and Z = the degree of confidence required.
For 95 per cent degree of confidence $Z = 1.96$

$$n = \frac{(1.96)^2(500)^2}{(100)^2}$$

$$= \frac{3.8416 \times \overset{25}{\cancel{250000}}}{\cancel{10000}}$$

$$= 96.04 \text{ or } 96 \text{ approx.}$$

(ii) $$n = \frac{Z^2\sigma^2}{E^2}$$

$$= \frac{(1.96)^2(1000)^2}{(200)^2}$$

$$= \frac{3.8416 \times \overset{25}{\cancel{1000000}}}{\cancel{40000}}$$

$$= 96.04 \text{ or } 96 \text{ approx.}$$

It will be seen that in both cases the sample size remains the same. This is because the doubling of the standard deviation is offset by the corresponding doubling of the allowable error. The other term Z remains unchanged.

Example 11.21 A random sample of size 9 is obtained from a normal population with $\mu = 25$. If the sample variance is 100, find the probability that the sample mean exceeds 31.2.

Solution We have to first calculate standard error of mean. As the sample variance is 100, $\sigma_{\bar{x}} = 10$.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{31.2 - 25}{10} = \frac{6.2}{10} = 0.62$$

Referring to Appendix Table 1, corresponding to $Z = 0.62$, the table gives an area of 0.2676. Subtracting this value from 0.5 comprising the right half of the table,

we get $0.5 - 0.2676 = 0.2324$

Hence, the probability that the sample mean exceeds 31.2 is 0.2324.

Example 11.22 A random sample of size 25 is taken from a normal population with $\mu = 49.5$. If sample variance is 1.69, find the probability that the mean of this sample falls between 48.75 and 50.10.

Solution As sample variance is 1.69, standard error of the mean $\sigma_{\bar{x}} = \sqrt{1.69} = 1.3$.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$= \frac{48.75 - 49.50}{1.3}$$

$$= \frac{-0.75}{1.3} = -0.576$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$= \frac{50.10 - 49.50}{1.3}$$

$$= \frac{0.6}{1.3} = 0.4615$$

Using Appendix Table 1, corresponding to $Z = -0.576$, (indicating the left-half of the normal area table) the area covered is 0.2810. Subtracting this value from 0.5,

we get $0.5 - 0.2810 = 0.219$.

Corresponding to $Z = 0.4615$, the area covered is 0.3228. Subtracting this value from 0.5,

we get $0.5 - 0.3228 = 0.1772$.

Adding the two values, we get the required probability as

$$0.2190 + 0.1772 = 0.3962$$

Example 11.23 Certain tubes, manufactured by a company, have mean lifetime of 800 hours and standard deviation of 60 hours. Find the probability that a random sample of 16 tubes taken from the group will have mean lifetime of (a) between 790 and 810 hours, (b) less than 785 hours, (c) more than 820 hours, and (d) between 770 and 830 hours.

Solution

$\because$ $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Therefore, $$\sigma_{\bar{x}} = \frac{60}{\sqrt{16}} = \frac{60}{4} = 15$$

(a) $$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$= \frac{790 - 800}{15}$$

$$= -\frac{10}{15} = -0.67$$

Further, $$\frac{810 - 800}{15} = \frac{10}{15} = 0.67$$

$Z = 0.67$ gives an area of 0.2514

Subtracting this from 0.5, we get

$$0.5 - 0.2514 = 0.2486$$

The values of $Z = -0.67$ and 0.67 for the left-half of the normal curve and the right-half of the normal curve.

Hence, by adding, we get $0.2486 + 0.2486 = 0.4972$

(b) $$Z = \frac{\bar{x} - \mu}{\sigma}$$

$$= \frac{785 - 800}{15} = \frac{-15}{15} = -1$$

$Z = 1$ gives an area of 0.1587

(c)

$$Z = \frac{\bar{x} - \mu}{\sigma}$$

$$= \frac{820 - 800}{15} = \frac{20}{15} = \frac{4}{3} = 1.33$$

$Z = 1.33$ gives an area of 0.0918

(d)

$$Z = \frac{\bar{x} - \mu}{\sigma}$$

$$= \frac{770 - 800}{15} = \frac{-30}{15} = -2$$

Again, $\frac{830 - 800}{15} = \frac{30}{15} = 2$

$Z = 2$ gives an area of 0.0228

$0.5 - 0.0228$

$= 0.4772$

As this applies to each half of the normal curve, the required probability would be

$0.4772 + 0.4772$

$= 0.9544$

GLOSSARY

Area sampling	A form of cluster sampling in which areas such as census tracts and blocks form the primary sampling units. The population is divided into mutually exclusive areas using maps. A random sample of the area is then selected.
Census	A measurement of each element in a group or population of interest.
Central limit theorem	A theory that states that as a sample size increases, the distribution of sample means tends to take the form of a normal distribution.
Cluster sampling	A sample design in which a cluster of elements is the primary sampling unit instead of individual elements in the population.
Convenience sample	A sample selected by the researcher on the basis of his convenience.
Finite population correction factor	A correction factor used while determining sample size from a finite population. The usual practice is to apply it when sample size is more than 5 per cent of the population. It is also known as *Finite Population Multiplier*.
Finite population	A population having a stated or limited size.
Infinite population	A population that is exceptionally large in size and as such it is impossible to cover all the elements comprising it.

Judgment sample	A non-probability sample based on the judgment of the researcher who thinks that the sample respondents thus selected would contribute to answering the question.
Multi-stage sampling	A sample design in which a sample is drawn in two or more stages sequentially. The sampling unit in each stage tends to be different.
Non-sampling error	An error that occurs in the collection, recording, tabulation and computation of data.
Parameter	A quantity that remains constant in each case considered but varies in different cases.
Precision	The desired size of the confidence interval when a population parameter is to be estimated. The concept is also useful in determining sample size.
Quota sample	A non-probability sample that contains a pre-specified quota of certain characteristics of a population.
Random sample	A sample that assigns some chance to each element of the population to be selected in the sample. It is also known as probability sample.
Representative sample	A sample that represents the characteristics of the population as closely as possible.
Sample	A subset or some part of a population.
Sampling distribution of a statistic	For a given population, a probability distribution of all the possible values that a statistic may take on for a given sample size.
Sampling distribution of the mean	The probability distribution of all the values of the mean calculated from all possible samples of the same size selected from a population.
Sampling error	The difference between the population parameter and the observed probability sample statistic.
Sampling fraction	The proportion of the number of elements included in a sample to the total number of elements contained in a population.
Sampling with replacement	A sampling procedure in which sample items are returned to the population; as a result, there is a possibility of their being chosen again in the sample.
Sampling without replacement	A sampling procedure in which sample items are not returned to the population, as a result none of these can be selected in the sample again.
Simple random sample	A probability sampling procedure where each element of the population has an equal chance of being selected.
Standard error of the mean	The standard deviation of the sampling distribution of the mean. It is calculated by dividing the population standard deviation by the square root of the sample size.
Standard error	The standard deviation of the sampling distribution of a statistic.

Statistic	A measure or characteristic of a sample.
Statistical inference	The process of deriving inference about the population from the information contained in the sample.
Strata	Groups within a population formed in such a way that each group is relatively homogeneous but wider variability exists among the separate groups.
Stratified sampling	A probability sampling method in which sub-samples are drawn from two or more strata comprising the population. The strata are more or less homogeneous.
Systematic sampling	A sampling method in which a sample is drawn in such a way that it is systematically spread over all the elements of population.

LIST OF FORMULAE

1. Standard error of the sample mean when the population is infinite

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ = standard error of the sample mean
σ = standard deviation of the population
n = number of elements or units in the sample

2. Standard error of the sample mean when the population is finite

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

where N = size of the population
n = number of elements or units in the sample

3. Finite population correction factor

$$\sqrt{\frac{N-n}{N-1}}$$

As used in formula 2 above, this multiplier is used when the population is small in relation to the sample size. When the sampling fraction $\left(\frac{n}{N}\right)$ is less than 0.05, this multiplier need not be used.

4. The Z value for $\bar{x}$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

where $\bar{x}$ = sample mean
$\sigma_{\bar{x}}$ = standard error of the sample mean
Z = number of standard errors from $\bar{x}$ to the population mean

Once the Z value has been obtained, the standard normal probability distribution table (Appendix Table 1) can be used. The table is organised in terms of standard units or Z values.

5. Sample proportion

$$\bar{p} = x/n$$

where $\bar{p}$ = sample proportion
x = number of elements in the sample that possess a specific characteristic
n = number of elements or units in the sample

6. Mean of the sample proportion $\bar{p}$

$$\mu_{\bar{p}} = p$$

7. Standard error of the sample proportion

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$(q = 1 - p)$

8. The Z value for sample proportion

$$Z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

where $\bar{p}$ = sample proportion
p = population proportion
$\sigma_{\bar{p}}$ = standard error of the sample proportion

QUESTIONS

11.1 Given below are twelve statements. Indicate in each case whether the statement is true or false:

(a) A parameter is a characteristic of a sample.
(b) In a random sample every element in the population has an equal chance of being selected.
(c) A cluster sample is a non-random sample.
(d) The standard error is different from the standard deviation of the distribution of sample means.
(e) A stratified random sampling is one where the population is divided into mutually exclusive and mutually exhaustive strata.
(f) As the sample size n increases, the standard error $\sigma_{\bar{x}}$ not necessarily decreases.
(g) Judgment sampling is not a representative sample.
(h) The standard error of the mean $\sigma_{\bar{x}}$ decreases in direct proportion to sample size n.
(i) The proportion of sample size to the population size is known as the sampling fraction.
(j) A theoretical sampling distribution implies that all the samples of a given size are considered.
(k) The precision of a sample depends on the proportion of the population sampled.
(l) If n is relatively very small as compared to N, then the final population correction factor need not be used.

Multiple Choice Questions (11.2 to 11.16)

11.2 Which of the following samples is not a probability sample design?
(a) Stratified sample (b) Multi-stage sample
(c) Cluster sample (d) Quota sample

11.3 Why is a census survey not popular?
(a) It is very costly.
(b) It takes more time.
(c) It requires a large number of investigators.
(d) All the above.

11.4 How many different sample of size 3 can be taken from the population comprising 5 elements?
(a) 7 (b) 12 (c) 5 (d) 10

11.5 When the sample size increases, which of the following is correct?
(a) The standard error remains unchanged.
(b) The standard error increases.
(c) The standard error declines.
(d) None of the above.

11.6 The application of Finite population multiplier is not required when the sampling fraction is
(a) Greater than 0.5 (b) Less than 0.5
(c) Greater than 0.05 (d) None of these

11.7 If the standard error of proportion is reduced by 50 per cent, the sample size becomes
(a) double (b) increases 6 times
(c) increases 4 times (d) none of the above

11.8 In case the population has a normal distribution, then the sampling distribution of the mean
(a) has a mean equal to the population mean (b) has a normal distribution
(c) both (a) and (b) (d) none of the above

11.9 In which of the following sample designs, maps rather than lists or registers are used as the sampling frame?
(a) simple random sample (b) cluster sample
(c) area sample (d) none of the above

11.10 As $\sigma_{\bar{x}}$ decreases in a sample distribution,
(a) the shape of the sample distribution remains the same
(b) becomes more skewed
(c) becomes absolutely normal
(d) tends to be normal

11.11 Suppose that for a certain population, $\sigma_{\bar{x}}$ is calculated as 30 when a sample of size 225 is taken. What would be the value of σ for this infinite population?
(a) 500 (b) 400 (c) 450 (d) 600

11.12 Suppose that a population with $N = 200$ has $\mu = 30$. What is the mean of the sampling distribution of the mean for samples of size 40?
(a) 40
(b) Not possible to determine as this information is inadequate
(c) 25
(d) 30

11.13 The use of the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ would be correct if
(a) sampling is from a finite population
(b) sampling is from finite population with replacement
(c) sampling is from infinite population
(d) (b) and (c)
(e) (a) and (b)

11.14 The precision of proportion as an estimator of the population proportion is the minimum when it is
(a) 0.75 (b) 0.25 (c) 0.50 (d) none of these

11.15 When there is a significant difference between the statistic and parametric values, it means that
(a) sample statistic is representative of the population parameter
(b) the difference is real
(c) statistic value is used to approximate parameter
(d) none of the above

11.16 For a sampling distribution of proportion, the most probable limits for population proportion are
(a) $p \pm 3\sigma_p$ (b) $\bar{x} \pm 3\sigma_{\bar{x}}$ (c) $\bar{p} \pm 3\sigma_{\bar{p}}$ (d) none of these

11.17 What are the advantages and limitations of sampling?

11.18 What is meant by 'representativeness' in a sample? Explain in what sense a simple random sample is representative of the population.

11.19 Distinguish between stratified sampling and systematic sampling.

11.20 Write short notes on the following, giving suitable examples:
(a) Probability sampling and non-probability sampling
(b) A cluster and area sampling
(c) Judgment and convenience sampling

11.21 Explain the terms 'sampling distribution' and 'standard error' of a statistic.

11.22 What is standard error? Explain its utility in sampling.

11.23 What is the finite population correction factor? When is it used? What is the value of this factor when **(a)** $n = 5$ and $N = 200$; and **(b)** $n = 100$ and $N = 5000$?

11.24 How many different samples of size 4 can be selected
(a) from a finite population of size 8?
(b) from a finite population of size 12?
(c) from a finite population of size 40?

11.25 How many different samples of size 2 can be selected
(a) from a finite population of size 6?
(b) from a finite population of size 10?
(c) from a finite population of size 25?

11.26 What is the probability of each possible sample if a random sample of size 5 is to be taken
(a) from a finite population of size 15?
(b) from a finite population of size 25?

11.27 A finite population consists of the following elements: 4, 5, 6, 7, 8 and 9.
(a) List all the samples of size 2 that can be taken from this finite population and calculate their respective means.

(b) Based on the data in **(a)**, construct a sampling distribution of the mean.
(c) Calculate the mean and the standard deviation of the probability distribution obtained in Part **(b)**.

11.28 A population comprises 4 numbers, 3, 5, 7 and 9.
(a) List all possible samples of size 2 that can be drawn from the population without replacement.
(b) Show that the mean of the sampling distribution of sample means is equal to the population mean.
(c) Calculate the standard deviation of the sampling distribution of sample means and show that it is less than the population standard deviation.

11.29 A machine produces 500 similar components with a mean weight of 4.03 *N* and a standard deviation of 0.20 *N*. A random sample of 60 components is selected from this group. What will be the probability of sample mean weight of (a) between 4.0 and 4.1 *N*; (b) more than 4.05 *N*? Assume that sampling is without replacement and the weights of the components are normally distributed.

11.30 The mean length of life of a certain cutting tool is 41.5 hours with 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?

11.31 Two electric light tubes of manufacturer A has a mean lifetime of 1,400 hours with a standard deviation of 100 hours corresponding figures for brand B tubes are 1200 hours and 100 hours. If a random sample of 125 tubes of each brand is tested, what is the probability that the brand A tubes will have a mean lifetime that is at least **(i)** 160 hours more than the brand B tubes; and **(ii)** 250 hours more than the brand B tubes?

11.32 A manufacturer of automobile batteries claims that the average duration of useful life for its grade A battery is 60 months. However, the guarantee on this brand is only 36 months. Assume that the standard deviation of the life-duration is known to be 10 months and that the frequency distribution of the life-duration is known to be mound–shaped.
(a) Approximately what percentage of the manufacturer's grade A batteries will last more than 50 months, assuming the manufacturer's claim about the mean life-duration is true?
(b) Approximately what percentage of the manufacturer's batteries will last less than 40 months, assuming that the manufacturer's claim about the mean life-duration is true?
(c) Suppose that your grade A battery lasts 37 months. What would you infer about the manufacturer's claim that the mean life-duration is 60 months?

11.33 Stress on the job is a major concern of a large number of people who go into managerial positions. Eighty per cent of all managers of companies suffer from stress. Let p be the proportion in a sample of 100 managers of companies who suffer from stress.
(a) What is the probability that the sample proportion is lower than the population proportion by 0.1 or more?
(b) What is the probability that the sample proportion is within 0.08 of the population proportion?
(c) What is the probability that the sample proportion is greater than the population proportion by 0.11 or more?
(d) What is the probability that the sample proportion is not within 0.08 of the population proportion?

11.34 Seventy per cent of adults favour some kind of government control on the prices of medicines. Assume that this percentage is true for the current population of all adults in a certain territory. Let p be the proportion of adults in a random sample of 400, who favour government control on the prices of medicines. Calculate the mean and the standard deviation of p and describe its sampling distribution.

11.35 Two firms A and B manufacture similar components with a mean breaking strength of 3,000 *N* and 2500 *N* and standard deviations of 200 *N* and 100 *N*, respectively. If random samples of 100 components of Firm A and 50 components of Firm B are tested, what is the probability that the components from Firm A will have a mean breaking strength which is at least (a) 450 *N*; (b) 575 *N* more than the components of Firm B?

11.36 Determine the sample size using the formula for the standard error of the mean, given that

(a) Level of precision = +5

(b) Confidence level = 95%

(c) The standard deviation of the population = 55.

11.37 Given that $n_1 = 400$, $x_1 = 250$, $s_1 = 40$ for one sample and $n_2 = 400$, $x_2 = 220$, $s_2 = 55$ for another sample, find the standard error of $x_1 - x_2$.

11.38 A random sample of size 16 has 53 as mean. The sum of squares of the deviations taken from the mean is 150. Can this sample be regarded as taken from the population having 56 as mean?

11.39 In measuring reaction time, a psychologist estimates that the standard duration is 0.05 second. 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 second?

11.40 The quantity of soft drink in a 12 oz can, manufactured by a company, is known to be well-approximated by a normal distribution, with a mean of 12 oz and a standard deviation of 0.025 oz. Find the probability of getting 4 cans all of which are less than 12 oz, if a random sample of 4 cans is selected.

11.41 A simple random sample of 50 ball-bearings, taken from a large population, has mean weight of 1.5 kg/bearing and standard deviation of 0.1 kg/bearing.

(i) Estimate the standard error of mean.

(ii) If the sample is taken from a production run of 150, estimate the standard error of mean.

(iii) If 10% of the bearings are known to be defective, calculate the standard error of proportions in a sample of 50.

(iv) If 10% of the bearings are known to be defective, and the production run of 150 is considered, calculate the standard error of proportions in a sample of 50.

11.42 At an inspection station, the probability of acceptance of the items is 0.9. If a sample of 10 items is taken, what is the probability that the number of accepted items is between 6 and 9?

11.43 The standard deviation of weights of a very large population of males is 10 kg. Samples, of 200 males each, are drawn from this population, and the standard deviations of the weights in each sample are computed. Find (a) the mean and (b) the standard deviation of the sampling distribution of standard deviations.

The McGraw-Hill Companies

CHAPTER 12 ESTIMATION

Learning Objectives

By the end of your work on this chapter, you should be able to

- recognise the need for making estimates, and differentiate between point and interval estimates
- know the criteria for a good estimator
- have a good understanding of the procedure involved in constructing confidence intervals for sample means and sample proportions
- determine the sample size in estimation.

Chapter Prerequisites

Before starting work on this chapter, make sure that you have a good understanding of all the material on probability and sampling covered in the three preceding chapters.

12.1 INTRODUCTION

Having discussed probability theory, the normal distribution, and the concept of sampling distribution of a statistic in the preceding chapters, we now turn to the main topics of Statistics, namely, estimation and testing hypotheses. In this chapter, estimation is discussed while hypothesis testing will be taken up in Chapter 13.

Inferential Statistics

We are now entering the area of inferential Statistics. It may be recalled that in Chapter 1 we defined inferential Statistics, which helps us in making decisions about some characteristics of a population based on the sample information. As such, inferential Statistics is always based on the sample information. Estimation is the first topic in inferential Statistics.

In business, there arise several situations when managers have to make quick estimates. Since their estimates have an impact on the success or failure of their enterprises, they have to take sufficient care to ensure that their estimates are not far away from the final outcome. The point to note is that such estimates are made without complete information and with a great deal of uncertainty about the eventual outcome.

In all such situations, it is the theory of probability that forms the basis for statistical inference. The term 'statistical inference' means *making a probability judgment concerning a population on the basis of one or more samples*. Based on probability theory, statistical inferences are made as a basis for making decisions. For example, an investor is interested to know whether he should subscribe for an investment consultancy service or not. On the basis of a sample, he has to examine whether the selection of his investment on the advice of the investment consultancy service has been more profitable than the selection based randomly, he may go in for this service. Likewise, a quality control engineer while examining the control chart finds that the production process has gone out of control. He may then look for the possible sources that have led to this situation. He may then take corrective measures to restore the production process under control.

In this chapter, we shall study the methods, which will enable us to estimate a population mean and the population proportion with a reasonable degree of accuracy. It may be noted that it is almost impossible to arrive at the exact population mean or population proportion in the wake of incomplete information. However, we will make an estimate and ascertain the extent of error that may be involved in such an estimate. We will also enforce some controls to minimise the error in our estimate. In other situations where we have more information, it will enable us to apply relevant statistical concepts. As a result, our estimates may turn out to be more accurate in such situations.

12.2 TYPES OF ESTIMATES

Let us first know the concept of 'estimate' as used in Statistics. According to some dictionaries, an estimate is a valuation based on opinion or roughly made from imperfect or incomplete data. This definition may apply, for example, when an individual who has an opinion about the competence of one of his colleagues. But, in Statistics the term estimate is not used in this sense. In Statistics too, the estimates are made when the information available is incomplete or imperfect. However, such estimates are made only when they are based on sound judgment or experience and when the samples are scientifically selected.

There are two types of estimates that we can make about a population: a *point estimate* and an *interval estimate*.

Point Estimate

A point estimate is a single number, which is used to estimate an unknown population parameter. Although a point estimate may be the most common way of expressing an estimate, it suffers from a major limitation since it fails to indicate how close it is to the quantity it is supposed to estimate. In other words, a point estimate does not give any idea about the reliability or precision of the method of estimation used. For instance, if someone claims that 40 per cent of all children in a certain town do not go to the school and are devoid of education, it would not be very helpful if this claim is based on a small number of households, say, 20. However, as the number of households interviewed for this purpose increases from 20 to 100, 500 or even 5,000, the claim that 40 per cent of children have no school education would become more and more meaningful and reliable. This makes it clear that a point estimate should always be accompanied by some relevant information so that it is possible to judge how far it is reliable.

Interval Estimate

The second type of estimate is known as the interval estimate. It is a range of values used to estimate an unknown population parameter. In case of an interval estimate, the error is indicated in two ways: first by the extent of its range; and second, by the probability of the true population parameter lying within that range. Taking our previous example of 40 per cent children not having a school education, the statistician may say that actual percentage of such children in that town may lie between 35 per cent and 45 per cent. Thus, he will have a better idea of the reliability of such an estimate as compared to the point estimate of 40 per cent.

Estimator and Estimate

When we make an estimate of a population parameter, we use a sample statistic. This sample statistic is an estimator.

For example, the sample mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$\bar{x}$ is a point estimator of the population mean μ. Many different Statistics can be used to estimate the same parameter. For example, we may use the sample mean or the sample median or even the range to estimate the population mean. The question here is: how can we evaluate the properties of these estimates, compare them with one another, and finally, decide which is the 'best'? The answer to this question is possible only when we have certain criteria that a good estimator must satisfy. These criteria are briefly discussed below.

12.3 CRITERIA OF A GOOD ESTIMATOR

There are four criteria by which we can evaluate the quality of a statistic as an estimator. These are: *unbiasedness*, *consistency*, *efficiency,* and *sufficiency*.

Unbiasedness

This is a very important property that an estimator should possess. If we take all possible samples of the same size from a population and calculate their means, the mean $\mu_{\bar{x}}$ of all these means will be equal to the mean μ of the population. This means that the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, the sample statistic is said to be an unbiased estimator.

Suppose we take the smallest sample observation as an estimator of the population mean μ, it can be easily shown that this estimator is biased. Since the smallest observation must be less than the mean, its expected value must be less than μ. Symbolically, $E(\text{Xs}) < \mu$, where Xs stands for the smallest item and E stands for the expected value. Thus, this estimator is biased downwards. The extent of bias is the difference between the expected value of the estimator and the value of the parameter. In this case, bias is equal to $E(\text{Xs}) - \mu$. In contrast, the bias for the sample mean $\bar{x}$ is zero.

Consistency

Another important characteristic that an estimator should possess is consistency. Let us take the case of the standard deviation of the sampling distribution of $\bar{x}$. In Chapter 11, we have used several times the following formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The formula states that the standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size increases and vice versa. When the sample size n increases, the population standard deviation σ is to be divided by a higher denominator. This results in the reduced value of sample standard deviation $\sigma_{\bar{x}}$. Let us take an example.

Example 12.1 A company has 4,000 employees whose average monthly wage comes to Rs 4,800 with a standard deviation of Rs 1,200. Let $\bar{x}$ be the mean monthly wage for a random sample of certain employees selected from this company. Find the mean and standard deviation of $\bar{x}$ for a sample size of **(a)** 40, **(b)** 100, and **(c)** 180.

Solution From the given information, for the population of all employees, $N = 4{,}000$ μ = Rs 4,800 σ = Rs 1,200

(a) The mean $\mu_{\bar{x}}$ of the sampling distribution of the $\bar{x}$ is $\mu_{\bar{x}} = \mu$ = Rs 4,800.
As $n = 40$ and $N = 4{,}000$, which gives $n/N = 0.01$. As this value is less than 0.05, the standard deviation of $\bar{x}$ is obtained by using the formula. Substituting the values,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or, } \sigma_{\bar{x}} = \frac{1{,}200}{\sqrt{40}} = \frac{1{,}200}{6.32} = \text{Rs } 189.87$$

(b) In this case, $n = 100$ and $n/N = 100/4{,}000 = 0.025$, which is also less than 0.05. The mean and the standard deviation $\bar{x}$ are

$$\mu_{\bar{x}} = \mu = \text{Rs } 4{,}800.$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or } \quad \sigma_{\bar{x}} = \frac{1{,}200}{\sqrt{100}} = \frac{1{,}200}{10} = \text{Rs } 120$$

(c) In this case, $n = 180$ and $n/N = 180/4{,}000 = 0.045$, which again is less than 0.05. The mean and the standard deviation of $\bar{x}$ are

$$\mu_{\bar{x}} = \mu = \text{Rs } 4{,}800.$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or } \quad \sigma_{\bar{x}} = \frac{1{,}200}{\sqrt{180}} = \frac{1{,}200}{13.42} = \text{Rs } 89.42$$

From the above three sets of calculation, it becomes clear that the mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population regardless of the sample size. But, in case of the standard deviation, we find the change. In the given example, we find that standard deviation of $\bar{x}$ decreased from Rs 189.87 to Rs 120 and then to Rs 89.42 as the sample size increased from 40 to 100 and then to 180.

Efficiency

Another desirable property of a good estimator is that it should be efficient. Efficiency is measured in terms of size of the standard error of the statistic. Since an estimator is a random variable, it is necessarily characterised by a certain amount of variability. This means that some estimates may be more variable than others. Just as bias is related to the expected value of the estimator, so efficiency can be defined in terms of the variance. In large samples, for example, the variance of the sample mean is $V(\bar{x}) = \sigma^2/n$. As the sample size n increases, the variance of the sample mean ($V_{\bar{x}}$) becomes smaller, so the estimator becomes more efficient. This criterion, when applied to large samples, gives better

estimates as compared to the small ones. In Fig. 12.1, the probability distributions of the two estimators are shown.

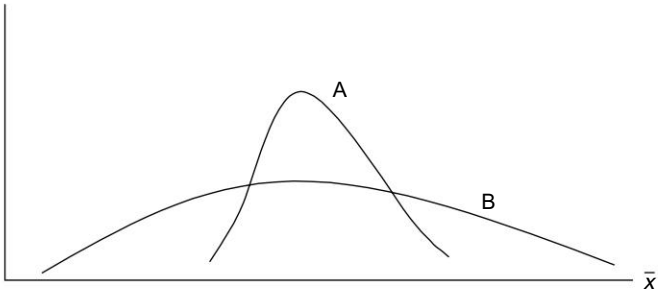


Fig. 12.1 The Sampling Distribution of Two Estimators

Curve A shows the distribution of sample means. It is a more precise estimator as compared to curve B.

It may be noted that although estimator A is biased, but it will yield an estimate that will be close to the true value, though it is likely to be wrong. Estimate B, though unbiased, can give estimates that are far away from the true value. As such, A would be the preferred estimator.

The efficiency of one estimator in relation to another estimator can be judged by comparing their sampling variances. Thus, efficiency relates to the size of the standard error. Given the same sample size, the statistic that has a smaller standard error is preferable as it is efficient in relation to another statistic that has a larger standard error. The sampling distribution of the mean and the median have the same mean, that is, the population mean. However, the variance of the sampling distribution of the means is smaller than the variance of the sampling distribution of the medians. As such, the sample mean is an efficient estimator of the population mean, while the sample median is an inefficient estimator.

Sufficiency

The fourth property of a good estimator is that it should be sufficient. A sufficient statistic such as $\bar{x}$ is an estimator, that utilises all the information a sample contains about the parameter to be estimated. $\bar{x}$, for example, is a sufficient estimator of the population mean μ. It implies that no other estimator of μ, such as the sample median, can provide any additional information about the parameter μ. Likewise, we can say that the sample proportion $\bar{p}$ is a sufficient estimator for the population proportion p.

Having looked into properties of a good estimator briefly, a pertinent question arises: how can we find estimators with these desirable properties? This brings us to the method of maximum likelihood.

12.4 METHOD OF MAXIMUM LIKELIHOOD (ML)

The maximum likelihood method provides estimators with the desirable properties such as efficiency, consistency and sufficiency, which we have just discussed. It usually does not give an unbiased estimate. Let us take an example to explain this method.

Example 12.2

Solution Suppose we want to estimate the average grade μ of a large number of students. A random sample of size $n = 64$ is taken and the sample mean $\bar{x}$ is found to be 90 marks. Now, the assumption on

which we have to base our reasoning is that the random sample of $n = 64$ is representative of the population. It may be recalled that this assumption is reasonable as was mentioned in Chapter 11 in which sampling distributions were discussed. We saw how samples that were similar to the population had greater probability of being selected.

Let us now reverse this reasoning as follows: we have before us a random sample size $n = 64$ and $\bar{x} = 90$ marks. From which population did it most probably come—a population with $\mu = 85$, 90 or 95? According to our earlier approach, we would think that it most probably came from a population with $\mu = 90$. Thus, it can be concluded that the population mean μ, based on our sample, is most likely to be $\mu = 90$ marks.

A point worth noting is that the population mean μ is either 90 or not; it has only one value. Hence, we have used the term *likely* instead of probably.

This technique to find the estimators was first used and developed by Sir R.A. Fisher in 1922, who called it the maximum likelihood method.

12.5 POINT ESTIMATES

The ML estimator of population mean μ is the sample mean $\bar{x}$. Hence,

$$\hat{\mu} = \frac{\Sigma x}{n} = \bar{x}$$

We have already seen that $\bar{x}$ is an unbiased, consistent and minimum variance estimator of μ. The ML estimator of population proportion p is

$$p = \bar{p} = 1/2$$

The ML estimator of variance σ^2 of a normal distribution is

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x - \bar{x})^2 = s^2$$

which is the sample variance. The expected value of s^2 is $E(s^2) = \frac{n-1}{n}\sigma^2$ and hence s^2 is a biased estimator of σ^2. This may be rewritten as $E(s^2) = \sigma^2 - \frac{\sigma^2}{n}$ and $-\sigma^2/n$ is the bias. In order to avoid this bias, it is necessary to make the following adjustment:

$$E\left(\frac{n}{n-1}s^2\right) = \sigma^2$$

It may be noted that the factor $n/(n-1)$ indicates that s^2 is underestimating σ^2. It can be seen that as n becomes larger, $n/(n-1)$ approaches 1 and s^2 becomes an unbiased estimator of σ^2.

12.6 INTERVAL ESTIMATES

An interval estimate consists of two values between which the true population value lies with some stated level of significance. Thus, an interval estimate of a parameter (say, the population mean μ) may be $a < \mu < b$ where a and b are the lower and upper points obtained from the sample observations. Many times, it is desirable to use an interval estimate rather than a point estimate. For example, it may be more useful to say that there is a 90 per cent probability that the average income of households in a certain city lies between Rs 6,000 and Rs 12,000 pm than to state that it is estimated to be Rs 10,000 pm.

Figure 12.2 illustrates the concept of interval estimation. It will be seen that our interval estimation is Rs 6,000 – Rs 12,000 which is likely to contain the population mean μ and that the mean income of all households in that territory is between Rs 6,000 and Rs 12,000. The value Rs 6,000 is called the lower limit of the interval and the value Rs 12,000 is called the upper limit of the interval.

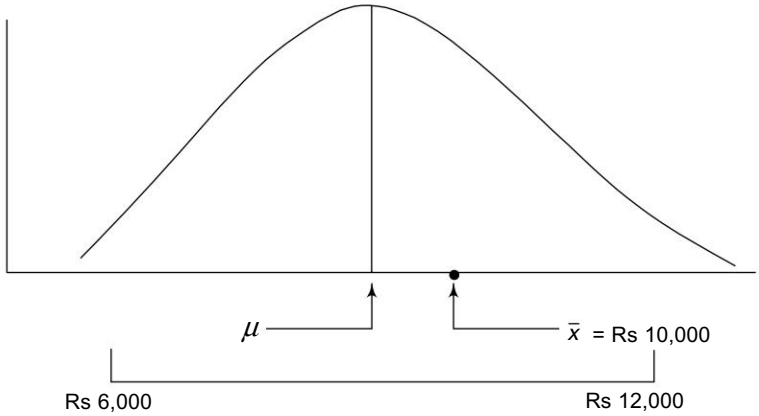


Fig. 12.2 Interval Estimation

It may be noted that each interval is constructed with regard to a given confidence level and is called a confidence interval. On the basis of the confidence level associated with a confidence interval, we can say as to how much confidence we have that the true population parameter lies in this interval.

The confidence level is denoted by $(1 - \alpha)100$ per cent. When we express it in terms of probability, it is called as the confidence coefficient and is denoted by $1 - \alpha$. The notation α (which is a Greek letter) is called the significance level.

Interval Estimation of a Population Mean: Large Samples

This section deals with the process of constructing an interval estimate of a population mean when the sample size is large. In other words, when n is 30 or more than 30, then the sample is regarded as a large sample. According to the central limit theorem, when the sample is large, then the sampling distribution of the sample mean $\bar{x}$ is almost normal regardless of the shape of the population from which the sample is selected. Accordingly, the normal distribution will be used in constructing a confidence interval for the population mean.

An interval estimate with a specified level of confidence is obtained from an interval formed by two points,

$\bar{x} - z\sigma_{\bar{x}}$ = lower point and

$\bar{x} + z\sigma_{\bar{x}}$ = upper point

where $\bar{x}$ is the mean of the sample, Z represents the number of standard errors for a specified level of significance and $\sigma_{\bar{x}}$ is the size of the standard error. When applied to the sampling distribution of the mean, the term standard error is used instead of standard deviation. It is customary to refer to the number of standard errors by a common term from the normal curve. We can say that a confidence level of 68.2 per cent is obtained when $Z = 1$, 95.4 per cent when $Z = 2$ and 99.8 per cent when $Z = 3$.

Each Z value indicates a specified level of confidence because the means of that percentage of samples that could be taken would lie between the lower and the upper points of the interval formed by that particular Z value.

As was mentioned in Chapter 11, there is a relationship between the population standard deviation and the sample standard error. This relationship is shown by the following formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Thus, we may substitute $\sigma_{\bar{x}}$, the standard error of the mean, by $\sigma/\sqrt{n}$. That is to say, the population standard deviation is to be divided by the square root of the sample size. We may now rewrite the lower point and the upper point as

$$\bar{x} - Z\frac{\sigma}{\sqrt{n}} = \text{lower point and}$$

$$\bar{x} + Z\frac{\sigma}{\sqrt{n}} = \text{upper point}$$

The value of Z would vary depending upon how much confidence we want to have for our interval estimate. We know from the normal area table that when $Z = 1.96$, it corresponds to a probability of 0.975. We may, therefore, write the interval estimate as

$$\text{P}\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Let $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} = a$ and $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}} = b$

Then $\text{P}(a < \mu < b) = 0.95$

As shown in Fig. 12.3, the random variable $\bar{x}$ takes three different values as $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$. Thus, there are three intervals:

(a) $\bar{x}_1 - 1.96\frac{\sigma}{\sqrt{n}}$ to $\bar{x}_1 + 1.96\frac{\sigma}{\sqrt{n}}$

(b) $\bar{x}_2 - 1.96\frac{\sigma}{\sqrt{n}}$ to $\bar{x}_2 + 1.96\frac{\sigma}{\sqrt{n}}$

(c) $\bar{x}_3 - 1.96\frac{\sigma}{\sqrt{n}}$ to $\bar{x}_3 + 1.96\frac{\sigma}{\sqrt{n}}$

It will be seen that the first two intervals include μ, but the third interval falls outside the two limiting values of $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$ $\left(\text{or } \bar{x} \pm 1.96\frac{\sigma}{\bar{x}}\right)$.

The probability that $\bar{x}$ will be in the interval $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$ is 0.95, which shows that there are 95 chances out of 100 that the $\bar{x}$ will be between $\mu - 1.96\frac{\sigma}{\sqrt{n}}$ and $\mu + 1.96\frac{\sigma}{\sqrt{n}}$, given that μ is in fact the

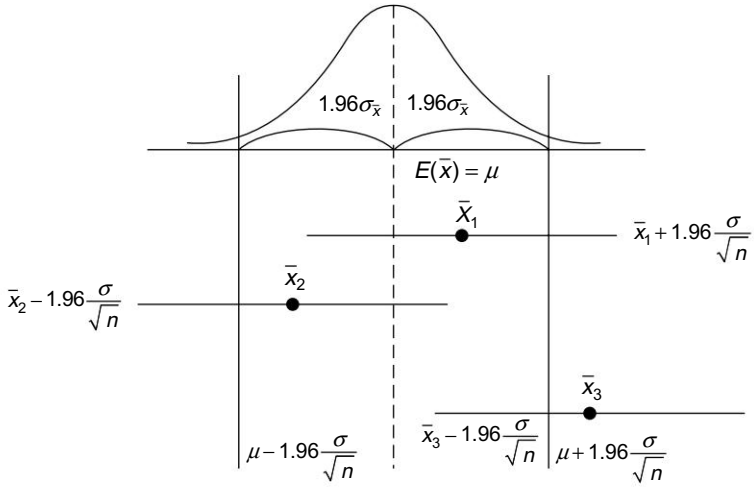


Fig. 12.3 Three Different Interval Estimates

true value of the parameter. Thus, we can see form Fig. 12.3 that when we construct our interval

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \text{ to } \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

we expect that 95 out of 100 such intervals will include μ.

We may take a few examples.

Example 12.3 A random sample of 400 firms was taken to find out the average sale per customer. The sample mean was found to be Rs 900 and the standard deviation Rs 200. Construct an interval estimate of the population mean with the confidence level of 95.44 per cent.

Solution Lower point as indicated earlier is $\bar{x} - Z\sigma_{\bar{x}}$, where $\sigma_{\bar{x}} = s/\sqrt{n}$, where s is the estimate of the standard deviation.

Thus,
$$\bar{x} - z\sigma_{\bar{x}} = 900 - 2\left(\frac{200}{\sqrt{400}}\right)$$
$$= 900 - \frac{2 \times 200}{20}$$
$$= 900 - 20 = 880$$

Upper point as indicated earlier is

$$\bar{x} + z\sigma_{\bar{x}} = 900 + 2\left(\frac{200}{\sqrt{400}}\right)$$
$$= 900 + 20 = 920$$

This can also be written as Rs 900 ± 20. We are 95.44 per cent confident that the population mean μ lies between Rs 880 and Rs 920.

Example 12.4 In the previous example, suppose we are interested in having an interval estimate with a higher confidence level, say, 99.8 per cent.

Solution The corresponding value of Z is 3. Using the same data as given in example 12.3 and taking $Z = 3$, we find

$$\bar{x} - z\sigma_{\bar{x}} < \mu < \bar{x} + z\,\sigma_{\bar{x}}$$

$$900 - 3\frac{200}{\sqrt{400}} < \mu < 900 + 3\,\frac{200}{\sqrt{400}}$$

$$900 - 30 < \mu < 900 + 30$$

$$870 < \mu < 930$$

In other words, the population mean lies between Rs 870 and Rs 930 and we are almost 100 per cent confident that it is so. Note that the interval between the lower and the upper points has widened as the level of confidence has increased. Conversely, if we reduce the level of confidence, we shall find that the interval between the two points has narrowed down.

Interval Estimation of a Population Mean: Small Samples

The preceding examples related to large samples where $n \geq 30$. When the sample size is less than 30, the sampling distribution is no longer normal. In such a case, the t distribution is used in place of the normal distribution. Before we apply the t distribution to a particular problem for interval estimation, we should know the main characteristics of the t distribution.

Characteristics of the t Distribution Both the normal and the t distributions are symmetrical but, in general, the latter is flatter than the former. Another point to note is that there is a different t distribution for each sample size. Further, as the sample size increases, the flatness of the t distribution reduces and it becomes approximately the normal distribution. When the sample size is 30 or more, the t distribution is almost the normal distribution.

The concept of degrees of freedom needs to be explained. When we use degrees of freedom, we mean a specified number of values that we can choose freely. Suppose we have 3 samples: a, b and c and we are told that their mean is 30. Symbolically, $(a + b + c)/3 = 30$. In such a case, we are free to choose values of a and b but have no freedom in choosing c. In our example, the mean of 3 values is 30, that is, the total of a, b and c is 90. Suppose we choose $a = 20$ and $b = 40$, then there is no choice left to us for choosing any value of c other than 30 as the total has to be 90 and the two values a and b add to 60. Thus, it becomes clear that when we have 3 sample values, we have $3 - 1 = 2$ degrees of freedom. When there are 20 sample values, then we have $(n - 1) = 20 - 1 = 19$ degrees of freedom, where n is the size of the sample, that is, the number of observations involved.

Appendix Table 2 gives the t distribution. It will be seen that this table gives t values for only some degrees of freedom. As mentioned earlier, there is a different t distribution for each number of degrees of freedom; the t table would be extremely lengthy to include a very large number of degrees of freedom, which physically is not possible. The common practice, therefore, is to include some frequently used values of the degrees of freedom.

Another point to note in respect of the t table is that it does not give the chance that a particular population parameter will be within a specified confidence interval. Instead, it shows the chance that the particular population parameter in which we are interested will not be within our confidence interval. For example, if we are interested to have an estimate at 95 per cent confidence level, then 0.05 column of the table (100 per cent – 95 per cent = 5 per cent) will be relevant. This 0.05 chance of error is denoted by α.

Finally, in using the t table, we should know that there should be a specific number of degrees of freedom. Without this number, it will not be possible for us to use the t table.

Let us now take an example.

Example 12.5 A firm has appointed a large number of dealers all over the country to sell its bicycles. It is interested in knowing the average sales per dealer. A random sample of 25 dealers is selected for this purpose. The sample mean is Rs 50,000 and the standard error is Rs 20,000. Construct an interval estimate with 95 per cent confidence.

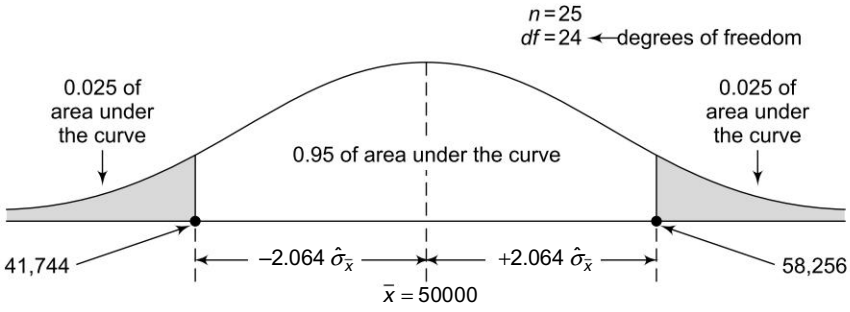


Fig. 12.4

A *t* distribution for 24 degrees of freedom, showing a 95 per cent confidence interval.

Solution From Appendix Table 2, the value of *t* for 25 – 1 = 24 degrees of freedom and corresponding to 95 per cent confidence level, which is 2.064, the interval estimates can now be calculated as follows:

$$\bar{x} - (t)\,(\hat{\sigma}_{\bar{x}}) \text{ to } \bar{x} + (t)\,(\hat{\sigma}_{\bar{x}})$$

We have to first find the numerical value of $\sigma_{\bar{x}}$ as follows:

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\text{Rs. } 20{,}000}{\sqrt{25}} = \text{Rs } 4{,}000$$

where s = standard error
n = number of observations

Interval Estimate = Rs 50,000 – (2.064) (Rs 4,000) to Rs 50,000 + (2.064) (Rs 4,000)
Rs 50,000 – Rs 8,256 to Rs 50,000 + Rs 8,256
= Rs 41,744 to Rs 58,256

We are 95 per cent confident that the interval estimate Rs 41,744 to Rs 58,256 contains the population mean.

Interval Estimate of a Population Proportion

So far, the discussion was confined to interval estimates of the population mean. In this section, we discuss interval estimate of a population proportion. The theory is identical to that used for constructing a confidence interval for the population mean. The procedure used for constructing interval estimates of proportions is, therefore, similar to that used for means.

First, the estimated standard error of the proportion, $\hat{\sigma}_{\bar{p}}$ is determined. Then the interval estimate is constructed around the sample proportion such that

$$\bar{p} - z\hat{\sigma}_{\bar{p}} = \text{lower point, and } \bar{p} + z\hat{\sigma}_{\bar{p}} = \text{upper point}$$

where z indicates the number of standard errors for the desired confidence level.

In case of a simple random sample where the population proportion is known, the standard error of the proportion is obtained by the following formula:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

where p is the proportion of items in the sample having a given characteristic, and n is the sample size. When the population proportion is not known, it can be estimated from the sample proportion $\bar{p}$ and the estimated standard error, $\hat{\sigma}_{\bar{p}}$, calculated from the following formula:

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Example 12.6 Suppose that a simple random sample of 400 families shows that 320 families own a television set and 80 do not. In other words, 80 per cent families own a television set. We have to construct a confidence interval with 95 per cent confidence.

Solution

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= \sqrt{\frac{(80)(100-80)}{400}}\,\% = \sqrt{\frac{(80\times 20)}{400}}\,\%$$

$$= \sqrt{\frac{1,600}{400}}\,\% = 2\%$$

The 95 per cent confidence interval would be

$$\bar{p} \pm z\hat{\sigma}_{\bar{p}} = 80 \pm 1.96\,(2)$$

$$= 80 \pm 3.92$$

$$= 76.08\% \text{ to } 83.92\%$$

Thus, one would be 95 per cent confident that the true percentage of television ownership lies between 76.08 and 83.92 per cent of the population.

Example 12.7 A firm manufacturing steel furniture desires to estimate the proportion of the population that uses its product. In a sample of 100 families, it is found that 35 families use its product. Estimate the proportion of the population using steel furniture of the firm, assuming the confidence level of 90 per cent.

Solution

1. Determine the proportion using steel furniture produced by the firm

$$\bar{p} = 35/100 = 0.35$$

2. Determine the estimated standard error

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= \sqrt{\frac{(0.35)(0.65)}{100}} = \sqrt{\frac{0.2275}{100}}$$

$$= \frac{0.4770}{10} = 0.048 \text{ approx.}$$

3. The value of z corresponding to 90 per cent confidence level from the normal area curve is 1.64.
4. Construct the interval estimate

$$\bar{p} \pm z\ \hat{\sigma}_{\bar{p}}$$

$0.35 \pm 1.64\ (0.048)$
0.35 ± 0.0787
0.2713 to 0.4287
27.13% to 42.87%

On the basis of these calculations, we are 90 per cent confident that the true population percentage lies within this interval.

Finite Correction Factor The relationship $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is valid when a sample is drawn from an infinite population. When the population is finite, a correction factor

$$\sqrt{\frac{N-n}{N-1}}$$

(as was mentioned in Chapter 11) is introduced and the relationship becomes

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

It can be seen that when the sample size is extremely small relative to the population, the correction factor approaches 1. However, when a sample forms a sizable proportion of the population, the correction factor assumes significance. Generally, when the sample size is more than 5 per cent of the population, the finite population correction factor is used.

Confidence Intervals for Differences

When we are asked to find the confidence intervals for differences between two population means, we have to use the following formula

$$\bar{x}_1 - \bar{x}_2 \pm Z\sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of sample 1 and sample 2, respectively, and σ_1 and N_1 and σ_2 and N_2 are the respective standard deviations and sizes of the two samples.

In order to understand the application of the formula, let us take an example.

Example 12.8 A sample of 100 brand A light bulbs showed mean lifetime of 1500 hours and standard deviation of 160 hours. A sample of 300 brand B light bulbs showed mean lifetime of 1400 hours and standard deviation of 120 hours. Find (a) 95%, and (b) 99% confidence limits for the difference of the mean lifetime of the two populations.

Solution

(a) Applying the formula given above to the data given in the above example, the 95% confidence limits are

$$1500 - 1400 \pm 1.96\sqrt{\frac{(160)^2}{100} + \frac{(120)^2}{300}}$$

$$= 100 \pm 1.96\sqrt{256 + 48} = 100 \pm (1.96)(17.49) = 100 \pm 34.28$$

(b) For 99% confidence limits, we have to take $Z = 2.58$.
Applying the same formula and using the given data, the 99% confidence limits for the difference between the two means are

$$1500 - 1400 \pm 2.58\sqrt{\frac{(160)^2}{100} + \frac{(120)^2}{300}}$$

$$= 100 \pm 2.58\sqrt{256 + 48}$$
$$= 100 \pm 2.58(17.49)$$
$$= 100 \pm 45.12$$

Confidence Internal Estimates for Standard Deviations

In order to calculate the confidence intervals for standard deviations, we use the following formula

$$s = Z\sigma_S$$

$$= \pm Z\frac{\sigma}{\sqrt{2N}}$$

where s stands for sample standard deviation.

This formula is applicable when the problem relates to normal population. Let us take an example.

Example 12.9 The standard deviation of the lifetimes of a sample of 400 tubes was found to be 250 hours. Find (a) 95%, and (b) 99% confidence limits for the standard deviation.

Solution

(a) For 95% confidence, $Z = 1.96$

$$s = \pm Z\frac{\sigma}{\sqrt{2N}}$$

Confidence limits would be

$$250 \pm 1.96\frac{250}{\sqrt{2 \times 400}}$$

or $$250 \pm 1.96\frac{250}{28.28}$$

or $$250 \pm (1.96)\ (8.84)$$

or $$250 \pm 17.33$$

(b) For 99% confidence, $Z = 2.58$

Confidence limits would be

$$250 \pm 2.58 \frac{250}{\sqrt{2 \times 400}}$$

or $250 \pm (2.58)\ (8.84)$

or 250 ± 22.81

12.7 DETERMINING THE SAMPLE SIZE IN ESTIMATION

In our discussion on Estimation, we have, so far, used the symbol n to indicate the sample size instead of a specific number. The question is: how large should the sample be? If the sample is very large, then it is obvious that we are wasting our resources when we collect the required data from that sample. In contrast, if the sample is too small, it means we are not getting adequate data, and this may defeat the very purpose for which the survey is undertaken. In view of these considerations, it becomes necessary to choose a sample of an appropriate size.

As we have not studied the entire population in a survey, there will arise some sampling error. If we have a high level of precision, then we have to choose a sample of large size to collect the required information. In such a case, the sampling error will be small. In general, we can say that when we require greater precision, we are required to take a larger sample and vice versa. In this context, we have some methods which enable us to determine sample size for any specified level of precision.

Sample Size for Estimating Mean

There are three considerations required to be checked when determining the sample size necessary to estimate the population mean. These are:

1. The extent of error or imprecision allowed
2. The degree of confidence desired in the estimate
3. The estimate of the standard deviation of the population

The first two considerations are matters of judgment involving the use of the data. The third consideration, the estimate of the standard deviation of the population, is the responsibility of the statistician. We may consider the problem of determining sample size in two different situations, namely, when the standard deviation of the population is known and when it is unknown.

Determination of Sample Size When Standard Deviation is Known

Extent of error The first consideration relates to the extent of error allowed. This is indicated by the standard error (i.e. the standard deviation of the sample means). We have to decide the magnitude of the standard error that we can tolerate. Although this is a difficult question, it is necessary to fix the limit of the standard error beyond which it should not exceed. The fixation of standard error should not be confined to overall results but should also be applied to various sub-groups. One way is to first determine the size of each sub-group on the basis of a given degree of precision. The total of the size of each sub-group could then be taken as the overall size of the sample.

The degree of confidence A second consideration is the degree of confidence that we want to have in the results of the study. In case we want to be 100 per cent confident of the results, we are left with no option but to cover the entire population. However, as this is often not possible on account of cost, time and other constraints, we should be satisfied with less than 100 percent confidence. Normally, three confidence levels, namely, 99 per cent, 95 per cent and 90 per cent are used. When a

99 per cent level is used, it implies that there is a risk of only one percent of the true population statistic falling outside the range indicated by the confidence interval. In the case of a 95 per cent confidence level, such a risk is of 5 per cent and in the case of 90 per cent confidence level, it is of 10 per cent. It should be noted that there is a trade off between extent of error permissible and a degree of confidence. For a given size of a sample, one can specify one of these two but not both of them at the same time.

The foregoing basic considerations involved in determining the sample size can be put in the form of the following formula:

$$n = \frac{Z^2\sigma^2}{E^2}$$

where n = sample size, E = the maximum error allowed and Z = the degree of confidence required. This can be better understood with the help of an example.

Example 12.10 Let us take the case where the population variance is known. Suppose that for a proposed survey we are given that the standard deviation is Rs 300 and that our estimate of sample mean should be within ± Rs 50 of the true population mean.

Solution This means that the total precision is 100 and half precision is 50. We shall use the latter value as we shall work out the calculations on the basis of one-half of the curve. In this way, certain calculations can be simplified as we know that the population mean μ divides the normal curve into two equal halves. Another point that needs to be decided relates to the degree of confidence in the result that we would like to have. Suppose that this degree of confidence is 95 per cent, which gives $z = 1.96$.

The formula to determine the size of n, as given earlier, can be written as

$$E = z\sigma_{\bar{x}}$$

$$= z\frac{\sigma}{\sqrt{n}} \left(\text{since } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ as given earlier} \right)$$

where E is the maximum error of estimate μ, which is 50 in the above example; $z = 1.96$ and σ is 300. Hence,

$$E = 1.96\frac{300}{\sqrt{n}}$$

or $$50 = 1.96\frac{300}{\sqrt{n}}$$

or $$50 = \frac{588}{\sqrt{n}}$$

or $50\sqrt{n} = 588$ or $\sqrt{n} = 11.76$. Therefore, $n = 138$ approx.

This calculation gives the sample size as 138. This indicates that when the standard deviation of population is Rs 300 and the extent of precision is ± Rs 50, a sample of 138 households needs to be selected.

Determination of Sample Size When Standard Deviation of Population is Unknown

Many a time, the standard deviation of the population is not known. In such cases too, the method followed is the same except that an estimate of the population standard deviation in place of its

previously known value is taken. Sometimes we have to undertake a pilot survey to ascertain the standard deviation. If this is not possible, we may have to use some alternative approach. As we know, the entire area under the normal curve falls within $\mu \pm 3\sigma$. This means that we should have some idea of the range of variation, that is, the difference between the highest and the lowest item. This range needs to be divided by 6 in order to get an estimate of the standard deviation.

Example 12.11 Suppose that in our previous example the minimum monthly income amongst households is Rs 3,000 and the maximum is Rs 18,000. We are asked to estimate the sample size.

Solution This gives a range of Rs 15,000, which divided by 6 yields a figure of Rs 2,500. This is the estimated value of σ. Taking other values as earlier, the sample size can be determined as shown below.

$$E = 1.96 \frac{2,500}{\sqrt{n}} \quad \text{or} \quad 50 = 1.96 \frac{2,500}{\sqrt{n}}$$

or $\sqrt{n} = 4,900/50 = 98$ or $n = (98)^2 = 9,604$

This shows that a sample of 9604 households should be taken.

Suppose a sample of 9,604 households gives a sample mean $\bar{x} = 10,000$ and a sample standard deviation s = 2,000, then the confidence interval would be $\bar{x} \pm z\ s_{\bar{x}}$.

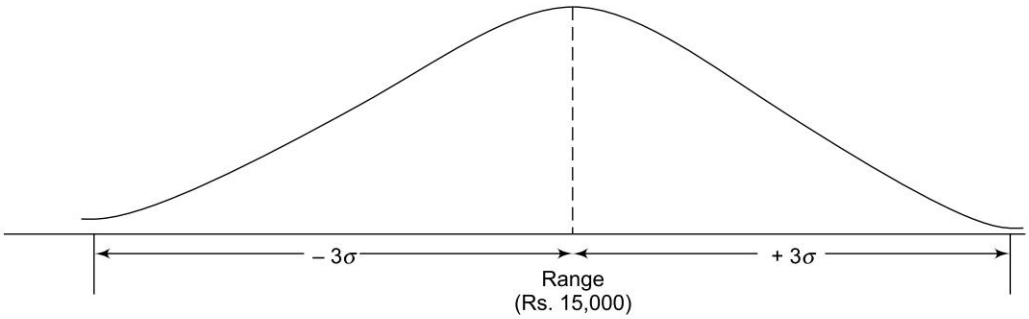


Fig. 12.5 Appropriate Relationship between the Range and Population Standard Deviation

$$10,000 \pm 1.96 \frac{2,000}{\sqrt{n}}$$

or $$10,000 \pm 1.96 \frac{2,000}{\sqrt{9,604}}$$

or $10,000 \pm 3,920/98$

or $10,000 \pm 40$

or $9,960 \leq \mu \leq 10,040$

This shows the extent of error as ± 40 as against ± 50 in the earlier example. Thus, the interval has become narrower than that envisaged earlier. This is because the sample standard deviation (2,000) is less than the estimated population standard deviation (2,500) in the earlier example. In other words, as the population standard deviation was overestimated as judged by the sample standard deviation, the confidence interval became narrower. Conversely, if the population standard deviation turns out to be underestimated, vis-à-vis sample standard deviation, the confidence interval will become wider.

Sample Size Decision for a Proportion

The foregoing discussion was carried out in relation to sample size for estimating mean values. At times, it is the proportion of population with a particular attribute that becomes more relevant than the mean value. For example, one may be more interested in knowing the proportion of households having a monthly income of, say, Rs 1,000 and less or of Rs 2,500 and above rather than in knowing the average income of the households.

The confidence interval formula for estimating population proportion is

$$\bar{p} \pm Z\sqrt{\frac{\bar{p}\bar{q}}{n}}$$

where $\bar{p}$ *= sample proportion and* $\bar{q} = 1 - \bar{p}$.

The above formula can be rewritten as $\bar{p} \pm E$, where E represents the maximum allowable sampling error, that is, the difference between the sample proportion and the population proportion. As

$$E = Z\sqrt{\frac{\bar{p}\,\bar{q}}{n}}$$

We can derive the formula for n as follows:

$$n = \frac{Z^2\bar{p}\,\bar{q}}{E^2}$$

The values of Z and E are predetermined, while the value of population proportion may be actual or estimated on the basis of the past experience. Let us take an example

Example 12.12 Suppose we are interested in estimating the proportion of households having a washing machine. We believe that the figure is about 60 per cent. Further, we decide that a standard error should not be more than 2 per cent, and we want to be 95 per cent confident in estimating the proportion of households having a washing machine.

Solution Applying the formula given earlier,

$$n = \frac{(1.96)^2\,(0.6 \times 0.4)}{(0.02)^2} = \frac{3.8416 \times 0.24}{0.0004}$$

$$= 2304.96 \text{ or } 2305 \text{ approx.}$$

In case we want to reduce the error from 2 per cent to 1 per cent, then the sample size will be

$$n = \frac{(1.96)^2\,(0.6 \times 0.4)}{(0.01)^2} = 9219.84 \text{ or } 9{,}220 \text{ approx.}$$

This shows that the reduction of error by one-half has raised the sample size by four times (i.e. from 2,305 to 9,220). In contrast, if we take a larger value of error, the sample size will be reduced.

In the same manner, we can see that if we want a higher degree of confidence in determining the sample size, then the value of Z will be higher. As a result, our sample size will be greater. In general, higher the degree of confidence, greater will be the sample size and vice versa.

It may be noted that we do not know the value of p. However, when we take p that will maximise $p(1 - p)$ and use it to calculate sample size n, we can be sure that the sample size will be large enough

to ensure that the error will be within 2 per cent. When $p = 1/2$, then $p(1-p)$ will be maximum, having its value as 1/4. On this basis, the earlier formula becomes

$$n = \frac{Z^2(1/4)}{E^2}$$

Let us take an example.

Example 12.13 A company engaged in selling ballpoint pens wishes to estimate the proportion of people who prefer its pens. It wishes to keep the error to 3 per cent with a risk of 0.0456. Determine the sample size for its proposed survey.

Solution Since the risk is set at 0.0456, the proportion in each tail of the normal curve is 0.0456/2 = 0.0228. We find from the normal area table that the Z, which corresponds to 0.0228, is $Z = 2$. Now, applying the formula

$$n = \frac{Z^2(1/4)}{E^2}$$

the sample size can be determined:

$$n = \frac{(2)^2(1/4)}{(0.03)^2} = \frac{4 \times 1/4}{0.0009} = 1111$$

If a random sample of size 1111 is taken, the error will be less than 3 per cent with a risk of 0.0456. In order to find an estimate of p, a random sample of size 1111 households should be taken. Then the maximum likelihood estimator $\hat{p} = \bar{p} = \Sigma x/n$. Assuming $\Sigma x = 400$, the ML estimator of p is

$$\hat{p} = \bar{p} = 400/1111 = 0.36 \text{ or } 36\%.$$

It may be reiterated that taking $p = 0.5$ gives $p(1-p)$ the largest value than any other value of p. This assures adequate sample size and meets the requirements placed on it.

Additional Examples

Example 12.14 A random sample of 100 students belonging to a college was taken. It was found that the mean height of these students was 168.75 cm. What should be the confidence intervals for estimating the mean height of the entire population of students at **(a)** 90%, and **(b)** 99% assuming the population standard deviation as 7.5 cm?

Solution

(a) For 90% confidence level $Z = 1.645$

$$\bar{x} - Z\frac{\sigma}{\sqrt{n}} \text{ lower point}$$

$$\bar{x} + Z\frac{\sigma}{\sqrt{n}} \text{ upper point}$$

$$168.75 - 1.645\frac{7.5}{\sqrt{100}} = 168.75 - (1.645 \times 0.75)$$

$$= 168.75 - 1.23$$

$$= 167.52$$

$$168.75 + 1.23 = 169.98$$

Hence, the confidence intervals are 167.52 to 169.98 cm.

(b) For 99% confidence level $Z = 2.58$

$$\bar{x} - Z\frac{\sigma}{\sqrt{n}} = 168.75 - 2.58\frac{7.5}{\sqrt{100}}$$
$$= 168.75 - (2.58 \times 0.75)$$
$$= 168.75 - 1.935$$
$$= 166.815$$
$$168.75 + 1.935 = 170.685$$

Hence, the confidence intervals are: 166.815 to 170.685 cm.

Example 12.15 Television advertisers mistakenly believe that most viewers understand most of the advertising that they see and hear. In this connection, a research study covering 2,300 viewers above the age of 20 years was taken. Each viewer looked at 30-second television advertising excerpts. It was found that 1,914 viewers misunderstood either the entire excerpt or a part of it. Determine a 95 per cent confidence interval for the proportion of all viewers (of which the sample is representative) that will misunderstand all or part of the television excerpts used in this study.

Solution

Lower point: $\bar{p} - Z\,\hat{\sigma}_{\bar{p}}$

Upper point: $\bar{p} + Z\,\hat{\sigma}_{\bar{p}}$

The formula for calculating $\hat{\sigma}_{\bar{p}}$ is

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}}$$

Substituting the values given in the problem in this formula,

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\frac{1914}{2300}\left(1-\frac{1914}{2300}\right)}{2300}}$$
$$= \sqrt{\frac{0.83(1-0.83)}{2300}}$$
$$= \sqrt{\frac{0.83\times 0.17}{2300}} = \sqrt{\frac{0.1411}{2300}} = 0.00783$$

$$\bar{p} - Z\,\hat{\sigma}_{\bar{p}} = 0.83 - (1.96 \times 0.00783)$$
$$= 0.83 - 0.0153$$
$$= 0.815$$
$$\bar{p} + Z\,\hat{\sigma}_{\bar{p}} = 0.83 + 0.0153$$
$$= 0.845$$

Hence, the confidence limits are 0.815 and 0.845.

Example 12.16 A random sample of $n = 500$ observations from a binomial population produced $x = 240$ successes.

(a) Find a point estimate for p, and place a 95% confidence interval.

(b) Find a 90 per cent confidence for p.

Solution

(a) $$\bar{p} = \frac{240}{500} = 0.48$$

where $\bar{p}$ is the sample proportion and is a convenient estimate of the population proportion p.

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= \sqrt{\frac{(0.48)(0.52)}{500}}$$

$$= \sqrt{\frac{0.2496}{500}}$$

$$= 0.022$$

Taking the value of z as 1.96 corresponding to 95 per cent confidence level.

Interval estimate

$$\bar{p} \pm Z\,\hat{\sigma}_{\bar{p}}$$

$$0.48 \pm 1.96\ (0.022)$$

$$0.48 \pm 0.043$$

$$0.437 \text{ to } 0.523$$

(b) For 90 per cent confidence $Z = 1.64$

$$\bar{p} + Z\,\hat{\sigma}_{\bar{p}}$$

$$0.48 \pm 1.64\ (0.022)$$

$$= 0.48 \pm 0.036$$

$$= 0.444 \text{ to } 0.516$$

Example 12.17 A person tosses a coin 40 times and obtains 24 heads. Find **(a)** 95 per cent and **(b)** 99.73 per cent confidence limits for the proportion of heads that he would obtain in an unlimited number of the tosses of the coin.

Solution

(a) The formula for finding the confidence interval is $\bar{p} \pm Z\,\hat{\sigma}_{\bar{p}}$ where Z at 95% level is 1.96.

We have to first find the value of $\hat{\sigma}_{\bar{p}}$, for which we apply the following formula:

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= \sqrt{\frac{0.6(1-0.6)}{40}} \qquad (\bar{p} = 24/40 = 0.6)$$

$$= \sqrt{\frac{0.24}{40}} = 0.077$$

$$\bar{p} \pm Z\,\hat{\sigma}_{\bar{p}} = 0.6 \pm 1.96\,(0.077)$$
$$= 0.6 \pm 0.15$$
$$= 0.6 - 0.15 = 0.45$$
$$0.6 + 0.15 = 0.75$$

Hence, the confidence limits are 0.45 to 0.75.

(b) $\bar{p} \pm Z\,\hat{\sigma}_{\bar{p}}$ where Z at 99.73% level is 3

$$= 0.6 \pm 3\,(0.077)$$
$$= 0.6 \pm 0.231$$
$$= 0.6 - 0.23 = 0.37$$
$$0.6 + 0.23 = 0.83$$

Hence, the confidence limits are 0.37 to 0.83.

Example 12.18 The standard deviation of the durability of an article is 144 hours. Determine the sample size that must be taken to be **(a)** 95%, **(b)** 99% confident so that the error in the estimated mean durability will not exceed **(i)** 15 hours, **(ii)** 20 hours.

Solution The formula to determine the sample size is

$$E = Z\,\sigma_{\bar{x}}$$

$$= Z\frac{\sigma}{\sqrt{n}} \quad \left(\text{where } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\right)$$

where E is the margin of error of the estimate μ.

(a) (*i*) $$15 = 1.96\frac{144}{\sqrt{n}}$$

or $$15 = \frac{282.24}{\sqrt{n}}$$

or $$15\sqrt{n} = 282.24$$

or $$\sqrt{n} = \frac{282.24}{15} = 18.816$$

$$\therefore \quad n = (18.816)^2 = 354.04 \text{ or } 354$$

(*ii*) $$E = Z\frac{\sigma}{\sqrt{n}}$$

$$20 = 1.96\frac{144}{\sqrt{n}}$$

or $$20\sqrt{n} = 282.24$$

or $\sqrt{n} = \frac{282.24}{20} = 14.112$

$\therefore \quad n = (14.112)^2 = 199.1485$ or 199

(b) (*i*) $E = Z\frac{\sigma}{\sqrt{n}}$

$15 = 2.58\frac{144}{\sqrt{n}}$

or $15\sqrt{n} = 371.52$

or $\sqrt{n} = \frac{371.52}{15}$

or $\sqrt{n} = 24.768$

$\therefore \quad n = (24.768)^2 = 613.45$ or 613

(*ii*) $E = Z\frac{\sigma}{\sqrt{n}}$

$20 = 2.58\frac{144}{\sqrt{n}}$

or $20\sqrt{n} = 371.52$

or $\sqrt{n} = \frac{371.52}{20} = 18.576$

$\therefore \quad n = (18.576)^2$

$= 345.06$ or 345

Example 12.19 You have been asked to determine the sample size for a proposed survey of households in a town. One of the main objectives of the proposed survey is to estimate the annual mean income of households. You have been told that the standard deviation of the household incomes is Rs 150.

(a) It is desired that the estimate should be ± Rs 20 of the true population value. If 95 per cent confidence level is desired, what should be the sample size?

(b) If you were to double the precision and at the same time to have 99 per cent confidence, what size of the sample would you take?

Solution

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$E = Z\,\sigma_{\bar{x}}$

$E = Z\frac{\sigma}{\sqrt{n}}$

(a) or $20 = 1.96\frac{150}{\sqrt{n}}$

or $20 = \dfrac{294}{\sqrt{n}}$

or $20\sqrt{n} = 294$

or $\sqrt{n} = \dfrac{294}{20}$

or $n = \left(\dfrac{294}{20}\right)^2 = (14.7)^2 = 216.09$ or 216

(b) In order to double the precision, we have to reduce E as ± 10. Again, the level of confidence is to be raised to 99%, which means $Z = 2.58$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$E = Z\ \sigma_{\bar{x}}$$

$$E = Z\ \frac{\sigma}{\sqrt{n}}$$

$$10 = 2.58\frac{150}{\sqrt{n}} \quad \text{or} \quad 10\sqrt{n} = 387$$

or $\sqrt{n} = \dfrac{387}{10}$ or $n = (38.7)^2 = 1498$

Example 12.20 A sample of 100 bolts from one machine showed that 12 were defective. A sample of 200 bolts from another machine showed that 15 were defective. Find (a) 95%, (b) 99%, and (c) 99.73% confidence limits for the difference in proportion of defective bolts from the two samples.

Solution

(a) $p_1 - p_2 \pm Z\sqrt{\left(\dfrac{p_1 q_1}{n_1}\right) + \left(\dfrac{p_2 q_2}{n_2}\right)}$

$$= 0.12 - 0.075 \pm 1.96\sqrt{\frac{(0.12)(0.88)}{100} + \frac{(0.075)(0.925)}{200}}$$

$$= 0.12 - 0.075 \pm 1.96\sqrt{0.001056 + 0.0003468}$$

$$= 0.045 \pm 1.96\sqrt{0.0014028}$$

$$= 0.045 \pm 1.96 \times 0.037$$

$$= 0.045 \pm 0.073$$

(b) $0.12 - 0.075 \pm (2.58)\ (0.037)$ (From the above calculations)

$$= 0.045 \pm 0.095$$

(c) $0.12 - 0.075 \pm (3)\,(0.037)$

$= 0.045 \pm 0.111$

Example 12.21 An urn contains an unknown proportion of red and white marbles. A random sample of 60 marbles, selected with replacement from the urn, showed that 70% were red. How large a sample of marbles should be taken in order to be (a) 95%, (b) 99.73% confident that the true proportion does not differ from the sample proportion by more than 5%?

Solution

(a) 95% confidence gives $Z = 1.96$

$$n = \frac{Z^2\,pq}{E^2}$$

$$= \frac{(1.96)^2\,(0.7)(0.3)}{(0.05)^2}$$

$$= \frac{3.8416 \times 0.21}{0.0025}$$

$$= 322.69$$

or At least 323

(b) 99.73% confidence gives $Z = 3$

Hence, $$n = \frac{(3)^2\,(0.7)(0.3)}{(0.05)^2}$$

$$= \frac{9 \times 0.7 \times 0.3}{0.0025} = \frac{1.89}{0.0025} = 756$$

that is, at least 756.

Increasing Precision and Accuracy In our examples of estimating a population mean or a population proportion on the basis of sample evidence, it became obvious that the precision of the estimate can be increased by giving up some probability of being right. On the other hand, the probability of being right can be increased by giving up some precision. It may be reiterated here that precision relates to the width of the confidence interval, whereas accuracy (or reliability) relates to the confidence coefficient. The only way to increase both precision and accuracy (or reliability) is to increase the size of the sample. This means that for a given sample size, the manner in which the conclusions are stated always involves striking a compromise between the desire for precision and the desire for accuracy (or reliability). As such the need to increase the sample size becomes obvious if one decides to have greater precision as well as greater accuracy (or reliability) at the same time. Increasing sample size obviously involves higher costs. In view of this in any business problem where sampling is involved, the management must carefully weigh alternative costs before taking a final decision on the sample size.

In this chapter, we have considered the problem of estimating a parameter on the basis of a statistic—the sample evidence. We are now ready to embark on another important topic, viz., tests of hypotheses, which forms the subject matter of the next chapter.

GLOSSARY

Confidence interval	A specified range of numbers within which a population parameter is likely to fall.
Confidence level	Denoted by $(1 - \alpha)$ 100 per cent, a confidence level states how much confidence we have so that the true population parameter lies within a confidence interval.
Confidence limits	The upper and lower boundaries of a confidence interval.
Consistent estimator	An estimator that gives values more closely approaching the population parameter as the sample size increases.
Degrees of freedom	The number of values in a sample that can be freely specified once something about a sample is known.
Efficient estimator	An estimator that has smaller standard error as compared to some other estimator of the population parameter.
Estimate	The value of a sample statistic that is used to find a corresponding population parameter.
Estimation	A procedure for assigning value or values to a population parameter based on the data collected from a sample.
Estimator	A sample statistic that is used to estimate a population parameter.
Interval estimate	The estimate of an interval in which an unknown population characteristic is expected to lie for a given level of significance.
Method of maximum likelihood	A method that provides estimators with the desirable properties, such as efficiency, consistency and sufficiency. It usually does not give unbiased estimators.
Parameter	The numerical value of a summary measure in the population, such as the mean μ or the standard deviation σ.
Point estimate	The value of a sample statistic pertaining to the corresponding population parameter.
Student's t distribution	A probability distribution used when the population standard deviation is unknown and the sample size $n < 30$.
Sufficient estimator	An estimator that uses all the available data pertaining to a parameter.
Unbiased estimator	When the expected value of the statistic used as an estimator is equal to the population parameter to be estimated, then the estimator is unbiased.

LIST OF FORMULAE

1. Point estimate of the population mean

$$\hat{\mu} = \frac{\Sigma x}{n} = \bar{x}$$

where $\hat{\mu}$ is an estimate of the population mean and $\bar{x}$ is an unbiased, consistent and minimum variance estimator of μ.

2. Estimated standard deviation of the population

$$\hat{\sigma} = s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

Sample standard deviation s can be used to estimate the population standard deviation.

3. Point estimate of the population variance

$$\hat{\sigma}^2 = s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

4. Point estimate of the population proportion

$$p = \bar{p}$$

where p is the estimate of the population proportion and $\bar{p}$ is the sample proportion.

5. Estimated standard error of the mean of a finite population

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N - n}{N - 1}}$$

6. Standard deviation of the population proportion

$$\sigma_p = \sqrt{\frac{p(1 - p)}{n}}$$

7. Estimated standard deviation of the population proportion. Standard error of the sample proportion is used when the population proportion is unknown.

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

8. Interval estimate of the population mean

$$\mu = \bar{x} \pm Z\ \sigma_{\bar{x}}$$

This could be written as

$$\bar{x} - Z\sigma_{\bar{x}} < \mu < \bar{x} + Z\sigma_{\bar{x}}$$

where Z is the standard normal random variable whose value is determined by the desired level of confidence

For 99 per cent confidence $Z = 2.58$

For 95 per cent confidence $Z = 1.96$

For 90 per cent confidence $Z = 1.645$

9. Interval estimate of population proportion

$$p = \bar{p} \pm z\sqrt{\frac{\bar{p}\bar{q}}{n}}$$

This can be written as $\bar{p} - z\sqrt{\frac{\bar{p}\bar{q}}{n}} \leq p \leq \bar{p} + z\sqrt{\frac{\bar{p}\bar{q}}{n}}$

where z is as stated in 8 above.

10. Formula for determining sample size for estimating μ

$$n = \frac{z^2\sigma^2}{E^2}$$

where n = sample size
E = the margin of error allowed
Z as in 8 above.

11. Formula for determining sample size for estimating population proportion p

$$n = \frac{z^2\bar{p}\bar{q}}{E^2}$$

where $\bar{p}$ = sample proportion
$\bar{q} = 1 - \bar{p}$
z as in 8 above.

QUESTIONS

12.1 Given below are twelve statements. Indicate in each case whether it is true or false:

(a) A point estimate is often insufficient because it is either right or wrong.
(b) An interval estimate is a range of values used to estimate an unknown population parameter.
(c) There are three types of estimates that we can make about a population.
(d) In a normal distribution, 100 per cent of the population lies within $\pm 3\sigma$ of the mean.
(e) There are three criteria that a good estimator of a population parameter must satisfy.
(f) The probability that a population parameter lies within a specified interval estimate is known as the confidence interval.
(g) If a statistic tends to have values higher than the population parameter as frequently as it tends to have values that are lower, then it is said that the statistic is an unbiased estimate of the parameter.
(h) The standard error of the proportion is calculated by the formula $\sqrt{npq}$.
(i) While determining the sample size, one can specify either the degree of precision or the degree of confidence but not both of them at the same time.
(j) An increase in the width of a confidence interval implies that the confidence level associated with the interval has also increased.
(k) The sample median is always the best estimator of the population median.
(l) The relationship $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is valid when a sample is drawn from an infinite population.

Multiple Choice Questions (12.2 to 12.12)

12.2 A good estimator should be
(a) unbiased (b) consistent (c) efficient (d) sufficient
(e) all of these

12.3 Suppose that 160 persons were asked whether they liked a certain T.V. programme. Of these, 40 persons said they liked it and 120 said they did not like it. Assuming 'like' means a success, which of the following is correct?
(a) $p = 0.75$ (b) $p = 0.25$ (c) $\bar{p} = 0.25$ (d) $\bar{p} = 0.33$

12.4 Assuming a sample is taken and its mean is calculated as 240. For a 90 per cent confidence interval for μ, the upper limit comes to 280, what is the lower limit of this confidence interval?
(a) 220 (b) 240 (c) 200
(d) cannot be determined from the information given

12.5 For using a t distribution table, which of the following is a necessary condition?
(a) the population is large (b) σ is not known but s is known
(c) n is small (d) all of these
(e) (b) and (c)

12.6 If the population proportion is not known, the standard error of the proportion can be estimated by using the formula
(a) $\sqrt{\bar{p}\bar{q}/n}$ (b) $\sqrt{npq}$ (c) $\sqrt{pq/n}$ (d) $\sqrt{n\bar{p}\bar{q}}$

12.7 The standard deviation of the sampling distribution of a statistic is called
(a) mean error (b) sampling error
(c) standard deviation (d) standard error

12.8 Which part of the area under the normal curve is represented by the coefficient Z?
(a) right tail (b) left tail
(c) both tails (d) none of the above

12.9 Maximum likelihood estimator of population mean of a normal distribution is
(a) $\bar{x}$ (b) $n\bar{x}$ (c) $\bar{x}/n$ (d) none of these

12.10 When the sample size is large and the standard deviation is known, then the interval estimate is given by the formula
(a) $\bar{x} \pm zs_{\bar{x}}$ (b) $\bar{x} \pm z\sigma_{\bar{x}}$ (c) $\bar{x} \pm t\sigma_{\bar{x}}$ (d) $\bar{x} \pm ts_{\bar{x}}$

12.11 Given $\bar{x} = 60$, $s = 10$ and $n = 100$, then the standard error of the sample mean is
(a) 1.96 (b) 1 (c) 10 (d) none of these

12.12 A sampling distribution is the distribution of
(a) mean (b) variance (c) statistic (d) parameter

12.13 Differentiate between **(a)** Estimate and Estimator; **(b)** Point estimate and Interval estimate.

12.14 What are the advantages of using an Interval estimate over a Point estimate?

12.15 What are the criteria of a good estimator? Explain each of the criteria briefly.

12.16 Based on knowledge about the desirable qualities of estimators, why is $\bar{x}$ considered as the "best" estimator of the population mean μ?

12.17 Write a note on maximum likelihood estimator.

12.18 What is the confidence coefficient? How is it useful in determining the confidence limits in interval estimation?

12.19 A random sample of $n = 250$ has given a mean $\bar{x} = 25$ and standard error $\sigma_{\bar{x}} = 4$. Construct the 95 per cent confidence limits of the mean. Comment on the result.

12.20 A random sample of 200 consumer accounts at a shop is selected for the purpose of estimating the mean number of transactions per year for each customer. The sample mean is 43. Determine 98 per cent confidence interval for the mean number of transactions of all consumer accounts with the shop. The population standard deviation is 2.5.

12.21 The mean weight of a random sample of size 100 from students' population is 65.8 kg, and the standard deviation is 4 kg. Set up 95 per cent confidence limits of the mean weight of the students' population.

12.22 In a sample of 81 items taken from a large consignment, some were found to be defective. If the standard error of the proportion of defective items in the sample is 1/18, find 95 per cent confidence limits of the percentage of defective items in the consignment.

12.23 Data below show the thickness of coating of a plastic paint taken for nine samples.

20.5 21.2 18.6 20.4 19.8 17.8 23.2 22.4 20.6

Compute the standard deviation and set up a 90 per cent confidence limits for the standard deviation of the population.

12.24 For the following sample sizes and confidence levels, find the appropriate t values for constructing confidence intervals:

	Sample Size (n)	*Confidence Interval (%)*
(a)	20	90
(b)	8	95
(c)	30	98
(d)	25	99
(e)	10	95
(f)	4	90

12.25 In 60 tosses of a coin, 35 heads were obtained. Find 90 per cent confidence limits for the proportion of heads, that would be obtained in an unlimited number of tosses of the coin.

12.26 Construct a 90 per cent confidence interval for the market share of a brand if the sample market share (number of shops considered 25) has been found out to be 30 per cent.

12.27 Find a 95 per cent confidence interval for a population mean μ for the following:

(a) $n = 36,\ \bar{x} = 13.1,\ s^2 = 3.42$

(b) $n = 64,\ \bar{x} = 2.73,\ s^2 = 0.1047$

(c) $n = 41,\ \bar{x} = 28.6,\ s^2 = 1.09$

12.28 A random sample of 400 television tubes was tested and 40 tubes were found to be defective. With confidence coefficient equal to 0.9, estimate the interval within which the true fraction defective lies.

12.29 Data below refer to the dividend yield from samples of corporations in electronic industry and textile industry. There is reason to believe that the population dividends are normally distributed with equal variances. Find a 95 per cent confidence interval for the population difference in dividends and assess whether the electronic industry pays higher than the textile industry.

	Electronics	*Textiles*
Sample size	10	12
Mean dividend (Rs)	24.6	17.5
Standard deviation (Rs)	5.65	3.95

12.30 A survey is conducted among random samples of employees in a large organisation as to whether the staff should be rotated between the two shifts or fixed by each shift. Out of 340 workmen in the first shift 187 preferred rotation while in an independent sample of 291 in the second shift, 192 preferred rotation. Find a 95 per cent confidence interval within which the true difference in the population proportions may be expected to lie. Can it be inferred that the preference for rotation of shift is higher for the second shift staff?

12.31 To estimate the proportion of unemployed workers in a certain city, a random sample of 400 persons from the working class was taken. Of these, 25 were unemployed.

(a) Estimate the true proportion of unemployed workers and place bounds on the error of estimation.

(b) How many persons must be sampled to reduce the bound on the error to 0.02?

12.32 A section manager wishes to estimate the mean number of seconds required by a worker to do a particular task. He observed the worker on 144 randomly selected occasions. The average number of seconds required in the 144 observations was 100 seconds and the standard deviation was 10 seconds.

What size of sample (i.e. how many observations) would be necessary to estimate the true mean within an error of 0.5 second with a 95 per cent confidence coefficient? (Use the standard deviation of the sample as the best available estimate of the standard deviation of the population).

12.33 As a business manager of a large company, you wish to check the inventory records against the physical inventories by a sample survey. You want to be almost sure that the maximum sampling error should not be more than 5 per cent above or below the true proportion of the inaccurate records. The proportion of the inappropriate records is estimated at 35 per cent from past experience. Determine the sample size.

12.34 Determine the confidence internal for 95% confidence limit for a sample of $n = 465$ which is drawn. These observations generate a mean of 180 and a sample standard deviation of 50.

12.35 Determine the confidence interval for 90% confidence limit for a sample of size 355 that has been taken, the proportion P is calculated as 0.55.

12.36 The output voltage of a power source is known to have a standard deviation of 10 V. Fifty readings are randomly selected, yielding an average of 118 V. Find a 95% confidence interval for the population mean voltage.

12.37 Two operators perform the same operation of applying plastic coating to a part. A random sample of 100 parts from the first operator shows that 6 are non-conforming. A random sample of 200 parts from the second operator shows that 8 are non-conforming. Find a 90% confidence interval for the difference in the proportion of non-conforming parts produced by the two operators.

12.38 A random sample of 100 teachers in a large metropolitan area revealed a mean weekly salary of Rs 4,870, with a standard deviation of Rs 480. With what degree of confidence, can we assert that the average weekly salary of all teachers in the metropolitan area is between Rs 4,720 and Rs 5,020?

12.39 For a particular consumer product, the sales per retail outlet last year in a sample of $n_1 = 10$ stores was x_1 = Rs 3,425 with s_1 = Rs 200. For a second product, the mean sales per outlet in

a sample of $n_2 = 12$ stores was $x_2 =$ Rs 3,250 with $s_2 =$ Rs 175. The sales per outlet are normally distributed. Estimate the difference between the mean level of sales per outlet last year using 95% confidence level. If the size of the sample s and $n_1 = 20$ and $n_2 = 24$, what would the difference in means be?

12.40 In a regression problem with a sample size of 17, the slope was found to be 3.73 and the standard error of estimate was 28.653. The quantity $(\Sigma X^2 - nX^2) = 871.56$. Find 98% confidence interval for the population slope.

12.41 The standard deviation of breaking strength of 100 cables tested by a company was 180 N. Find (a) 95%, (b) 99%, and (c) 99.73% confidence limits for the standard deviation of all cables produced by the company.

The McGraw-Hill Companies

CHAPTER 13 TESTING HYPOTHESES

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the concepts of hypothesis, null and alternative hypothesis, and the procedure involved in testing them
- distinguish between Type I and Type II errors in hypothesis testing and the relationship between them
- formulate and test an appropriate null hypothesis in situations involving means, proportions and the differences between means and proportions in large or small samples
- understand the power of a hypothesis test
- specify the most appropriate test of hypothesis in a given situation, apply the procedure and make inferences from the result.

Chapter Prerequisites

Before starting work on this chapter, make sure you are quite familiar with all the material on sampling and sampling distribution covered in Chapter 11.

13.1 INTRODUCTION

This chapter deals with problems that are quite similar to those of the preceding chapter on estimation. However, the procedure to examine them is different. The estimation of population parameters and the testing of hypotheses concerning those parameters are similar techniques, but at the same time there are major differences in the interpretation of results arising from each method. When we are concerned with measurement, say, of expenditure on entertainment, the appropriate method would be the process of estimation. When we are involved in decision-making such as whether we should raise the price of our product by 5 per cent or not, it is the hypothesis testing that would enable us to take a proper decision. In addition, hypothesis testing is very helpful in examining the validity or otherwise of theories such as wage increase leads to rising prices. It may, however, be noted that sometimes such situations arise that it may be difficult to interpret correctly the results emerging from hypothesis testing.

Concept of Hypothesis

Before we proceed to discuss the procedure involved in hypothesis testing, we should be familiar with the concept of hypothesis. A hypothesis is a proposition that we want to verify. For example, we think that companies manufacturing washing machines spend at least 10 per cent of their annual profits on advertising. This is a statement or proposition that we would like to verify whether it is true or not. For this purpose, we have to collect the relevant information, process it using statistical techniques and then test the above hypothesis. It may be mentioned that while a hypothesis is useful, it is not always necessary. Many a time, we may be interested in collecting and analysing data, indicating the main characteristics without a hypothesis excepting the one that we may suggest incidentally during the course of our study. However, in a problem-oriented study, it is necessary to formulate a hypothesis or hypotheses in as clear terms as possible. In such studies, hypotheses are generally concerned with the causes of a certain phenomenon or a relationship between two or more variables under estimation.

Concept of the Null Hypothesis

A null hypothesis is a statement about a population parameter (such as μ), and the test is used to decide whether or not to accept the hypothesis. A null hypothesis, identified by the symbol H_0, is always one of status quo or no difference. If the null hypothesis is false, something else must be true. In view of this possibility, whenever a null hypothesis is specified, an alternative hypothesis identified by symbol H_1, must also be specified. It is the opposite of the null hypothesis. It should be clear that both null and alternative hypotheses cannot be true and only one of them must be true. When we are involved in such an exercise, our conclusion must result into the acceptance of one hypothesis and the rejection of the other.

Let us take a non-statistical example of these two concepts. Suppose that a person is facing a legal trial for committing a crime. The judge looks into all the evidence for and against it, listens very carefully the prosecution's and defendant's arguments, and then decides the case and gives his verdict. Now, the verdict could be

1. That the person has not committed the crime.
2. That the person has committed the crime.

As we all know that the judge will not give his judgment unless he is very sure about his proposed decision. In Statistics, the first verdict that the person has not committed the crime is called the *null hypothesis* and the verdict that the person has committed the crime is called the *alternative hypothesis*. When the hearing in the court commences, it is assumed that the person has not committed the crime. Similarly, in Statistics, at the beginning when a certain problem is being looked into, the null hypothesis is generally the hypothesis that is assumed to be true. In the example given above, the two hypotheses are stated as below:

Null hypothesis H_0 The person has not committed the crime.

Alternative hypothesis H_1 The person has committed the crime.

Let us take an example that belongs to Statistics. Suppose that a drug manufacturing company has installed a machine that fills automatically 5 grams in a small bottle. However, in reality, this may or may not be true. To begin with we shall assume that what the company claims is true. Thus, our null hypothesis will be

$H_0 : \mu = 5$ grams (the company's claim is true.)

It may be noted that this null hypothesis can be written in a different way as well:

$H_0 : \mu \geq 5$ grams

Although the company has claimed that on an average each small bottle has 5 grams of the drug, but when we include '>'(greater than) symbol, the claim of the company remains unaffected. It should be clear that the charge against the company would be made when it is found that on an average, less than 5 grams of drug is contained in small bottles. This means that our hypothesis test will not be affected whether the null hypothesis uses an '=' or a '≥' sign as long as the alternative hypothesis uses a < sign.

There are two points to be noted: First, there can be two or more alternative hypotheses though only one alternative hypothesis can be tested at one time against the null hypothesis. Second, both in the null and alternative hypothesis, the sample statistic such as $\bar{x}$ or $\bar{p}$ is not used. Instead the population parameter such as μ or p is used.

13.2 PROCEDURE IN HYPOTHESIS TESTING

There are five steps involved in testing a hypothesis. These are briefly discussed below:

1. Formulate a Hypothesis The first step is to set up two hypotheses instead of one in such a way that if one hypothesis is true, the other is false. Alternatively, if one hypothesis is false or rejected, then the other is true or accepted. We have already seen these two hypotheses, that is, H_0 and H_1 earlier.

2. Set up a Suitable Significance Level Having formulated the hypothesis, the next step is to test its validity at a certain level of significance. The confidence with which a null hypothesis is rejected or accepted depends upon the significance level used for the purpose. A significance level of, say 5 per cent, means that in the long run, the risk of making the wrong decision is about 5 per cent. In other words, one is likely to be wrong in accepting a false hypothesis or in rejecting a true hypothesis on 5 out of 100 occasions. A significance level of, say, 1 per cent implies that there is a risk of being wrong in accepting or rejecting the hypothesis on 1 out of every 100 occasions. Thus, a 1 per cent significance level provides greater confidence to the decision than a 5 per cent significance level.

3. Select Test Criterion The next step in hypothesis testing is the selection of an appropriate statistical technique as a test criterion. There are many techniques from which one is to be chosen. For example, when the hypothesis pertains to a large sample of more than 30, the *Z*-test implying normal distribution is used. When a sample is small (less than 30), the *t*-test will be more suitable. The test criteria that are frequently used in hypothesis testing are *Z, t, F* and χ^2.

4. Compute After having selected the statistical technique to test the hypothesis, the next step involves various computations necessary for the application of that particular test. These computations include the testing statistic as also its standard error.

5. Make Decisions The final step in hypothesis testing is to draw a statistical decision, involving the acceptance or rejection of the null hypothesis. This will depend on whether the computed value of the test criterion falls in the region of acceptance or in the region of rejection at a given level of significance. It may be noted that the statement rejecting the hypothesis is much stronger than the statement accepting it. It is much easier to prove something false than to prove it true. Thus, when we say that the null hypothesis is not rejected, we do not categorically say that it is true.

13.3 TWO TYPES OF ERRORS IN HYPOTHESIS TESTING

At this stage, it is worthwhile to know that when a hypothesis is tested, there are four possibilities:

1. The hypothesis is true but our test leads to its rejection.

2. The hypothesis is false but our test leads to its acceptance.
3. The hypothesis is true and our test leads to its acceptance.
4. The hypothesis is false and our test leads to its rejection.

Of these four possibilities, the first two lead to an erroneous decision. The first possibility leads to a Type I error and the second possibility leads to a Type II error. This can be shown as follows:

Table 13.1 Types of Errors in Hypothesis Testing

	State of Nature	
Decision	*H_0 is True (S_1)*	*H_0 is False (S_2)*
Accept H_0 (A_1)	Correct decision	Type II error (β)
Reject H_0 (A_2)	Type I error (α)	Correct decision

Table 13.1 indicates that one of the two conditions (states of nature) exists in the population, that is, either the null hypothesis is true or it is false. Similarly, there are two decision alternatives: accept the null hypothesis or reject the null hypothesis. Thus, two decisions and two states of nature result into four possibilities.

In any hypothesis testing, there is a risk of committing Type I and Type II errors. In case we are interested in reducing the risk of committing a Type I error, we should reduce the size of the rejection region or level of significance, indicated in Table 13.1, by α. When $\alpha = 0.10$, it means that a true hypothesis will be accepted in 90 out of every 100 occasions. Thus, there is a risk of rejecting a true hypothesis in 10 out of every 100 occasions. To reduce this risk, we may choose $\alpha = 0.01$, which implies that we are prepared to take 1 per cent risk. That is, the probability of rejecting a true hypothesis is merely 1 per cent instead of 10 per cent as in the previous case.

It may be noted that the choice of a level of significance involves a compromise between two kinds of risk. A reduction in the probability of committing a Type I error increases the risk of committing a Type II error, that is, the probability of accepting a null hypothesis when it is false, increases. On the other hand, if we choose a significance level ($\alpha = 0.05$) of 5 per cent as compared to 1 per cent, it means that we are taking a greater risk of committing a Type I error, though this would reduce the risk of committing a Type II error. An increase in the sample size is the only way to reduce the risk of committing both the types of errors. Let us take an example to understand Type I and Type II errors.

A business firm wants to introduce another product in the market. Thus, it has to choose one of the two decisions—to introduce the product or not to introduce it. Now, the states of nature are two, namely, the failure of the product and the success of the product. The firm thus runs the risk of a wrong decision in two ways. The first risk is that the firm does not introduce the product though it would have succeeded had it been introduced. This type of risk is denoted by a Type II (β) error. The second risk is that the firm introduces the product but it does not succeed. This type of risk is denoted by a Type I (α) error. Table 13.2 shows these two types of errors.

Table 13.2 Examples of Type I and Type II Errors

	State of Nature	
Decision	*H_0 is True (S_1) (Product fails)*	*H_0 is False (S_2) (Product succeeds)*
Do not introduce the product (A_1)	Correct decision	Type II error (β)
Introduce the product (A_2)	Type I error (α)	Correct decision

It should be noted that α and β errors would depend on how we select H_0, H_1, A_1 and A_2. Conventionally speaking, α is taken as the one that we would like to avoid in particular and we would formulate our hypotheses accordingly. In the above example too, the firm would like to avoid the α error as it is more serious than the β error. It should be obvious that the firm would have to sustain heavy losses if it introduces the product that fails. This is more serious than the other alternative before the firm of not introducing the product though it would have succeeded had it introduced. In this case, the firm would just miss an opportunity of increasing its profitability but it would not incur any loss.

13.4 TAILS OF A TEST

In Statistics, the rejection region in hypothesis testing can be on both sides of the curve with the non-rejection region in between the two rejection regions. A hypothesis test with two rejection regions is called a *two-tail* test and a test with one rejection region is called a *one-tail* test. The one rejection region can be on either side—right or left. If the rejection region is on the right side of the curve, then the test is known as the right-tail test. When the test has a rejection region on the left side, then it is known as the left-tail test.

Now, the question is: how to find out that a particular test is a two-tail or a one-tail test and if it is a one-tail test then whether it is a right-tail or a left-tail test. We can find out this with the help of the sign in the alternative hypothesis. If the alternative hypothesis has a '$\neq$' (not equal) sign, it is a two-tail test. For example, we may say that the average monthly income of households in a certain town is Rs 5,000. We may set up the two hypotheses as follows:

$H_0 : \mu = \text{Rs } 5{,}000$
$H_1 : \mu \neq \text{Rs } 5{,}000$

This is shown in Fig. 13.1.

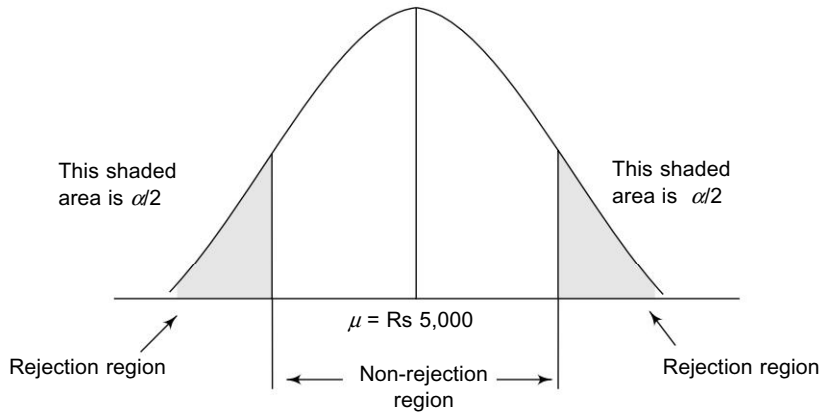


Fig. 13.1 Example of a Two-tail Test

The two values 2 are called the critical values.

One-tail Test

Let us now take an example to explain a one-tail test. Suppose we have undertaken a survey of a certain section of a city to ascertain their consumption of Coca-Cola. We hold the view that the average

weekly consumption of Coca-Cola per family is 10 litres. With this information, we may set up a left-tail test as follows:

$H_0 : \mu = 10$ litres

$H_1 : \mu < 10$ litres

Here , the alternative hypothesis has a '<' sign indicating 'less than' the quantity stated in the null hypothesis. Figure 13.2 shows a left-tail test.

In the left-tail test, the rejection region is always in the left-tail of the curve.

Let us continue with the earlier example. We retain the null hypothesis as it was formulated earlier. But, now we say that the average weekly consumption of Coca-Cola is more than 10 litres. This is shown in Fig. 13.3.

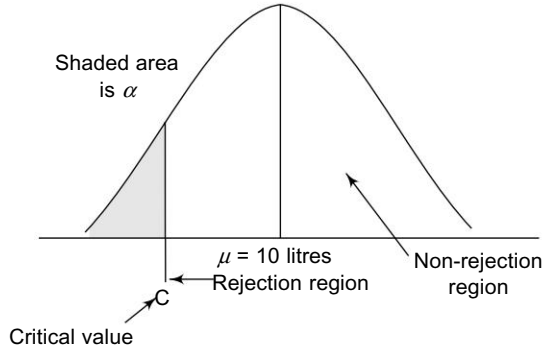


Fig. 13.2 **Example of a Left-tail Test**

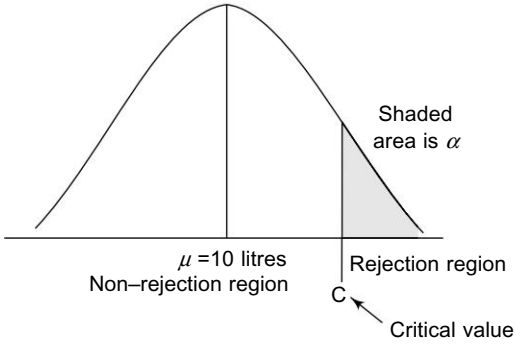


Fig. 13.3 **Example of a Right-tail Test**

It will be seen that when we set up our hypotheses as

$H_0 : \mu = 10$ litres

$H_1 : \mu > 10$ litres,

the alternative hypothesis has a '>' sign indicating 'greater than' the quantity stated in the null hypothesis. When this sign is found in the alternative hypothesis, then the hypothesis test is a right-tail test. We may now summarise this discussion in the form of Table 13.3.

Table 13.3 **Signs in the Tails of a Test**

	Two-tail Test	*Left-tail Test*	*Right-tail Test*
1. Sign in the null hypothesis H_0	=	= or ≥	= or ≤
2. Sign in the alternative hypothesis H_1	≠	<	>
3. Rejection region	In both tails	In the left tail	In the right tail

Table 13.4 gives critical values of Z for both one-tail and two-tail tests at different levels of significance. The critical values of Z as given here are commonly used. One can find critical values of Z for other values from the table of standard normal distribution.

Table 13.4 Critical Values of Z for Selected Levels of Significance

Levels of Significance, α	*0.10*	*0.05*	*0.01*	*0.005*	*0.002*
Critical values of *Z* for one-tail test	–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 values of *Z* for two-tail test	–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

13.5 HYPOTHESIS TEST ABOUT A POPULATION MEAN: LARGE SAMPLES

We may now discuss hypothesis tests concerning a population mean where sample size is large.

Example 13.1 A company manufacturing automobile tyres finds that tyre-life is normally distributed with a mean of 40,000 km and standard deviation of 3,000 km. It is believed that a change in the production process will result in a better product and the company has developed a new tyre. A sample of 100 new tyres has been selected. The company has found that the mean life of these new tyres is 40,900 km. Can it be concluded that the new tyre is significantly better than the old one, using the significance level of 0.01?

Solution In a problem of this type, we are interested in testing whether or not there has been an increase in the mean life of tyres. In other words, we would like to test whether the mean life of new tyres has increased beyond 40,000 km.

The various steps in testing the hypothesis are outlined below:

1. Null hypothesis and alternative hypothesis are:

   $H_0 : \mu = 40{,}000$ km

   $H_1 : \mu > 40{,}000$ km

2. The significance level is 0.01. That is, in 1 out of every 100 occasions, there is a risk of being wrong in accepting or rejecting the hypothesis.
3. The test criterion is the *Z*-test.
4. Computations: Substituting the value of standard deviation $\sigma = 3{,}000$ km in the formula

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \sigma_{\bar{x}} = \frac{3{,}000}{\sqrt{100}} = 300$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{40{,}900 - 40{,}000}{300} = 3$$

5. At 0.01 level of significance, the critical (table) value of z is ±2.33. As can be seen from Fig. 13.4, as the computed value of $z = 3$ falls in the rejection region, we reject the null hypothesis that $\mu = 40{,}000$ km. That is, the alternative hypothesis that μ is greater than 40,000 km is accepted. We, therefore, conclude that the new tyre is significantly better than the old one.

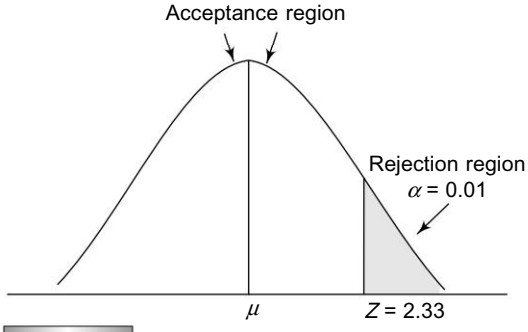


Fig. 13.4 Hypothesis Test Concerning New Tyres

Example 13.2 An insurance agent has claimed that the average age of policyholders who insure through him is less than the average for all agents, which is 35 years. A random sample of 40 policyholders who have insured through him gave an average of 32 years with a standard error of 2 years. Using α at 5 per cent level of significance, ascertain whether the insurance agent's claim is justifiable.

Solution The null hypothesis, $H_0 : \mu = 35$ years and the alternative hypothesis, $H_1 : \mu < 35$ years.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{32 - 35}{2} = -1.5$$

We are interested in testing whether or not the insurance agent's claim of average age of policyholders, who insure through him, is justified. Thus, it is a left-tail test as the alternative hypothesis $\mu < 35$ years.

At $\alpha = 0.05$ level of significance, the critical value of z is –1.64 for a one-tail test. The computed value of $z = -1.5$ falls in the acceptance region, as shown in Fig. 13.5. Thus, we accept the null hypothesis and conclude that the claim made by the insurance agent is not justifiable.

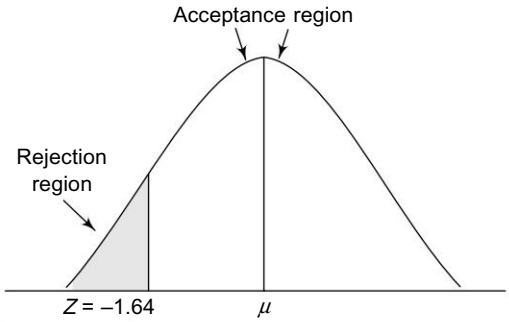


Fig. 13.5 **Hypothesis Test Concerning Insurance Agent's Claim**

Example 13.3 A company is engaged in the packaging of a superior quality tea in jars of 500 gm each. The company is of the view that as long as jars contain 500 gm of tea, the process is in control. The standard deviation is 50 gm. A sample of 225 jars is taken at random and the sample average is found to be 510 gm. Has the process gone out of control?

Solution To begin with, we assume that the process has not gone out of control, that is, the null hypothesis is that the population mean is 500 gm ($\mu = 500$). The alternative hypothesis is that the population mean is not 500 gm, that is, $\mu \neq 500$. This implies that the population mean can be either more than or less than 500 gm. It is assumed that the company considers it equally important to avoid overfilling or under-filling a jar. Such a hypothesis involves a two-tail test. Symbolically,

$H_0 : \mu = 500$ gm
$H_1 : \mu \neq 500$ gm

In order to solve this problem, we have to assume a level of significance. Let us assume that $\alpha = 5$ per cent. This is shown in Fig. 13.6. The calculation for the test is as follows:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{510 - 500}{50/\sqrt{225}} = \frac{10 \times 15}{50} = 3$$

Figure 13.6 shows the significance level of 0.05 as the two shaded regions, each containing 0.025 of the area. The 0.95 acceptance region contains two equal area of 0.475 each. From Table 13.4, we find that the appropriate Z value 0.05 level of significance for a two-tail test is 1.96. As the calculated value of $Z = 3$ is higher than 1.96, the null hypothesis is rejected. This means that the process is not in control.

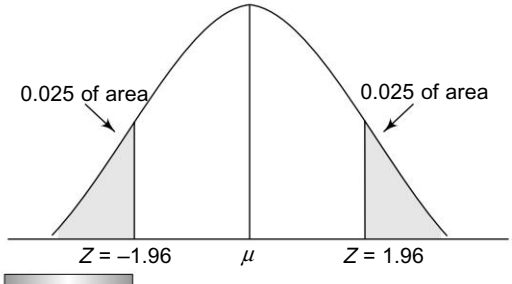


Fig. 13.6 **Hypothesis Test Concerning Tea Jars**

More frequently, however, we may be interested only in extreme values to one side of the mean. For example, when we have introduced a new process, we are hoping that it is better than the earlier one. In such a case, the test will be a right-tail test, the critical region being the area equal to the level of significance.

13.6 THE POWER OF STATISTICAL TEST

We have earlier discussed Type I error and Type II error. These errors are denoted by α and β, respectively. The question is: how to know that our test of a hypothesis has been good? We can measure the goodness of a test by the probabilities of making a Type I error or a Type II error.

The power of a statistical test, given as $1 - \beta$ = P (reject H_0 when H_0 is false), measures the ability of the test to perform as required. This $1 - \beta$ is called the *power of the function*. This means that greater the power of the function the better would be the decision rule.

Suppose that the null hypothesis is false. In such a case, we would like to reject our null hypothesis. Whenever null hypothesis is false, each time the hypothesis needs to be rejected. *But, as hypothesis test cannot always be foolproof, at times our calculations may suggest that a null hypothesis should not be rejected though it is false*. This means that μ, the true population mean, does not equal μ_{H_0}, the hypothesised population mean. This implies that μ equals some other value. For each possible value of μ where H_1 (the alternative hypothesis) is true, there is a different β, the probability of accepting a null hypothesis when it is false. All the same, our endeavour should be to minimise β. In other words, we should aim at $1 - \beta$ to be as large as possible.

The foregoing explanation suggests that a low value of $1 - \beta$ (a value very close to zero) indicates that our hypothesis test is working poorly. In contrast, when we have a high value of $1 - \beta$ (a value very close to 1), we can be sure that our hypothesis test is working quite well. If the different values of $1 - \beta$ corresponding to the values of μ for which the alternative hypothesis is true, are plotted on a graph, then the curve thus emerging is known as a *power curve*.

13.7 HYPOTHESIS TEST ABOUT A POPULATION MEAN: SMALL SAMPLES

The *Z*-test, used earlier, is based on the assumption that the sampling distribution of the mean is a normal distribution. This is applicable when the sample is large, that is, more than 30. When a sample is small, the assumption of normal distribution does not hold good and, as such, the *Z*-test will not be appropriate. Instead, another test known as the *t*-test is used. The procedure of testing the hypothesis is the same except that instead of the *Z*-value the *t*-value is used.

A few examples on the use of the *t*-test in hypothesis testing are given below:

Example 13.4 A manufacturer of electric batteries claims that the average capacity of a certain type of battery that the company produces is at least 140 ampere-hours with a standard deviation of 2.66 ampere-hours. An independent sample of 20 batteries gave a mean of 138.47 ampere-hours. Test at 5 per cent significance level the null hypothesis that the mean life is 140 ampere-hours against alternative that it is lower. Can the manufacturer's claim be sustained on the basis of this sample?

Solution In this problem, we are given a specific null hypothesis and an alternative hypothesis to be tested.

H_0 : The mean life of batteries is 140 ampere-hours.
H_1 : The mean life of batteries is < 140 ampere-hours.

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{138.47 - 140}{2.66/\sqrt{20}} = \frac{-1.53}{\frac{2.66}{4.47}} = -2.57$$

This t has a t distribution with $n - 1 = 20 - 1 = 19$ degrees of freedom. Taking α, the risk of Type I error at 5 per cent for 19 degrees of freedom, the critical value of t is ± 1.729 (Appendix Table 2). The calculated value of t being less than the critical value, falls within the rejection region. We, therefore, reject the null hypothesis and conclude that the sample mean is less than 140 ampere-hours.

Example 13.5 A company is engaged in the manufacture of car tyres. Their mean life is 42,000 km with a standard deviation of 3,000 km. A change in the production process is believed to improve the quality of tyres. A test sample of 28 new tyres has a mean life of 43,500 km. Do you think that the new car tyres are significantly superior to the earlier one? Test the hypothesis at 5 per cent level of significance.

Solution Null hypothesis, $H_0 : \mu \leq 42{,}000$ km
Alternative hypothesis, $H_1 : \mu > 42{,}000$ km
As the sample is small, we use the t-statistic to test the hypothesis.

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{43{,}500 - 42{,}000}{3{,}000/\sqrt{28}} = \frac{1{,}500}{567.11} = 2.64$$

Level of significance: $\alpha = 0.5$

Decision: The critical value of t at $\alpha = 0.5$ level of significance for $28 - 1 = 27$ degrees of freedom is 1.703. Since the computed value of t is greater than the critical value, the null hypothesis is rejected. In other words, the company's claim in respect of its new tyres is quite valid.

13.8 HYPOTHESIS TEST CONCERNING PROPORTION

Having discussed hypothesis test about a population mean, we now turn to hypothesis test concerning proportion, in this section.

Example 13.6 A pharmaceutical company, engaged in the manufacture of a patent medicine claimed that it was 80 per cent effective in relieving an allergy for a period of 15 hours. A sample of 200 persons, who suffered from allergy, were given this medicine. It was found that the medicine provided relief to 150 persons for at least 12 hours. Do you think that the company's claim is justified? Use 0.05 level of significance.

Solution Let p denote the proportion of getting relief from the allergy by using the medicine. We now set up the two hypotheses:

$H_0 : p = 0.8$, and the company's claim is justified
$H_1 : p < 0.8$, and the company's claim is false.

It may be noted that we have decided to have a one-tail test as we are interested in ascertaining whether the proportion of persons relieved by the medicine is lower than the proportion specified.

The sample proportion $\bar{p} = 150/200 = 0.75$. Using the Z-test, we find that

$$Z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.75 - 0.80}{\sqrt{\frac{0.8\,(1-0.8)}{200}}} = -\frac{0.05}{\sqrt{\frac{0.16}{200}}} = -1.786$$

The critical value of Z for one tail-test at 0.05 level of significance is –1.645. Since the calculated Z is –1.786, which is less than the critical value, it lies in the rejection region. We, therefore, conclude that the null hypothesis is rejected and, as such, the claim of the company that its patent medicine is 80 per cent effective is not justified.

Example 13.7 It is known from the past data that 10 per cent of the families in a certain locality subscribe to a periodical called *Outlook*. Of late, there has been some apprehension that this subscription rate has declined. In order to test whether or not there has been a decline, a random sample of 100 families is chosen. It is found that the sample proportion $\bar{p}$ is 0.07, that is, 7 per cent. Can it be concluded that the subscription rate has really declined, assuming a 5 per cent level of significance?

Solution As we are interested to test whether or not there has been a decline in the subscription rate, let us assume that there has been no decline. The null and alternative hypothesis are:

$H_0 : p = 10$ per cent

$H_1 : p < 10$ per cent.

The sample proportion is $\bar{p} = 0.07$. Using the Z-test, we find that

$$Z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.07 - 0.10}{\sqrt{\frac{(0.10)(0.90)}{100}}} = \frac{-0.03}{\sqrt{\frac{0.09}{100}}} = -1$$

The level of significance is 0.05, which means that $Z = \pm 1.64$ for a one-tail test. Further, as $Z = -1$ falls within the acceptance level as shown in Fig. 13.7, we accept the hypothesis that the subscription rate has not dropped. The management need not be unnecessarily alarmed and it may be concluded that the same subscription rate continues as before.

This example is related to a left-tail test. We now take up another example to illustrate a right-tail test.

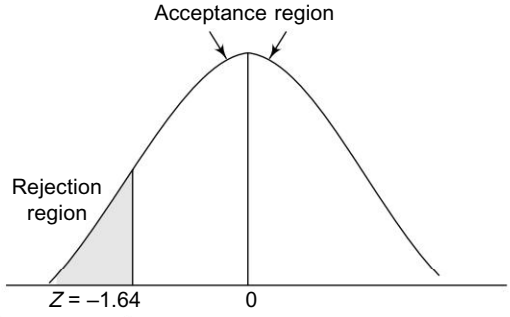


Fig. 13.7 **Hypothesis Test Concerning the Subscription to *Outlook***

Example 13.8 In 1983, 15 per cent of households in a certain city indicated that they owned a sewing machine. In 1985, there was a reason to believe that there was some increase in its per centage. A survey based on random sample of 900 households was taken and it was found that 189 households had a sewing machine. Can we conclude that there has been a significant increase in the sale of sewing machines, assuming 0.05 level of significance?

Solution As we are interested to test whether or not there has been a significant increase in the sale of sewing machines, we set up a null and alternative hypothesis as follows:

$H_0 : p = 15$ per cent

$H_1 : p > 15$ per cent.

This is a right-tail test. The calculations are as follows:

$$Z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.21 - 0.15}{\sqrt{\frac{0.15\,(1-0.15)}{900}}} = \frac{0.06}{\sqrt{\frac{0.1275}{900}}} = 5.04$$

As the value of Z at $\alpha = 0.05$ is 1.64 for a one-tail test, the calculated value of Z falls in the rejection region. We, therefore, reject the null hypothesis and conclude that there has been a significant increase in the sale of sewing machines in 1985 as compared to 1983. This is also revealed by Fig. 13.8.

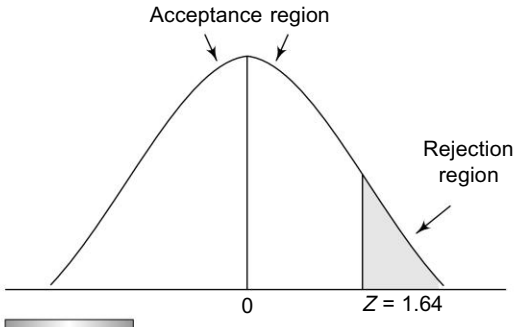


Fig. 13.8 **Hypothesis Test Concerning the Sale of Sewing Machines**

The preceding two examples were on one-tail test. Let us take another example where we have to use a two-tail test.

Example 13.9 A company engaged in the manufacture of superior quality diaries, which are primarily meant for senior executives in the corporate world. It claims that 75 per cent of the executives employed in Delhi use its diaries. A random sample of 800 executives was taken and it was found that 570 executives did use its diary when the survey was undertaken. Verify the company's claim, using 5 per cent level of significance.

Solution As we are interested in verifying the company's claim, it has to be a two-tail test and we set up the null and alternative hypothesis as follows:

$H_0 : p = 75$ per cent

$H_1 : p \neq 75$ per cent

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - P}{\sqrt{\frac{P(1-P)}{n}}} \quad \text{Here} \quad \bar{p} = 570/800 = 0.7125$$

$$= \frac{0.7125 - 0.75}{\sqrt{\frac{0.75\,(1-0.75)}{800}}} = \frac{-0.0375}{\sqrt{\frac{0.19}{800}}} = -2.45$$

As the calculated value of $z = -2.45$ is less than the critical value of $Z = -1.96$, it falls in the rejection region. This is shown in Fig. 13.9. We, therefore, reject the null hypothesis and conclude that the claim of the company is exaggerated and is not supported by our statistical test.

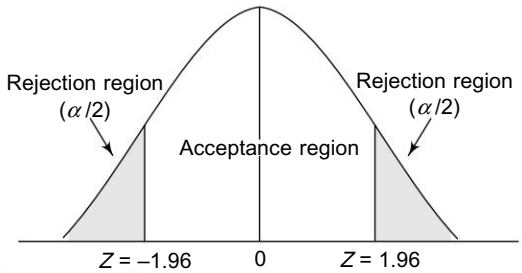


Fig. 13.9 **Hypothesis Test Concerning Diaries for Executives**

It may be noted that as the level of significance is 0.05 and as the test is a two–tail test, each shaded area in Fig. 13.9 is 0.05/2 = 0.025, being equal on both sides. Then the area between 0 and $Z_1 = 0.5000 - 0.0250 = 0.4750$ and $Z = 1.96$ as taken from the Appendix Table 1. (This is also given in Table 13.4.)

13.9 HYPOTHESIS TEST CONCERNING THE DIFFERENCES BETWEEN TWO POPULATION MEANS

So far our hypothesis tests have been concerned with individual means and proportions—often referred to as *one-sample test*. However, it may sometimes be claimed that there is no difference between the means or proportions of two groups. In such a case, we need samples from each group: *two-sample tests*.

The procedure for testing the hypothesis is similar to that used in one-sample tests. Here, we have two populations and our concern is to test the claim as to a difference in their sample means. For example, a company may claim that there is no difference between the average salaries of its male and female executives. Thus, we have average salaries of males (μ_1) and females (μ_2). We take random samples of size n_1 and n_2 and determine their sample means ($\bar{x}_1$) and ($\bar{x}_2$) along with sample standard deviations (s_1) and (s_2). In order to test the hypothesis, we require the standard error (SE) of the sampling distribution of the difference between two means, which is divided into the difference between the population means subtracted from the difference between the sample means.

When n > 30, the Z statistic takes the following form:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (\text{when } H_0 : \mu_1 = \mu_2)$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

where $\bar{x}_1 - \bar{x}_2$ is the difference between two sample means, $\mu_1 - \mu_2$ is the difference between two population means, σ_1^2 and σ_2^2 are the variances of population 1 and population 2, respectively and n_1 and n_2 are sample sizes from population 1 and population 2, respectively.

Let us take an example to explain this.

Example 13.10 A potential buyer wants to decide which of the two brands of electric bulbs he should buy as he has to buy them in bulk. As a specimen, he buys 100 bulbs of each of the two brands—A and B. On using these bulbs, he finds that brand A has a mean life of 1,200 hours with a standard deviation of 50 hours and brand B has a mean life of 1,150 hours with a standard deviation of 40 hours. Do the two brands differ significantly in quality? Use $\alpha = 0.05$.

Solution Let us set up the null hypothesis that the two brands do not differ significantly in quality:

$H_0 : \mu_1 = \mu_2$ and an alternative hypothesis:

$H_1 : \mu_1 \neq \mu_2$

where μ_1 = mean life of brand A bulbs, and μ_2 = mean life of brand B bulbs.

We now construct the Z statistic.

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{1,200 - 1,150}{\sqrt{\frac{(50)^2}{100} + \frac{(40)^2}{100}}}$$

$$= \frac{50}{\sqrt{25 + 16}} = 7.81 \text{ approx.}$$

As we are given $\alpha = 0.05$, the value of Z for a two–tail test is ± 1.96. As the calculated value of Z (7.81) falls in the rejection region, we reject the null hypothesis and, therefore, conclude that the bulbs of two brands differ significantly in quality.

When the sample is small ($n < 30$), the t–test is used instead of Z test. This will be clear from the example given below:

Example 13.11 Two salesmen, A and B, are employed by a company. Recently, it conducted a sample survey yielding the following data:

	Salesman A	*Salesman B*
Number of sales	20	22
Average weekly sales (Rs lakh)	30	25
Standard deviation (Rs lakh)	10	7

Is there any significant difference between the average sales of the two salesmen?

Solution We first set up the null hypothesis that there is no significant difference between the average sales of the two salesmen. That is,

$H_0 : \mu_1 = \mu_2$ and an alternative hypothesis:

$H_1 : \mu_1 \neq \mu_2$

The t statistic is calculated as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

However, we can use this formula when we know $\hat{\sigma}$. Hence, we first calculate $\hat{\sigma}$ by applying the following formula:

$$= \hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(20)(10)^2 + (22)(7)^2}{20 + 22 - 2} = \frac{2,000 + 1,078}{40}$$

$$= 76.95$$

$$\hat{\sigma} = \sqrt{76.95} = 8.77$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{30 - 25}{8.77} \sqrt{\frac{20 \times 22}{20 + 22}}$$

$$= 0.57 \times \sqrt{10.4762} = 0.57 \times 3.23669 = 1.84 \text{ approx.}$$

At $\alpha = 0.05$ level of significance, the critical value of t for (20 + 22 – 2) that is, 40 degrees of freedom for a two-tail test is 2.021 (Appendix Table 2). As our calculated t-value, being 1.84, is less than the critical value, it falls in the acceptance region. Thus, the null hypothesis is accepted. In other words, the average sales by the two salesmen are not significantly different.

The above examples are based on observations randomly drawn from two independent populations. At times, it is useful to set up the hypothesis differently and compare the effects of two different 'treatments' applied to the same objects or persons. Such a test is known as a 'before' and 'after' type test. For example, a company may be interested to ascertain whether there is a significant difference in the sale of its product as a result of introducing the new packaging. Another example may relate to the reorganisation of its sales department and then to examine the change in its overall sales.

An example with the necessary calculations is given below:

Example 13.12 A company has reorganised its sales department. The following data show its weekly sales both before and after reorganisation. The period for comparison is taken from January to March in two successive years.

	Week No.									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Sales prior to reorganisation (Rs lakh)	12	15	13	11	17	15	10	11	18	19
Sales after reorganisation (Rs lakh)	16	17	14	13	15	14	12	11	17	22

Can it be concluded that the reorganisation of the sales department of the company has resulted in a significant increase in its sales?

Solution We set up the null hypothesis:

H_0 : the reorganisation of the sales department has not resulted in improved sales.

H_1 : the reorganisation of the sales department has resulted in improved sales.

Since the observations are paired together, the paired t-test may be applied. For this purpose, the t statistic is $t = \bar{d} / \sqrt{(s^2 / n)}$, where $d = X_2 - X_1$.

The computations are shown in Table 13.5.

Table 13.5 Worksheet for the Paired *t*–Test

Week No.	*Sales before Re-organisation (Rs lakh)* X_1	*Sales after Re-organisation (Rs lakh)* X_2	*Deviations* $d = X_2 - X_1$	*Deviations Square* $d^2 = (X_2 - X_1)^2$
1	12	16	4	16
2	15	17	2	4
3	13	14	1	1
4	11	13	2	4
5	17	15	–2	4
6	15	14	–1	1
7	10	12	2	4
8	11	11	0	0
9	18	17	–1	1
10	19	22	3	9
		Total	**10**	**44**

$$\bar{d} = \frac{\Sigma d}{n} = \frac{10}{10} = 1$$

$$s^2 = \frac{1}{10-1}\left(44 - \frac{(10)^2}{10}\right)$$

$$= \frac{1}{9}\left(\frac{440-100}{10}\right) = 3.78$$

$$t = \bar{d}/\sqrt{(s^2/n)} = 1/\sqrt{(3.78/10)} = 1/0.61 = 1.64 \text{ approx.}$$

The critical value of t for 10 – 1 = 9 degrees of freedom at 5 per cent level of significance is 1.83 (one-tail test). As the calculated t value is less than the critical value, it falls in the acceptance region. We, therefore, accept the null hypothesis and conclude that the reorganisation of the sales department did not have significant increase in the company's sales.

It may be noted that when we are dealing with two independent samples, these two samples may not be random in true sense. In that case, the hypothesis test will not reflect properly the effect of training in the subsequent period. Besides, there is a possibility of poor workers being reluctant to participate in the training programme or the management might have selected better workers for training, which may be taken as some kind of reward for their good work. Even efficient workers might have taken an initiative and volunteered themselves to undergo the training programme. While all these possibilities are there in a well-designed training programme, there should be no scope for these possibilities to happen otherwise our results will be misleading.

13.10 HYPOTHESIS TESTS OF DIFFERENCES BETWEEN TWO PROPORTIONS

At times we come across certain problems where we are interested to know whether or not there exist significant differences between the proportions of two groups of, say, consumer in respect of a certain activity. Such problems may relate to the difference in the proportion of male and female employees working with the company for, say, five years or more, or difference in proportion of young and old consumers satisfied with a product. Again, we may like to know if there is a difference between consumers in regard to their taste for a particular product.

Let us take an example to explain the procedure to be used in hypothesis testing in such cases.

Example 13.13 You obtain a large number of components to an identical specification from two sources. You may notice that some of the components are from the supplier's own plant in Pune and some are from the plant located in Bangalore. You would like to know whether the proportions of defective components are the same or there is a difference between the two. You take a random sample of 600 components from each plant and find that the rejection rate $\bar{p}_1$ is 0.015 for Pune components as compared to $\bar{p}_2 = 0.017$ for Bangalore components. Set up the null hypothesis and test it at 5 per cent level of significance.

Solution We set up the two hypotheses:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

where p_1 and p_2 are the proportions of defective components from Pune and Bangalore, respectively. This is a two-tail test.

Level of significance $\alpha = 0.05$, both sample sizes are large.

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{600(0.015) + 600(0.017)}{600 + 600}$$

$$= \frac{9 + 10.2}{1,200} = 0.016$$

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}} = \frac{0.015 - 0.017 - 0}{\sqrt{0.016(1 - 0.016)(1/600 + 1/600)}}$$

$$= \frac{-0.002}{\sqrt{(0.016 \times 0.984)(1/300)}} = \frac{-0.002}{\sqrt{0.005248}}$$

$$= -0.0276$$

Since the calculated value of Z is –0.0276 as against the critical value of Z at 5 per cent level of significance for a two-tail test is ± 1.96, it does not fall in the rejection region. As such, our conclusion is that there is no difference in the rejection rates of components from Pune and Bangalore plants.

Figure 13.10 illustrates the acceptance and rejection regions pertaining to this problem.

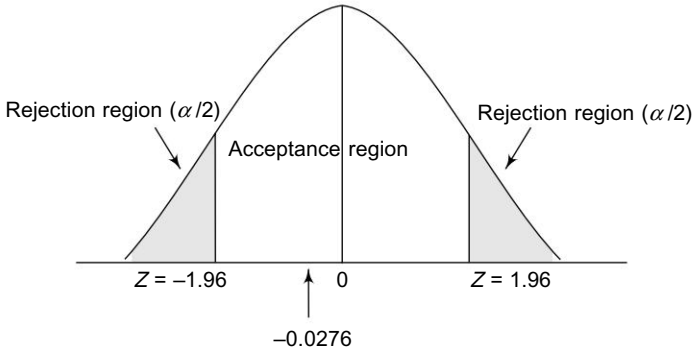


Fig. 13.10 Hypothesis Test Concerning Difference in Proportions of Defective Components

F-Distribution

In order to test for equality of two population variances, we have to use the *F*-test. It is necessary to know what the *F*-distribution is. The *F*-distribution is named after the English statistician Sir Ronald A Fisher.

As in the case with *t* and chi-square distributions, the shape of a particular *F*-distribution curve depends on the number of degrees of freedom. A major difference between *t* and chi-square distributions and *F*-distribution is that the former two distributions have only one number of degrees of freedom whereas the latter has two numbers of degrees of freedom. This is because it has degrees of freedom for the numerator as also for the denominator. These two numbers of the degrees of freedom

are the parameters of the *F*-distribution. The two degrees of freedom taken together give a particular *F*-distribution curve. It may be noted that the *F*-distribution is also a continuous distribution like normal, *t* and chi-square distribution. It takes only positive values. The *F*-distribution curve is skewed to the right. But, as the number of degrees of freedom increases, the skewness decreases.

Figure 13.11 shows three *F*-distribution curves for three different sets of degrees of freedom.

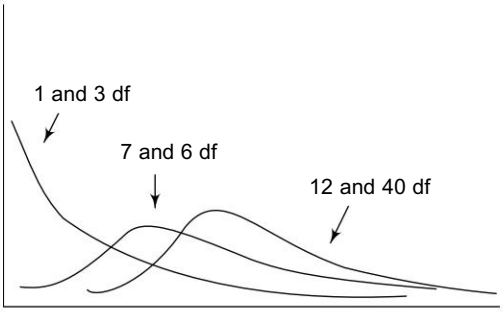


Fig. 13.11 Three *F*-Distribution Curves

It will be seen that the first set of degrees of freedom is 1 and 3, the second set, 7 and 6, and the third set is 12 and 40. The first number gives the degrees of freedom associated with the numerator while the second number gives the degrees of freedom associated with the denominator. Figure 13.11 also shows that when the degrees of freedom are small, the *F*-distribution curve is highly skewed. However, as the degrees of freedom increase, the skewness decreases.

Appendix Table 6 shows the values of *F* for the *F*-distribution. In order to use this table, we need to know three quantities, viz. (i) the degrees of freedom for the numerator, (ii) the degrees of freedom for the denominator, and (iii) an area in the right tail of an *F*-distribution curve. It may be noted that the *F*-distribution table is read only for an area in the right tail of the *F*-distribution curve. Further, the Appendix Table 6 has three parts, giving the *f* values for an area of 0.01 and 0.05, respectively, in the right tail of the *F*-distribution curve.

13.11 *F*-TEST FOR DIFFERENCES IN TWO VARIANCES

At times it is necessary to ascertain whether the two independent populations have the same variability. For this purpose, we use the *F*-test, which measures the ratio of the two sample variances. If each population is assumed to be normally distributed, then the ratio s_1^2/s_2^2 follows the *F*-distribution. The crucial values of the *F*-distribution are given in Appendix Table 6 on two sets of degrees of freedom. The degree of freedom in the numerator of the ratio pertain to the first sample, and the degrees of freedom in the denominator pertain to the second sample.

The F-test statistic for testing the equality of two variances is given below:

$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 = *variance of sample* 1

s_2^2 = *variance of sample* 2

The test statistic F follows an F-distribution with $n_1 - 1$ *and* $n_2 - 1$ *degrees of freedom.*

For a given level of significance α, we set up the null hypothesis of equality of variances as

$$H_0 : \sigma_1^2 = \sigma_2^2$$

against the alternative hypothesis that the two population variances are not equal

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

The null hypothesis is rejected if the computed F-test statistic is greater than the critical value of F. The critical value of F is found from Appendix Table 6 on the basis of $n_1 - 1$ degrees of freedom from sample 1 in the numerator and $n_2 - 1$ degrees of freedom from sample 2 in the denominator.

Let us take an example.

Example 13.14 Suppose a company manufacturing light bulbs is using two different processes A and B. The life of the light bulbs of process A has a normal distribution with mean μ_1 and standard deviation σ_1. Similarly, for process B, it is μ_2 and σ_2. The data pertaining to the two processes are given below:

Sample A	*Sample B*
$n_1 = 16$	$n_2 = 21$
$\bar{x}_1 = 1200$ hr	$\bar{x}_2 = 1300$ hr
$\sigma_1 = 60$ hr	$\sigma_2 = 50$ hr

We have to test the hypothesis that the variability of the two processes is the same.

Solution

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

We test the null hypothesis with the level of significance $\alpha = 0.05$.

For this, we first calculate the F statistic:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

$$= \frac{n_1 s_1^2 / n_1 - 1}{n_2 s_2^2 / n_2 - 1}$$

$$= \frac{16(60)^2/16 - 1}{21(50)^2/21 - 1}$$

$$= \frac{57600/15}{52500/20}$$

$$= \frac{3840}{2625} = 1.46$$

Since this is a two-tail test, this F is compared against $F(15, 20)$ for $\alpha = 0.05$. The critical value of $F(15, 20)$ is 2.20. As this value is greater than the computed $F(15, 20)$ being 1.46, we accept the null hypothesis that $\sigma_1^2 = \sigma_2^2$ indicating that there is no significant difference in the variability of the two samples.

13.12 THE *P*-VALUE OF A TEST

So far our discussion on hypothesis testing shows that we have examined a hypothesis at a specified level of significance, say 5%. Now this 5% level of significance is arbitrarily fixed as we could have used any other level of significance. As such mere acceptance or rejection of a hypothesis fails to show the full strength of the sample evidence. An alternative is the use of *P*-value approach. The *P*-value is the observed level of significance, which is the smallest value at which H_0 can be rejected. Let us discuss this in some detail.

It may be noted that setting the value of α affects both type I and type II error probabilities. Hence, setting a value for α involves a compromise for which it is necessary for us to know the cost of each type of error. However, it is very difficult to estimate the cost of errors as they depend on the unknown actual value of the parameter being tested. In view of this major difficulty, it is better to follow an intuitive approach of assigning a value to α. Normally, one of the three values, viz. 10%, 5% and 1% are assigned to α. In this intuitive approach, we try to estimate the relative costs of the two types of errors. For example, suppose our company is engaged in manufacturing automobile tyres and we have to ensure that tyres must satisfy the minimum norm. Here type I error will result in rejecting a good batch of tyres while type II error will result in accepting a batch of poor quality tyres. In case type I error is more costly than the type II error, it is advisable to choose a small value for α, say 1%. On the other hand, if type II error happens to be more costly, it is advisable to choose a large value for α, say 10%. However, in several cases we may find it difficult to decide as to which type of error is more costly. When the costs of type I and type II errors are more or less equal, or when we do not know about their relative costs, then α may be taken at 5%.

P-value—Application

Now, coming to the application of *P*-value, the following two rules must be noted:

1. If the *P*-value is greater than or equal to α, H_0 is not rejected.
2. If the *P*-value is less than α, H_0 is rejected. Let us take an example to explain this.

Example 13.15 Suppose a company manufacturing automobile tyres finds that tyre-life is normally distributed with a mean of 40, 000 km and standard deviation of 3000 km. The company has developed a new tyre. On the basis of a sample of 100 new tyres, the company has found that the mean life of new tyres is 40,450 km. Can it be concluded that the new tyre is significantly better than the old one, using the significance level of 0.05?

$H_0 : \mu = \mu_0$ (= 40,000 km)

$H_1 : \mu \neq \mu_0$

As it is a two tail-test, the critical value of Z for 0.05 level of significance is 1.96.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3000}{\sqrt{100}} = 300$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{40,450 - 40,000}{300} = \frac{450}{300} = 1.5$$

As the calculated value of Z is less than the critical value of Z (1.96), null hypothesis H_0 is accepted.

We find from Appendix Table 1 that the probability of Z being 1.5 is 0.0668. This is the probability of obtaining a Z value below –1.50. The probability of obtaining a value below +1.50 is 1 – 0.9332 = 0.0668. Thus, the P-value for this two-tail test is 0.0668 + 0.0668 = 0.1336. Now, we find that the P-value is greater than $\alpha = 0.05$. As such, according to the first rule given above, the null hypothesis is not rejected.

In the above example, the observed sample mean is 40450 km, which is 450 km above the hypothesized value. Thus, if the population mean is 40,000 km, there is a 13.36 per cent chance of obtaining a sample mean more than 450 km away from 40,000 km (i.e., greater than or equal to 40,450 km or less than or equal to 39,550 km).

Additional Examples

Example 13.16 A strength test carried on sample of two yarns spun of the same count gives 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 the sources from which the samples are drawn at 1 per cent level of significance?

Solution Since the problem relates to small samples, we have to use the t-test.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

The t statistic is calculated by using the following formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

However, this formula can be used when we know $\hat{\sigma}$. Hence we first calculate $\hat{\sigma}$ by applying the following formula:

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(4-1)42 + (9-1)56}{4 + 9 - 2}$$

$$= \frac{(3 \times 42) + (8 \times 56)}{11}$$

$$= \frac{126 + 448}{11} = \frac{574}{11} = 52.18$$

$$t = \frac{52 - 42}{\sqrt{52.18}} \sqrt{\frac{4 \times 9}{4 + 9}}$$

$$= \frac{10}{\sqrt{52.18}}\sqrt{\frac{36}{13}}$$

$$= \frac{10}{\sqrt{52.18}} \times 1.664$$

$$= \frac{16.64}{\sqrt{52.18}} = \frac{16.64}{7.22} = 2.305$$

The critical value of t for 11 degrees of freedom for a two-tail test at 1 per cent level of significance is 2.718. As the calculated t-value is less than the critical value, H_0 is accepted. Hence, the difference in the mean strengths is not significant.

Example 13.17 Two urns, A and B, contain equal number of marbles, but the proportion of red and white marbles in each of the urns is unknown. A sample of 50 marbles selected with replacement from each of the urns revealed 32 red marbles from A and 23 red marbles from B. Using a significance level of 0.05, test the hypothesis that **(a)** the urns have equal proportions of red marbles and **(b)** A has a greater proportion of red marbles than B.

Solution

(a) H_0 : The urns have equal proportions of red marbles.

H_1 : The urns do not have equal proportions of red marbles.

Symbolically,

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

First, we have to calculate $\bar{p}$:

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \quad \bar{p}_1 = \frac{32}{50} = 0.64 \quad \bar{p}_2 = \frac{23}{50} = 0.46$$

$$= \frac{(50 \times 0.64) + (50 \times 0.46)}{50 + 50}$$

$$= \frac{32 + 23}{100} = \frac{55}{100} = 0.55$$

We now calculate Z statistic:

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.64 - 0.46}{\sqrt{0.55(1-0.55)\left(\frac{1}{50} + \frac{1}{50}\right)}}$$

$$= \frac{0.18}{\sqrt{(0.55 \times 0.45)(0.02 + 0.02)}}$$

$$= \frac{0.18}{\sqrt{0.2475 \times 0.04}}$$

$$= \frac{0.18}{\sqrt{0.0099}} = \frac{0.18}{0.099} = 1.818 \quad \text{or} \quad 1.82$$

The critical value of $Z_{0.05}$ for a two-tail test is 1.96. Since the calculated value of Z is less then the critical value, H_0 is accepted. This means that the two urns have equal proportions of red marbles.

(b) H_0 : Urn A has a greater proportion of red marbles.

H_1 : Urn A does not have a greater proportion of red marbles.

The critical value of $Z_{0.05}$ for a one-tail test is 1.645. As the calculated value of Z is greater than the critical value, H_0 is rejected. This means that urn A does not have a greater proportion of red marbles than urn B.

Example 13.18 A company is considering two different television advertisements for 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 per cent level of significance is used?

Solution

H_0 : There is no significant difference in the effectiveness of the two advertisements A and B

Symbolically, $H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

$$\bar{p}_1 = \frac{x_1}{n_1} = \frac{18}{60} = 0.30, \quad n_1 = 60$$

$$\bar{p}_2 = \frac{x_2}{n_2} = \frac{22}{100} = 0.22, \quad n_2 = 100$$

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{18 + 22}{60 + 100} = \frac{40}{160} = 0.25$$

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\hat{\sigma}_{\bar{p}_1 - \bar{p}_2}}$$

$$= \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.30 - 0.22 - 0}{\sqrt{0.25(1-0.25)\left(\frac{1}{60} + \frac{1}{100}\right)}}$$

$$= \frac{0.08}{\sqrt{(0.25 \times 0.75)\left(\frac{5+3}{300}\right)}}$$

$$= \frac{0.08}{\sqrt{(0.1875)(0.0267)}}$$

$$= \frac{0.08}{\sqrt{0.00500625}} = \frac{0.08}{0.0707} = 1.13$$

As the calculated value of $Z = 1.13$ is less than the critical value of $Z = 1.645$ at 5 per cent level of significance, the null hypothesis is accepted. The conclusion is that there is no significant difference in the effectiveness of the two advertisements.

Example 13.19 A product is produced in two ways. A pilot test of 64 items from each method indicates that the product of method 1 has a sample mean tensile strength of 106 pounds and a standard deviation of 12, whereas in method 2, the corresponding values are 100 and 10 pounds, respectively. Greater tensile strength is preferable. Use an appropriate large sample test at 5 per cent level of significance to test whether or not method 1 is better for processing the product.

Solution Let us set up the null hypothesis that the mean tensile strength in both the methods do not differ significantly. In other words,

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

$$= \frac{106 - 100 - 0}{\sqrt{\frac{(12)^2}{64} + \frac{(10)^2}{64}}}$$

$$= \frac{6}{\sqrt{\frac{144}{64} + \frac{100}{64}}}$$

$$= \frac{6}{\sqrt{\frac{244}{64}}} = \frac{6}{1.95} = 3.08$$

The critical value of $Z_{0.05}$ for one-tail test is 1.645. Since the calculated value is greater than the critical value of Z, H_0 is rejected. The conclusion is that method 1 is better than method 2.

Example 13.20 An auto company claims that the petrol consumption of the new car introduced by it is 9.5 km per litre. In an experiment with 50 cars, it was found that the mean petrol consumption was 10 km per litre, with a standard deviation of 3.5 km per litre. Test the validity of the claim of the car company using 5% level of significance.

Solution

H_0 : Consumption is 9.5 km per litre
H_1 : Consumption is more than 9.5 km per litre

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

$$= \frac{10 - 9.5}{3.5}$$

$$= \frac{0.5}{3.5} = 0.14$$

The critical value of Z @ 5% level of significance for one-tail test is 1.64. As the calculated value of $Z = 0.14$ is less than 1.64, H_0 is accepted.

Example 13.21 Examine the claim of a battery producer that the batteries will last for 100 days, given that a sample study about their life, of the batteries on 200 batteries, showed mean life of 90 days with a standard deviation of 15 days. Assume normal distribution, and test at 5% level of significance.

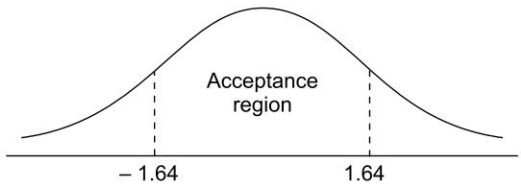
Solution

H_0 : Batteries will last for 100 days
H_1 : Batteries will last for less than 100 days

$$Z = \frac{\bar{X} - \mu}{\sigma}$$

$$= \frac{90 - 100}{15}$$

$$= \frac{-10}{15} = -0.67$$



At 5% level of significance, critical value of Z is –1.64. The calculated value of Z being $Z = –0.67$, it falls in the acceptance region. Hence, the claim of the battery producer is acceptable.

Example 13.22 A manufacturer claimed that at least 95% of the equipment, 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 of (a) 0.01, (b) 0.05.

Solution

H_0 : $p \geq 95$ per cent
H_1 : $p < 95$ per cent

The sample proportion is $\bar{p} = \frac{200 - 18}{2} = 91$ per cent

Using the Z-test, we find that

$$Z = \frac{\bar{p} - p}{\sigma_{\bar{p}}}$$

$$= \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.91 - 0.95}{\sqrt{\frac{0.95(1-0.95)}{200}}}$$

$$= \frac{-0.04}{\sqrt{\frac{0.0475}{200}}} = \frac{-0.04}{0.015} = -2.67$$

The critical value of Z for one-tail test at 0.01 level of significance is –2.33, and at 0.05 level of significance is –1.64. The calculated value of Z = –2.67 lies in the rejection region. As such, the manufacturer's claim at both levels is rejected.

Example 13.23 The personnel manager of a firm is recruiting its graduate trainees from colleges affiliated to two universities. The performance index (0 – 10 scale) of the trainees from each university follows normal distribution. The variance of performance index of trainees from University I is 9 and that of trainees from University II is 4. The manager feels that the mean performance index of trainees from University I is less than that of trainees from University II. To test his intuition, he has selected a sample of 49 trainees from University I, and their mean performance index is found to be 7. Similarly, he has selected a sample of 64 trainees from University II, and their mean performance index is found to be 6.

Test the intuition of the manager at a significant level of 0.05.

Solution

Let us assume that the mean performance index of students in the two universities is the same.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{7 - 6}{\sqrt{\frac{9}{49} + \frac{4}{64}}} = \frac{1}{\sqrt{\frac{576 + 196}{3136}}}$$

$$= \frac{1}{\sqrt{\frac{772}{3136}}}$$

$$= \frac{1}{0.496}$$
$$= 2.016$$

The critical value of Z @ 5% level of significance for one-tail test is 1.64. As the calculated value of $Z = 2.016 > 1.64$, H_0 is rejected. The manager's intuition, that the mean performance index of trainees from University I is less, is valid.

Example 13.24 Two urns, A and B, contain equal number of marbles, but the proportion of red and white marbles in each of the urns is unknown. A sample of 50 marbles, selected with replacement from each of the urns, revealed 32 red marbles from A and 23 red marbles from B. Test the hypothesis that (a) the two urns have equal proportion of red marbles, and (b) A has greater proportion of red marbles than B. Use a significance level of 0.05.

Solution

(a) $H_0 : p_A = p_B$
$H_1 : p_A \neq p_B$

$$\bar{p} = \frac{n_A \bar{p}_A + n_B \bar{p}_B}{n_A + n_B}$$
$$= \frac{50\,(0.64) + 50\,(0.46)}{50 + 50}$$
$$= \frac{32 + 23}{100} = \frac{55}{100} = 0.55$$
$$Z = \frac{(\bar{p}_A - \bar{p}_B) - (p_A - p_B)}{\sqrt{\bar{p}\,(1-\bar{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{0.64 - 0.46 - 0}{\sqrt{0.55\,(1-0.55)\left(\frac{1}{50} + \frac{1}{50}\right)}}$$
$$= \frac{0.18}{\sqrt{(0.2475)(0.04)}} = \frac{0.18}{\sqrt{0.0099}} = \frac{0.18}{0.099} = 1.82$$

Since it is a two-tail test, the critical value of $Z_{0.05}$ is 1.96. As it is > calculated $Z(1.82)$, H_0 cannot be rejected.

(b) $H_0 : p_A$ is not greater than p_B.
$H_1 : p_A > p_B$

Since it is a one-tail test, the critical value of Z at 0.05 level of significance is 1.64. As the calculated value of Z is greater than the critical value, H_0 is rejected. In other words, urn A has a greater proportion of red marbles than urn B.

13.13 COMMENTS ON THE THEORY OF HYPOTHESIS TESTS

We have seen that there is a clear-cut procedure laid down for testing a null hypothesis with measured risk α and β. Unfortunately, the theoretical framework does not help us at times in practical situations.

The basic problem in testing hypotheses requires that the experimenter is able to specify an alternative hypothesis, which allows the calculation of the probability β of a Type II error for all alternative

values of the parameter(s). However, there may be certain situations when it is difficult to specify clearly alternatives to null hypothesis, H_0, that have practical importance. All the same, this difficulty does not undermine the utility of hypothesis tests. In a way, it suggests that we have to be very careful in rejecting a null hypothesis on the basis of insufficient information. If we are convinced that the data are insufficient, we should prefer to get additional data and then apply the test.

We should also note that our choice in favour of a one-tail or a two-tail test will depend on the alternative value of the parameter that we are trying to find. Suppose that our parameter of interest is population mean μ. In case we were to incur heavy financial loss if μ_1 were greater than μ_0 but not if it were less, then we would focus our attention on the detection of values of μ_1 greater than μ_0. In that case we would use a right-tail test to reject the hypothesis. In case we are interested in detecting value of μ_1, which is either greater than or less than μ_0, then we would use a two-tail test. By applying such a reasoning, we can use an appropriate test—right-tail or left-tail or two-tail test.

Another point to note is that sometimes we may find that the assumptions upon which the test is based are not valid. In such a case, it would be wrong to use the test discussed in this chapter. In order to overcome this problem, we may use an appropriate non–parametric test. Such tests are known as distribution-free tests and have a few assumptions, if any. We shall discuss a major non-parametric test (chi-square) in Chapter 15 and a number of other non-parametric tests in Chapter 20.

GLOSSARY

Alpha (α)	The significance level of a test of hypothesis that denotes the probability of rejecting a null hypothesis when it is actually true. In other words, it is the probability of committing a Type I error.
Alternative hypothesis	A hypothesis that takes a value of a population parameter different from that used in the null hypothesis.
Beta (β)	The probability of not rejecting a null hypothesis when it actually is false. In other words, it is the probability of committing a Type II error.
Critical region	The set of values of the test statistic that will cause us to reject the null hypothesis.
Critical value	The 'first' (or 'boundary') value in the critical region.
Decision rule	If the calculated test statistic falls within the critical region, the null hypothesis H_0 is rejected. In contrast, if the calculated test statistic does not fall within the critical region, the null hypothesis is not rejected.
F-distribution	A continuous distribution that has two parameters (*df* for the numerator and *df* for the denominator). It is mainly used to test hypotheses concerning variances.
F-ratio	In ANOVA, it is the ratio of between-column variance to within-column variance.
Hypothesis	An unproven proposition or supposition that tentatively explains a phenomenon.

Null hypothesis	A statement about a status quo about a population parameter, that is being tested.
One-tail test	A statistical hypothesis test in which the alternative hypothesis is specified such that only one direction of the possible distribution of values is considered.
Power of the hypothesis test	The probability of rejecting the null hypothesis when it is false.
Significance level	The value of α that gives the probability of rejecting the null hypothesis when it is true. This gives rise to Type I error.
Test criteria	Criteria consisting of (i) specifying a level of significance α, (ii) determining a test statistic, (iii) determining the critical region(s), and (iv) determining the critical value(s).
Test statistic	The value of Z or t calculated for a sample statistic such as the sample mean or the sample proportion.
Two-tail test	A statistical hypothesis test in which the alternative hypothesis is stated in such a way that it includes both the higher and the lower values of a parameter than the value specified in the null hypothesis.
Type I error	An error caused by rejecting a null hypothesis that is true.
Type II error	An error caused by failing to reject a null hypothesis, that is not true.

LIST OF FORMULAE

1. (a) The test statistic Z for $\bar{x}$ when σ is known

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

(b) The test statistic Z for $\bar{x}$ when σ is unknown but $n > 30$

$$Z = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

where $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ is the estimated standard error of the mean

2. The test statistic t for a small sample when σ is unknown but $n \leq 30$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

3. The test statistic z for p (proportion) for a large sample

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$, where $\bar{p}$ = sample proportion, p = population proportion

4. The test statistic z for test concerning differences between two population means

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

when $H_0 : \mu_1 - \mu_2 = 0$, then the test statistic z is

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

when σ_1 and σ_2 are unknown, then s_1 and s_2 are used.

5. The test statistic z for test concerning differences between two population proportions

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1 - \bar{p})\,(1/n_1 + 1/n_2)}}$$

when $H_0 : p_1 = p_2$, then the test statistic z is

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\,(1/n_1 + 1/n_2)}}$$

Pooled sample proportion $(\bar{p})$ for two samples

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}, \text{ where } \bar{p}\,(1 - \bar{p}) = \bar{p}\,\bar{q}$$

6. The power of a statistical test

$1 - \beta$

Ability of the test to perform as required. The greater the value of $1 - \beta$, the better is the decision.

7. The test statistic t when paired observations are involved

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

where $\bar{d} = \frac{\Sigma d}{n}$ and $d = x_2 - x_1$

s = estimate of σ.

8. The F statistic

$$F = \hat{\sigma}_1^2/\hat{\sigma}_2^2 = s_1^2/s_2^2$$

where s_1^2 = Variance of Sample 1

s_2^2 = Variance of Sample 2

QUESTIONS

13.1 Given below are twelve statements. Indicate in each case whether the statement is true or false:

(a) We commit a Type I error when we reject a null hypothesis when it is really true.

(b) It is easier to select an appropriate significance level than to select a proper test to use.

(c) The normal probability distribution is always the basis in testing hypotheses.

(d) In testing a hypothesis, one can make three types of error.

(e) If our null hypothesis is $H_0 : \mu = 50$ and the alternative hypothesis is $H_1 : \mu < 50$, then the test is known as the left-tail test.

(f) An exercise in hypothesis testing enables us to draw conclusions about the estimated parameters.

(g) The value $(1 - \beta)$ is known as the power of the test.

(h) If the critical value of Z is 1.96, then the significance level of one-tail test is 0.05.

(i) On the basis of a hypothesis test, we can determine whether a population mean is 55 or 50 (i.e. $H_0 : \mu = 55$; $H_1 : \mu = 50$).

(j) For a given level of significance, we find that as the sample size increases, the critical values of t get closer to zero.

(k) If the standardised sample mean exceeds the critical value, we should accept H_0.

(l) A 1 per cent significance level provides greater confidence to the decision than a 5 per cent significance level.

Multiple Choice Questions (13.2 to 13.14)

13.2 When the null hypothesis is H_0: $\mu = 50$, the alternative hypothesis can be

(a) H_1: $\mu = 49$ (b) H_1: $\mu \neq 49$

(c) H_1: $\mu < 50$ (d) H_1: $\mu \geq 50$

13.3 For a two-tail test of a hypothesis at $\alpha = 0.05$, the acceptance region is

(a) between the two critical values

(b) outside the two critical values

(c) to the left of the positive critical value

(d) to the right of the negative critical value

13.4 When a null hypothesis is rejected, then it is possible that

(a) A Type II error has been made

(b) A correct decision has been made

(c) neither (a) nor (b) has occurred

(d) both (a) and (b) have occurred

(e) none of these

13.5 For a two-tail test when n is large, the value of Z at 0.05 level of significance is

(a) 1.645 (b) 2.58 (c) 1.96 (d) none of these

13.6 Supposing that while testing a hypothesis we think that Type I error will be much costlier than Type II error, which of the following values of α should we adopt?

(a) 0.5 (b) 0.1 (c) 0.05 (d) 0.01

13.7 When α is changed from 0.5 to 0.1, the probability of rejecting a null hypothesis that is actually true

(a) Increases (b) Remains the same

(c) Decreases (d) All of these

13.8 A choice of on appropriate significance level is made by examining the cost of
(a) A Type II error (b) A Type I error
(c) Performing the test (d) (b) and (c)
(e) (a) and (b) (f) (a) and (c)

13.9 When a test of hypothesis involves the mean of a normal population with a known standard deviation, the comparison should be between:

(a) The observed value of $\bar{X}$ with the critical value of Z
(b) The observed value of Z with the critical value of $\bar{X}$
(c) The observed value of Z with the critical value of Z
(d) The observed value of $\bar{X}$ with the critical value of $\bar{X}$
(e) Either (c) or (d)

13.10 In performing a hypothesis test, when we choose a significance level of 0.05, it means that
(a) Our risk in accepting a hypothesis which is false is 5%.
(b) Our minimum standard for acceptable probability is 5%.
(c) Our risk in rejecting a hypothesis which is true is 5%.
(d) (a) and (b) only
(e) (b) and (c) only
(f) (a) and (c) only

13.11 Assuming that we want to test whether a population mean is significantly smaller or larger than 50, what should be our alternative hypothesis?
(a) $\mu < 50$
(b) $\mu > 50$
(c) $\mu \neq 50$
(d) From the information given, it is not possible to determine.

13.12 For a hypothesis test, we have taken a sample of 20 and $\bar{X}$ computed. At a significant level of 0.05, where should we look for the critical value?
(a) The Z table, where 0.95 of the area is to the left of the Z value.
(b) The Z table, where 0.90 of the area is to the left of the Z value.
(c) The t table, where, with 19 degrees of freedom, the column heading is 0.05.
(d) The t table, where, with 19 degrees of freedom, the column heading is 0.10.

13.13 In a hypothesis test, $\alpha = 0.05$ and $\beta = 0.10$, the power of the test is
(a) 0.95 (b) 0.10 (c) 0.90 (d) none of these

13.14 For testing $P_1 = P_2$ in a large sample, the proper test is
(a) t test (b) Z test (c) F test (d) none of these

13.15 What is a hypothesis? What steps are involved in statistical testing of a hypothesis?

13.16 Differentiate between Type I error and Type II error in hypothesis testing. Explain the relationship between the two errors.

13.17 Distinguish between a one-tail and a two-tail test, giving a diagram and an example in each case.

13.18 Explain how hypothesis testing is useful to management.

13.19 What do you understand by the power of a statistical test? Explain the concept by giving a suitable example.

13.20 When would you prefer **(a)** a one-tail test and **(b)** a two-tail test?

13.21 "A hypothesis can only be rejected but it can never be accepted". Do you agree with this statement? Why or why not?

13.22 State the null and alternative hypotheses regarding population mean that lead to **(i)** left-tail test, **(ii)** right-tail test and **(iii)** two-tail test.

13.23 Differentiate the following pairs of concepts:

(a) Critical region and Acceptance region

(b) Null hypothesis and Alternative hypothesis

(c) One-tail test and Two-tail test

(d) Type I error and Type II error

13.24 A manufacturer for a new motorcycle claims for it an average mileage of 60 km per litre with a standard deviation of 2 km under city conditions. However, the average mileage in 64 trials is found to be 57 km. Is the manufacturer's claim justified?

13.25 An owner of the press claims that the life of its largest web press is 14,500 hours with a standard deviation of 2,100 hours. From a sample of 25 presses, the company finds a sample mean of 13,000 hours. Do you think that the average life of the presses is different from 14,500 hours as claimed by the company? Verify it at 0.01 level of significance.

13.26 A radio shop sells, on an average, 200 radios per day with a standard deviation of 50 radios. After an extensive advertising campaign, the management will compute the average sales for the next 25 days to see whether an improvement has occurred. Assume that the daily sales of radios are normally distributed.

(i) Write down the null and alternative hypotheses

(ii) Test the hypothesis at 5 per cent level of significance if $\bar{x} = 216$.

(iii) How large must $\bar{x}$ be in order that the null hypothesis is rejected at 5 per cent level of significance?

13.27 An aircraft manufacturer needs to buy aluminium sheets of 0.05 inch in thickness. Thinner sheets would not be appropriate and thicker sheets would be too heavy. The aircraft manufacturer takes a random sample of 100 sheets and finds that their average thickness is 0.048 inch and their standard deviation is 0.01 inch. Should the aircraft manufacturer buy the aluminium sheets from his supplier if he wants to make the decision at

(i) 5 per cent level of significance?

(ii) 1 per cent level of significance?

13.28 An advertising company feels that 20 per cent of the population in the age group of 18 to 25 years in a town watch a specific serial. To test this assumption, a random sample of 2,000 individuals in the same age group was taken of which 440 watched the serial. At 5 per cent level of significance, can we accept the assumption laid down by the company.

13.29 The credit manager of a department store chain believes that the average age of charge account customers is less than 30 years. A random sample of 100 charge account customers reveals a mean age of 27 years and a standard deviation of 10 years. Do these data provide sufficient evidence to support the credit manager's belief? Let level of significance be $\alpha = 0.05$.

13.30 A stock-broker claims that she can predict with 85 per cent accuracy whether a stock's market value will rise or fall during the coming month. As a test, she predicts the outcome of 60 stocks and is correct in 45 of the predictions. Do these data support the stock-broker's claim?

13.31 An insurance agent has claimed that the average age of policyholders who insure through him is less than the average for all agents, which is 30.5 years.

A random sample of 100 policyholders who had insured through him gave the following age distribution.

Age Last Birthday (Years)	*No. of Persons Insured*
16 – 20	12
21 – 25	22
26 – 30	20
31 – 35	30
36 – 40	16
Total	**100**

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at the 5 per cent level of significance.

13.32 A random sample of boots worn by 36 soldiers in a desert region shows an average life of 1.08 years with the standard deviation of 0.6 years. Under the standard condition, the boots are known to have an average life of 1.28 years. Is there a reason to assert at 1 per cent level of significance that use in desert causes the mean life of such boots to decrease? What will be your conclusion if the level of significance is 5 per cent? Assume that the life of boots is normally distributed.

13.33 Two types of batteries are tested for their length of life and the following data are obtained:

	Sample Size	*Mean Life*	*Variance (hour)*
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means?

13.34 You are given the following data about the life of two brands of bulbs:

	Mean Life	*Standard Deviation*	*Sample Size*
Brand A	2,000 hours	250 hours	12
Brand B	2,230 hours	300 hours	15

Is there a significant difference in the mean life of two brands?

13.35 Two types of scooters produced in India are tested for mileage. One group consisting of 36 scooters averaged 24 km per litre of petrol, while the other group consisting of 72 scooters averaged 22.5 km per litre of petrol.

(i) What test statistic is appropriate if $\sigma_1^2 = 1.5$ and $\sigma_2^2 = 2.0$?

(ii) Test, whether statistically there exists a significant difference in the petrol consumption of these two types of scooters. (Use $\alpha = 0.01$.)

13.36 The cinema-goers were 800 people out of a sample of 1,000 persons during the period of a fortnight in a town where no TV programme was available. The cinema-goers were 700 people out of a sample of 2,800 persons during a fortnight in another town where TV programme was available. Do you think that there has been a significant decrease in proportion of cinema-goers after introducing TV sets? Test the hypothesis that there is no difference in proportion of cinema-goers. Use a level of significance of 0.01.

13.37 A random sample of 100 mill-workers at Kanpur showed their mean wage to be Rs 3,500 with a standard deviation of Rs 280. Another random sample of 150 mill-workers in Mumbai showed the mean wage to be Rs 3,900 with a standard deviation of Rs 400. Do the mean wages of workers in Mumbai and Kanpur differ significantly? Use 0.05 level of significance.

13.38 A man buys 50 electric bulbs of 'Philips' make and another 50 of the 'GE' make. He finds that the Philips bulbs give an average life of 1,500 hours with a standard deviation of 60 hours and the GE bulbs give an average of 1,512 hours with a standard deviation of 80 hours. Is there a significant difference between the two makes?

13.39 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	20	15
Average sales (in Rs)	170	200
Standard deviation (in Rs)	20	25

13.40 Ten persons were appointed in officer cadre in an office. Their performance was evaluated by giving a test and the marks were recorded out of 100. They were given two months' training and another test was held and marks were recorded out of 100.

Employees	*Before Training*	*After Training*
A	80	84
B	76	70
C	92	96
D	60	80
E	70	70
F	56	52
G	74	84
H	56	72
I	70	72
J	56	50

By applying the *t*-test, can it be concluded that the employees have benefited by the training? Use 5 per cent level of significance.

13.41 Samples of two different types of bulbs were tested for length of life, and the following data were obtained:

	Sample Size	*Sample Mean*	*Sample S.D.*
Type 1	8	1,234 hours	36 hours
Type 2	7	1,136 hours	40 hours

Is the difference in means significant?

13.42 500 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 per

cent level of significance that production at the second factory is better than in the first factory?

13.43 Judge the correctness of the following:
To test $H_0 : \sigma = \sigma_0$ (given); $H_1 : \sigma < \sigma_0$, given that μ is known and sample size is 10, the relevant test statistic follows a *t*–distribution with degrees of freedom as 10 and the value under H_0 turns out to be –3.06.

13.44 Judge the correctness of the following:
Given that μ and σ are both unknown for testing the hypotheses $H_0 : \mu = \mu_0$ (given), $H_1 : \mu > \mu_0$ with the sample size as 15, the value of the test statistic under H_0 should be between 3.5 and 7.2 so as to reject H_0 at 5 per cent level of significance and to accept H_0 at 1 per cent level of significance.

13.45 In a study of advertising effectiveness, a sample of 50 housewives rated a new bleach product when it first appeared. After a month of intensive television advertising, the same women were asked to try the product again. Using a scoring system based on perceptions of product effectiveness, the difference in scores given to the product had mean 1.6 and standard deviation 2.0. Is there evidence that perceptions of the product's effectiveness changed during the period of the advertisement? Carry out the test at 0.05 level of significance.

13.46 A large corporation finds that 63 per cent of the 150 salespeople who have never had a self-improvement course would like such a course. The firm had done a similar study 10 years ago. Then, only 58 per cent of 160 salespeople wanted a self-improvement course. At the 0.05 level of significance, test the null hypothesis that salespeople are no more eager for self-improvement course this year than they were 10 years ago. The groups are assumed to constitute two independent simple random samples.

13.47 Wansley, Rosenfeldt, and Cooley (1983) compared the profiles of a sample of 44 firms that merged during 1975–76 with those of a sample of 44 firms that did not merge. The table displays information obtained on the firm's price-earning ratios.

	Merged Firms	*Non-merged Firms*
Sample mean	7.295	14.666
Sample standard deviation	7.374	16.089

(i) The analysis performed by Wensley, Rosenfeldt, and Cooley indicated that "merged firms generally have smaller price-earning ratios". Do you agree? Test using $\alpha = 0.05$.

(ii) Report the *p*-value of the test you conducted in part (i).

(iii) What assumption(s) was necessary to make in order to perform the test in part (i)?

(iv) Do you think that the distribution of the price-earning ratios for the population from which these samples were drawn are normally distributed? Why or why not?

13.48 One way by which corporations raise money for expansion is to issue bonds, which are loan agreements to repay the purchaser a specified amount of money with a fixed rate of interest paid periodically over the life of the bond. The sale of the bonds is usually handled by a underwriting firm. In a study described in the *Harvard Business Review* (July–August 1979), D. Logue and R. Rogalski ask the question, "Does it pay to shop for your bond underwriter?" The reason for the question is that the price of a bond may rise or fall after its issuance.

Therefore, whether a corporation receives the market price for a bond depends on the skill of the underwriter. The mean change in the price of 27 bonds over a 12-month period by one underwriter and in the prices of 23 bonds handled by another are given in the table.

	Underwriter I	*Underwriter II*
Sample size	27	23
Sample mean	– 0.0491	– 0.0307
Sample variance	0.009800	0.002465

Do the data provide sufficient evidence to indicate a difference in the mean change in bond prices handled by the two underwriters? Test at $\alpha = 0.05$.

13.49 In a survey on consumption pattern, 400 women shoppers are chosen at random in Supermarket A, located in a certain section of Mumbai city. Their average monthly food expenditure is Rs 250, with standard deviation of Rs 40. For 400 women shoppers, chosen at random in Supermarket B, in another section of the city, the average food expenditure is Rs 220, with standard deviation of Rs 55. Test at 1% level of significance whether the average food expenditure of the two populations of shoppers, from which the samples were obtained, are equal.

13.50 The mean produce of wheat of a sample of 100 farmers comes to 200 kg per hectare, with standard deviation of 10 kg. Another sample of 150 farmers gives the mean at 220 kg, with standard deviation of 12 kg. Assuming that the population standard deviation is 11 kg, examine whether the difference in mean production in the two samples is significant at 5% level of significance.

13.51 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 50 cars was 10 km per litre, with standard deviation of 35 km per litre. Test for the company at 5% level of significance, whether the claim that the new car's petrol consumption is 9.5 km per litre, on an average, is acceptable.

13.52 A manufacturer of sports equipment has developed a new synthetic fishing line that he claims has a mean breaking strength of 8 kg, with standard deviation of 0.5 kg. Test the hypothesis that $\mu = 8$ kg against alternative hypothesis $\mu \neq 8$ kg, if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kg. Use a 0.01 level of significance.

13.53 A study shows that 16 of the 200 tractors produced on one assembly line required extensive adjustments before they could be shipped, while the same was true for the 400 tractors produced on another assembly line. At the 0.01 level of significance, does this support the claim that the second production line does superior work?

13.54 The mean weekly wage for a sample of $n_1 = 30$ employees in a large firm is Rs 2800, with a sample standard deviation, $s_1 =$ Rs 140. In another firm, a sample of $n_2 = 40$ employees have a mean wage of Rs 2700, with standard deviation, $s_2 =$ Rs 100. The standard deviations of the populations are not assumed to be equal. Test the hypothesis that there is no difference between the mean weekly wage amounts of the two firms at 5% significance level.

CHAPTER 14 CHI-SQUARE DISTRIBUTION

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the meaning and uses of chi-square test
- understand the shape of the chi-square distribution
- calculate chi-square by using different formulae and interpret the result
- know the precautions to be taken while using the chi-square test

Chapter Prerequisites

Before starting work on this chapter, make sure you are quite conversant with

1. all the material on testing hypotheses covered in Chapter 13
2. the computation of standard deviation and variance

14.1 INTRODUCTION

In the preceding chapter, some tests of significance were discussed with a number of examples. These tests were based on the assumption that the samples were drawn from normally distributed population or, more appropriately, that the sample means were normally distributed. These tests are known as *parametric tests* as the testing procedure involves the assumption about the type of population of parameters. There are, however, many situations where it is not possible to assume a particular type of population distribution from which samples are drawn. This limitation has led to the development of *a number of alternative techniques that are known as non-parametric or distribution-free methods. Chi-square test is one of the important non-parametric methods*.

This chapter is devoted to chi-square tests and deals with three types of such tests. These are:

1. Tests of hypotheses for experiments with more than two categories, called *goodness-of-fit* tests.
2. Tests of hypotheses about contingency tables, called *independence* and *homogeneity* tests.
3. Tests of hypotheses about the variance and standard deviation of a single population.

As all these tests are performed by using the chi-square distribution, we should first know about the chi-square distribution before we use a chi-square test.

The Chi-square Distribution

At the outset, we should know that the chi-square distribution has only one parameter called the 'degrees of freedom' (*df*) as is the case with the *t*-distribution. The shape of a particular chi-square distribution depends on the number of degrees of freedom. This can be seen from Fig. 14.1.

How are these degrees of freedom calculated? We shall come to this query when we deal with specific problems. It may also be noted that the random variable chi-square assumes positive values only. As such, a chi-square curve starting from the origin (zero point) lies entirely to the right of the vertical axis. Figure 14.1 shows three chi-square distribution curves. They are for 2, 7 and 12 degrees of freedom, respectively.

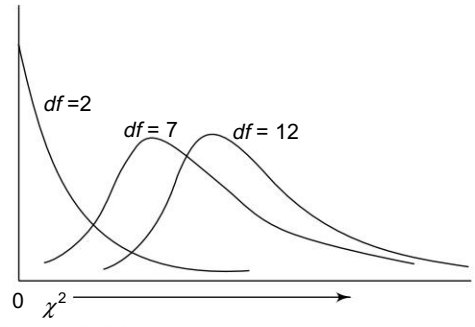


Fig. 14.1 **Chi-square Distribution with Three Different Curves**

As can be seen from Fig. 14.1, the shape of a chi-square distribution curve is skewed for very small degrees of freedom. As the degrees of freedom increase, the shape also changes. Eventually, for large degrees of freedom, the curve looks similar to the curve of a normal distribution. A point worth noting is that the total area under a chi-square distribution curve is 1.0 as is the case in all other continuous distribution curves.

Chi-Square Distribution—Properties The properties of the chi-square distribution are given below:

1. Chi-square is non-negative in value; it is either zero or positively valued.
2. It is not symmetrical; it is skewed to the right as can be seen from Fig. 14.1.
3. There are many chi-square distributions. As with the *t*-distribution, there is a different chi-square distribution for each degree-of-freedom value.

If we know the degrees of freedom and the area in the right tail of a chi-square distribution, we can find the value of chi-square (χ^2) from Appendix Table 5. We give below two examples to show how the value of χ^2 can be obtained from this table.

Example 14.1 Find the value of χ^2 for 10 degrees of freedom and an area of 0.05 in the right tail of the chi-square distribution curve.

Solution In order to find the required value of χ^2, we first locate 10 in the column for degrees of freedom (*df*) and 0.05 in the top row in Appendix Table 5. The required chi-square value is given by the entry at the intersection of the row for 10 and the column for 0.05. This value is 18.307.

Example 14.2 Find the value of χ^2 for 15 degrees of freedom and an area of 0.10 in the left tail of the chi-square distribution curve.

Solution It will be seen that this example is different from the earlier one. Here, the area in the left tail of the chi-square distribution curve is given. In such a case, we have to first find the area in the right tail. This is obtained as follows:

Area in the right tail = 1 – area in the left tail. In respect of this particular example,
Area in the right tail = 1 – 0.10 = 0.90.

Now we can find, in the same manner as used earlier, the value of χ^2. We locate 15 in the column for *df* and 0.90 in the top row in the Appendix Table 5. The required value of χ^2 is 8.547.

14.2 THE GOODNESS-OF-FIT TEST

We shall consider the goodness-of-fit test as the first application of chi-square. This test is a test of the agreement (or conformity, or consistency) between a hypothetical and a sample distribution. A number of times we find that the results obtained in samples are not consistent with the theoretical results expected according to the rules of probability. For example, theory says that when we toss a fair coin 100 times, we should expect 50 heads and 50 tails, but in reality we may not find this result; either heads or tails may be more than 50 and thus they may seldom equal. Before we take up an example, we give below the procedure involved in the chi-square test.

Procedure for Conducting Chi-square Test There are six steps as explained below:

1. State the null hypothesis, which is usually as follows: the sample distribution agrees with the hypothetical or theoretical distribution.
2. Calculate the number in each category on the assumption that the null hypothesis is correct. Thus, for each observation, we shall have observed frequency and expected frequency.
3. Determine the level of significance, that is, how much risk of the Type I error we are prepared to take.
4. Calculate the chi-square by using the following formula.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where χ^2 = chi–square
O_i = observed frequency in the i^{th} category
E_i = expected frequency in the i^{th} category
k = number of categories

5. Determine the number of degrees of freedom. For the specified level of significance and the degrees of freedom, find the critical or theoretical value of χ^2.
6. Compare the calculated value of χ^2 with the theoretical value and determine the region of rejection. In case the calculated value of χ^2 is less than the theoretical (or critical) value, the null hypothesis is accepted. If, on the other hand, the calculated value of χ^2 is greater than the theoretical value, the null hypothesis is rejected.

Let us take an example involving the use of χ^2 test.

Example 14.3 Suppose a coin is tossed 200 times with the following results:

Event	*Frequency*
Head	90
Tail	110
Total	**200**

Is this a fair coin?

Solution We set up the null hypothesis.

$H_0 : O_i = E_i$ and alternative hypothesis is
$H_1 : O_i \neq E_i$

where O is the observed frequency and E is the expected frequency. Since a fair coin should have shown 50 per cent head and 50 per cent tail in the total process, the expected frequencies should have shown Head—100 and Tail—100.

Table 14.1 shows the calculations of chi-square.

Table 14.1 Worksheet for Calculating Chi-square

	O	E	$O - E$	$(O - E)^2$	$(O - E)^2 / E$
Head	90	100	–10	100	1
Tail	110	100	10	100	1
				Total	**2**

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = 2$$

Since there are 2 rows and 2 columns, the degree of freedom is $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$. Referring to the Appendix Table 5, we find that the critical value of chi-square for α at 5 per cent level of significance and for 1 degree of freedom is 3.841. Since the calculated value of chi-square is less than 3.841, the null hypothesis is not rejected. In other words, we conclude that the coin is fair.

Yates' Correction for Continuity

Chi-square distribution is a continuous distribution. When results for continuous distribution are applied to discrete data, certain corrections for continuity can be made. Yates' correction is similar to the continuity correction that is applied to the normal approximation to the binomial distribution.

The correction requires rewriting of chi-square as follows:

$$\chi^2 \text{ (corrected)} = \frac{(|O_1 - E_1| - 0.5)^2}{E_1} + \frac{(|O_2 - E_2| - 0.5)^2}{E_2} + \cdots + \frac{(|O_k - E_k| - 0.5)^2}{E_k}$$

It may be noted that this adjustment is used when there is only 1 degree of freedom. Again, when several values of χ^2 are combined, it should not be used. Applying this to our earlier example, we get the following calculations:

$$\chi^2 = \frac{(|90 - 100| - 0.5)^2}{100} + \frac{(|110 - 100| - 0.5)^2}{100}$$

$$= \frac{(9.5)^2}{100} + \frac{(9.5)^2}{100}$$

$$= \frac{90.25}{100} + \frac{90.25}{100} = 1.805$$

It can be seen that the χ^2 in the example, without the continuity correction, was 2. As a result of Yates' continuity correction, there has been a downward correction of the χ^2 value. When n is large, the continuity correction has little effect but it becomes important when n is small.

Example 14.4 Let us take an example of normal distribution. Suppose we are given the following data: An observed distribution and a fitted normal distribution (n = 200, *mean* = 9 and standard deviation = 3).

No. of Readings	*Values*						*Total No. of readings*
	< 3	*3–6*	*6–9*	*9–12*	*12–15*	*15+*	
Observed	7	30	61	73	25	4	200
Theoretical	5	27	68	68	27	5	200

Find out whether the normal distribution gives a close degree of fit.

Solution First of all, we set up the two hypotheses:

H_0 : Population sampled follows a normal distribution.

H_1 : Population sampled does not follow a normal distribution.

Now, we have to decide the degrees of freedom. Degrees of freedom = $p - k - 1$, where p is the number of groupings for which observed and theoretical frequencies are compared; and k is the number of parameters that had to be estimated in fitting the theoretical model (in our example, the mean m and standard deviation s). Further, one degree of freedom is to be subtracted because it has been used up in fixing the sample size n. Thus, in our case, the degrees of freedom are 6 – 2 – 1 = 3.

We now look up the critical value of χ^2 with 3 degrees of freedom at 5 per cent level of significance in Appendix Table 5. This comes to 7.81. Now, we have to calculate χ^2 from the data given in the above table.

Table 14.2 Calculation of Chi-square

O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
7	5	2	4	0.80
30	27	3	9	0.33
61	68	–7	49	0.72
73	68	5	25	0.37
25	27	–2	4	0.15
4	5	–1	1	0.20
				$\chi^2 = 2.57$

Comparing the calculated value of χ^2 with the critical value of 7.81, we find that as the former is less than the latter, we accept the null hypothesis. We, therefore, conclude that the population sampled follows the normal distribution. The discrepancies from the fitted normal distribution may, therefore, be only due to sampling errors.

Example 14.5 A brand manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey in which the country was divided into four geographical regions, a random sampling of 100 consumers in each region was surveyed, with the following results:

	Region				
	NE	*NW*	*SE*	*SW*	*Total*
Purchase the brand	40	55	45	50	190
Do not purchase	60	45	55	50	210
Total	**100**	**100**	**100**	**100**	**400**

Develop a table of observed and expected frequencies for this problem.

(i) Calculate the sample chi-square value.

(ii) State the null and alternative hypotheses.

(iii) At $\alpha = 0.05$, test whether brand share is the same across the four regions.

Solution In order to calculate the expected frequencies for the corresponding observed frequenices, we have to apply the formula:

(Row total × Column total)/Grand total

For example, the observed frequency in row 1 and column 1 is 40. Its expected frequency will be E = (190 × 100)/400 = 47.5. In this manner, expected frequencies are calculated and shown in Table 14.3.

Table 14.3 Calculation of Sample Chi-square

		O	*E*	*O* – *E*	$(O-E)^2$	$(O-E)^2/E$
Row 1	NE	40	47.5	–7.5	56.25	1.18
	NW	55	47.5	7.5	56.25	1.18
	SE	45	47.5	–2.5	6.25	0.13
	SW	50	47.5	2.5	6.25	0.13
Row 2	NE	60	52.5	7.5	56.25	1.07
	NW	45	52.5	–7.5	56.25	1.07
	SE	55	52.5	2.5	6.25	0.12
	SW	50	52.5	–2.5	6.25	0.12
						$\chi^2 = 5.00$

The two hypotheses are as follows:

H_0 : The brand share is evenly distributed.

H_1 : The brand share is not evenly distributed.

The degrees of freedom = $(r-1)(c-1) = (2-1)(4-1) = 3$.

The critical value of χ^2 at $\alpha = 0.05$ level for 3 degrees of freedom from Appendix Table 5 is 7.815.

Since the calculated χ^2 is less than the critical value of 7.815, the null hypothesis is accepted. In other words, the brand share is evenly distributed in all the four regions of the country.

14.3 THE TEST OF INDEPENDENCE

Another application of a chi-square distribution is the test of independence. The population and sample are now classified according to several attributes, but the probability distributions of these classifications are not given. In such cases, we are interested in ascertaining whether there is any dependency

relationship between the two attributes. When we apply the chi-square distribution, it will indicate whether or not the two attributes are independent. The test cannot indicate the degree of association or the direction of dependency. If our chi-square test indicates independence, it means that the observed data are consistent with the hypothesis that the two attributes are independent.

Example 14.6 From the data given in the following table, find out whether there is any relationship between sex and the preference of colour.

Colour	*Male*	*Female*	*Total*
Red	30	40	70
Blue	50	20	70
Green	40	20	60
Total	**120**	**80**	**200**

Solution We have to first calculate the expected value for the observed frequencies. These are shown in Table 14.4 along with observed frequencies.

Table 14.4 Worksheet for Calculating Chi-square

		O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
Red	M	30	42	– 12	144	3.42
	F	40	28	12	144	5.14
Blue	M	50	42	8	64	1.52
	F	20	28	– 8	64	2.28
Green	M	40	36	4	16	0.44
	F	20	24	– 4	16	0.67
						$\chi^2 = 13.47$

H_0 : There is no relationship between sex and preference of colour.
H_1 : There is relationship between sex and preference of colour.

The degree of freedom are $(r - 1)\,(c - 1) = (3 - 1)\,(2 - 1) = 2$.

The critical value of χ^2 for 2 degrees of freedom at 0.05 level of significance from Appendix Table 5 is 5.991.

Since the calculated $\chi^2 = 13.47$ exceeds the critical value of χ^2, the null hypothesis is rejected. Hence, the conclusion is that there is a definite relationship between sex and preference of colour.

Example 14.7 A company has introduced a new drug B to cure common cold. It is being compared against an existing drug A. The relevant data are shown below:

	Helped	*Harmed*	*No effect*	*Total*
Drug A	44	10	26	80
Drug B	52	10	18	80
Total	**96**	**20**	**44**	**160**

Is the new drug more effective in curing cold?

Solution

Let us first calculate the χ^2 with the help of expected frequencies. This is shown in Table 14.5.

Table 14.5 Worksheet for Calculating Chi-square

	O	E	$O - E$	$(O - E)^2$	$(O - E)^2 / E$
R_1C_1	44	48	– 4	16	0.333
R_1C_2	10	10	0	0	0
R_1C_3	26	22	4	16	0.727
R_2C_1	52	48	4	16	0.333
R_2C_2	10	10	0	0	0
R_2C_3	18	22	– 4	16	0.727
					$\chi^2 = 2.12$

We set up the two hypotheses:

H_0 : There is no difference in the effectiveness of the two drugs.
H_1 : There is difference in the effectiveness of the two drugs.

In order to test the null hypothesis, we should first decide the degrees of freedom. Our problem contains two rows and three columns. Hence, the degrees of freedom will be $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$.

For 2 degrees of freedom, the critical value of χ^2 at 0.05 level of significance is 5.991 (Appendix Table 5). Since the calculated value of χ^2 being 2.12 is less than the critical value, the null hypothesis cannot be rejected. Our conclusion is that in terms of effectiveness in the treatment of common cold, there is no difference between the two drugs.

Simple Formulae for Computing Chi-square (χ^2) At this stage, we may introduce a simpler approach for calculating χ^2 in a 2×2 table where the cell frequencies and the marginal totals are as shown below:

a	b	$(a + b)$
c	d	$(c + d)$
$(a + c)$	$(b + d)$	N

Chi-square can be calculated by a simple formula, which involves only the observed frequencies. The formula is

$$\chi^2 = \{N(ad - bc)\}^2 / \{(a + c)(b + d)(c + d)(a + b)\}$$

In case there is a 2×3 table, the alternative formula for calculating χ^2 will be different and is given below:

a	b	c	$(a + b + c)$
d	e	f	$(d + e + f)$
$(a + d)$	$(b + e)$	$(c + f)$	N

The formula for using the values on the basis of the above table is as follows:

$$\chi^2 = \frac{N}{a+b+c}\left(\frac{a^2}{a+d}+\frac{b^2}{b+e}+\frac{c^2}{c+f}\right)+\frac{N}{d+e+f}\left(\frac{d^2}{a+d}+\frac{e^2}{b+e}+\frac{f^2}{c+f}\right)-N$$

where the general result $\chi^2 = \Sigma\ \{(O_j^2)\ /\ (E_j^2)\} - N$ is used.

Let us take a numerical example to illustrate the application of the above alternative formula. This formula is applied to the data given in Example 14.7.

$$\chi^2 = \frac{160}{80}\left(\frac{(44)^2}{96}+\frac{(10)^2}{20}+\frac{(26)^2}{44}\right)+\frac{160}{80}\left(\frac{(52)^2}{96}+\frac{(10)^2}{20}+\frac{(18)^2}{44}\right)-160$$

$$= 2\left(\frac{1,936}{96}+\frac{100}{20}+\frac{676}{44}\right)+2\left(\frac{2,704}{96}+\frac{100}{20}+\frac{324}{44}\right)-160$$

$$= 2(20.17 + 5 + 15.36) + 2\ (28.17 + 5 + 7.36) - 160$$

$$= 2(40.53) + 2(40.53) - 160 = 81.06 + 81.06 - 160 = 2.12$$

It will be seen that both the methods give the same value of χ^2, which is 2.12.

14.4 CHI-SQUARE (χ^2) AS A TEST OF HOMOGENEITY

This type of application of χ^2 test can be regarded as an extension of the χ^2 test of independence. Such tests indicate whether two or more independent samples are drawn from the same population or from different populations. Suppose a test is given to students in two different higher secondary schools. The sample size in both the cases is the same. The question we have to ask is: is there any difference between the two higher secondary schools? In order to find the answer, we have to set up the null hypothesis that the two samples came from the same population. The word 'homogeneous' is used frequently in Statistics to indicate 'the same' or 'equal'. Accordingly, we can say that we want to test in our example whether the two samples are homogeneous. Thus, the test is called a *test of homogeneity*.

Let us take an example.

Example 14.8 A company has two factories—one located in Delhi and another in Mumbai. It is interested to know whether its workers are satisfied with their jobs or not at both places. To get proper and reliable information, it has undertaken a survey at both the factories, and the data obtained are shown as follows:

Table 14.6 Number of Workers by Degree of Satisfaction

	Delhi	*Mumbai*	*Total*
Fully satisfied	50	70	120
Moderately satisfied	90	110	200
Moderately dissatisfied	160	130	290
Fully dissatisfied	200	190	390
Total	**500**	**500**	**1,000**

Solution It may be noted that in this example the column total are fixed. This means that it is decided in advance to take a sample of 500 workers each from Delhi and Mumbai. However, the row total of 120, 200, 290 and 390 are determined randomly by the outcomes of the two samples. A point that must be noted is that when either the column totals or the row totals are fixed, we use a test of homogeneity. When both the row and column totals are determined randomly, we use a test of independence.

Let us revert to our example of job satisfaction.

We first state the null and alternative hypotheses.

Step 1

H_0: The proportions of workers who belong to the four job satisfaction categories are the same in both Delhi and Mumbai.

H_1: The proportions of workers who belong to the four job satisfaction categories are not the same in Delhi and Mumbai.

Step 2

We select the test criterion. In this case, the chi–square distribution will be used to test the homogeneity.

Step 3

The rejection and non-rejection regions are to be determined. The significance level is 5 per cent. As the homogeneity test is right-tail, the area of the rejection region is 0.05, which lies in the right-tail of the chi-square distribution curve.

As this example contains four rows and two columns, the degrees of freedom are: $df = (r - 1)(c - 1) = (4 - 1)(2 - 1) = 3$. From Appendix Table 5, the critical value of χ^2 for 3 df and $\alpha = 0.05$ area in the right-tail of the chi-square distribution curve is 7.815.

Step 4

We have to calculate the value of the test statistic. For this purpose, it is now necessary to calculate the expected frequencies (E). The expected frequencies are calculated on the basis of the following formula.

$$E = \frac{(\text{Row total})\,(\text{Column total})}{\text{Total number of observations}}$$

These frequencies are shown in Table 14.7.

Table 14.7 Worksheet for Calculating Chi-square

	Delhi		*Mumbai*	
	O	*E*	*O*	*E*
Fully satisfied	50	60	70	60
Moderately satisfied	90	100	110	100
Moderately dissatisfied	160	145	130	145
Fully dissatisfied	200	195	190	195

The value of the test statistic is computed by the following formula:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(50-60)^2}{60} + \frac{(90-100)^2}{100} + \frac{(160-145)^2}{145} + \frac{(200-195)^2}{195} +$$

$$\frac{(70-60)^2}{60} + \frac{(110-100)^2}{100} + \frac{(130-145)^2}{145} + \frac{(190-195)^2}{195}$$

$$= 1.667 + 1.000 + 1.552 + 0.128 + 1.667 + 1.000 + 1.552 + 0.128$$

$$= 8.694$$

Step 5

Make a decision—the value of the test statistic $\chi^2 = 8.694$ is greater than the critical value of $\chi^2 = 7.815$ for 3 degrees of freedom at 0.05 level of significance. As such, the calculated value falls in the rejection region. We, therefore, reject the null hypothesis and conclude that the distribution of job satisfaction for workers in Delhi and Mumbai is not homogeneous.

Example 14.9 A test in mathematics was given to students of two higher secondary schools in Delhi. The sample size in each case was 100, that is, $n_1 = 100$ and $n_2 = 100$. The break-up of students on the basis of grades obtained by them is shown below:

Table 14.8 The Break-up of Students

Grade		Sample		Total
		HSS 1	*HSS 2*	
p_1	A	15	10	25
p_2	B	25	15	40
p_3	C	35	40	75
p_4	D	25	35	60
		$n_1 = 100$	**$n_2 = 100$**	**200**

**HSS = Higher Secondary School*

The question is: is there any difference between the grades of the two higher secondary schools?

Solution In order to answer this question, we have to first set up null and alternative hypotheses.

H_0 : The two samples came from the same population.

H_1 : The two samples came from different populations.

This means that we are interested in testing whether the data or samples are homogeneous in this problem.

We now set up the worksheet for the calculation of χ^2.

Table 14.9 Worksheet for Calculating Chi-square

Grade	*O*	*E*	*O – E*	$(O-E)^2$	$(O-E)^2/E$
A_1	15	12.5	2.5	6.25	0.50
B_1	25	20.0	5.0	25.00	1.25
C_1	35	37.5	–2.5	6.25	0.17
D_1	25	30.0	–5.0	25.00	0.83
A_2	10	12.5	–2.5	6.25	0.50
B_2	15	20.0	–5.0	25.00	1.25
C_2	40	37.5	2.5	6.25	0.17
D_2	35	30.0	5.0	25.00	0.83
					$\chi^2 = 5.50$

Degrees of freedom $(r-1)(c-1) = (4-1)(2-1) = 3$.

The critical value of χ^2 for 3 degrees of freedom at 5 per cent level of significance is 7.815. As the calculated value of χ^2 is less than 7.815, the null hypothesis cannot be rejected. In other words, it can be concluded that the two samples came from the same population.

As we are going to ascertain whether the two samples came from the same population, there is a single probability distribution or the attribute in the population. In case of the present example, it is p_1, p_2, p_3 and p_4 as shown in Table 14.9. For example, p_1 denotes the proportion of A's in the population. It may be noted that we do not actually know this probability distribution.

We may restate the null hypothesis. Instead of saying that the two samples came from the same population, we may now say that both samples have the same probability distribution, which is the probability distribution of the population.

Additive Property of χ^2

At this stage, we may like to know that χ^2 has an additive property. This will be clear from an example.

Example 14.10 Suppose we have conducted repeated experiments based on the same data. The results of three experiments can be denoted by χ_1^2, χ_2^2 and χ_3^2. The values of these chi-squares are:

$$\chi_1^2 = 2.08;\ \chi_2^2 = 2.51 \text{ and } \chi_3^2 = 3.63.$$

Each of these corresponds to 1 degree of freedom. Find an overall chi-square.

Solution In order to obtain an overall χ^2, we add all the three values. Thus,

$$\chi^2 = 2.08 + 2.51 + 3.63 = 8.22 \text{ with } 1 + 1 + 1 = 3 \text{ degrees of freedom.}$$

The critical value of χ^2 for 3 degrees of freedom at 0.05 level of significance is 7.815. Since the calculated $\chi^2 = 8.22$ is greater than the critical value of χ^2, we may have to reject the null hypothesis, whatever it may be. It may be noted that in each of the three experiments, the value of χ^2 is 1 degree of freedom, which was less than the critical value of 3.841 at 0.05 level of significance. This means that in case of individual experiments, null hypothesis was not rejected. But, when the results of the three experiments are combined, then the null hypothesis has to be rejected. Such a situation may occur at times while adding individual chi-square values.

Additional Examples

Example 14.11 Two hundred bolts were selected at random from the output of each of the five machines. The numbers of defective bolts found were 4, 8, 12, 6 and 5. Is there a significant difference among the machines? Use 0.05 level of significance.

Solution As there are five machines, the total number of defective bolts should be equally distributed among these machines. That give us expected frequencies. Calculations are shown below.

Table 14.10 Worksheet for Calculating Chi-square

Machine	*O*	*E*	*O – E*	$(O-E)^2$	$(O-E)^2/E$
1	4	7	–3	9	1.29
2	8	7	1	1	0.14
3	12	7	5	25	3.57
4	6	7	–1	1	0.14
5	5	7	–2	4	0.57
					5.71

$df = 5 - 1 = 4$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 5.71$$

H_0 : There is no significant difference among the machines.

H_1 : There is significant difference among the machines.

The critical value of χ^2 at 0.05 level of significance for 4 degrees of freedom is 9.488. As the calculated value of $\chi^2 = 5.71$ is less than the critical value, H_0 is accepted. In other words, the difference among the five machines in respect of defective bolts is not significant.

Example 14.12 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, the manager reports the following number of customers in the stores as:

3,000, 2,960, 3,100, 2,780 and 3,160

Test the divisional manager's belief at 5 per cent level of significance.

Solution We set up the worksheet for calculating the chi-square.

Table 14.11 Worksheet for Calculating Chi-square

Stores	*O*	*E*	*O – E*	$(O-E)^2$	$(O-E)^2/E$
1	3000	3000	0	0	0
2	2960	3000	–40	1600	0.533
3	3100	3000	100	10000	3.333
4	2780	3000	–220	48400	16.133
5	3160	3000	160	25600	8.533
					28.532

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 28.532$$

H_0 : There is no significant difference in the number of customers entering the five stores.

H_1 : There is significant difference in the number of customers entering the five stores.

The critical value of χ^2 at 5 – 1 = 4 *df* at 0.05 level of significance is 9.488. As the calculated value of χ^2 is far greater than the critical value, the null hypothesis is rejected. The conclusion is that there is a significant difference in the number of customers entering each of the five stores.

Example 14.13 A production supervisor is interested in knowing if number of breakdowns on four machines is independent of the shift using the machines. Test this hypothesis based on the following sample information:

Shift	*Machine*			
	A	*B*	*C*	*D*
Morning	15	10	18	12
Evening	12	8	15	10

Solution We set up the two hypotheses as follows:

H_0 : Breakdowns on four machines are independent of the shift using the machines,

H_1 : Breakdowns on four machines are not independent of the shift using the machines.

Table 14.12 Worksheet for Calculating Chi-square

Shift	*O*	*E*	*O – E*	$(O-E)^2$	$(O-E)^2/E$
Morning	15	14.85	0.15	0.0225	0.0015
	10	9.90	0.10	0.0100	0.0010
	18	18.15	–0.15	0.0225	0.0012
	12	12.10	–0.10	0.0100	0.0008
Evening	12	12.15	–0.15	0.0225	0.0019
	8	8.10	–0.10	0.0100	0.0012
	15	14.85	0.15	0.0225	0.0015
	10	9.90	0.10	0.0100	0.0010
					0.0101

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 0.0101$$

$df = (r-1)(c-1) = (2-1)(4-1) = 3$

The critical value of χ^2 with 3*df* at 5% level of significance is 7.815. As the calculated value of χ^2 is less than the critical value, the null hypothesis is accepted. The conclusion is that the breakdowns are independent of the shift using the machines.

Example 14.14 A survey of 200 firms found the following evidence regarding profitability and market share.

	Market share			
Profitability	*Upto 10%*	*11–25%*	*> 25%*	*Total*
Low	36	14	16	66
Medium	26	22	16	64
High	16	24	30	70
Total	**78**	**60**	**62**	**200**

Do you find that the above data give you significant evidence to conclude that market share and profitability are related? Test the hypothesis at 0.01 level of significance.

Solution We set up the two hypotheses as follows:

H_0 : Market share and productivity are unrelated.

H_1 : Market share and productivity are related.

Now, we set up the worksheet.

Table 14.13 Worksheet for Calculating Chi-square

Profitability	*O*	*E*	*O – E*	$(O-E)^2$	$(O-E)^2/E$
Low	36	25.74	10.26	105.267	4.090
	14	19.80	–5.80	33.640	1.699
	16	20.46	–4.46	19.892	0.972
Medium	26	24.96	1.04	1.082	0.043
	22	19.20	2.80	7.840	0.408
	16	19.84	–3.84	14.745	0.743
High	16	27.30	–11.30	127.690	4.677
	24	21.00	3.00	9.000	0.429
	30	21.70	8.30	68.890	3.175
					16.236

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 16.236 \quad df = (3-1)(3-1) = 4$$

The critical value of χ^2 for 4 df at 0.01 level of significance is 13.277. As the calculated value of χ^2 is greater than the critical value of χ^2, H_0 is rejected. Hence, market share and productivity are related.

Example 14.15 Five coins are tossed 3,200 times and the number of heads appearing each time is noted. At the end, the following results were obtained:

Number of heads	0	1	2	3	4	5
Frequencies	80	570	1100	900	500	50

Use chi-square test of goodness-of-fit to determine whether the coins are unbiased. Use 5 per cent level of significance.

Solution The two hypotheses are:

H_0 : The coins are unbiased.

H_1 : The coins are biased.

In a problem of this type, the distribution of heads will be in accordance with the binomial distribution. For this purpose, we have to calculate expected frequencies as follows:

Table 14.14

No. of heads	$f(x) = {}^nC_x\, p^x\, q^{n-x}$	*Expected frequencies* $= N\,(fx)$
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	$^5C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 = 5\left(\frac{1}{2}\right)^5$	$3200 \times 5 \times \frac{1}{32} = 500$
5	$^5C_5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = 1\left(\frac{1}{2}\right)^5$	$3200 \times \frac{1}{32} = 100$

Having calculated the expected frequencies in this manner, the remaining calculations will be carried out exactly in the same manner as in the earlier problems. This is done by setting up the worksheet as shown below.

Table 14.15 Worksheet for Calculating Chi-square

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
80	100	–20	400	4.0
570	500	70	4900	9.8
1100	1000	100	10000	10.0
900	1000	–100	10000	10.0
500	500	0	0	0
50	100	–50	2500	25.0
				58.8

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$$

$$= 58.8$$

$$df = n - 1 = 6 - 1 = 5$$

The critical value of χ^2 with 5 degrees of freedom at 5% level of significance is 11.07. As the calculated value of χ^2 is greater than the critical value, the null hypothesis is rejected. The conclusion is that the five coins are biased.

Example 14.16 An automobile company gives you the following information about age groups and the liking for a particular model of car that it plans to launch:

Persons	*Below 25*	*25–50*	*Above 50*
Who liked the car	45	30	25
Who disliked the car	55	20	25

On the basis of the above data, can it be concluded that the model appeal is independent of the age group?

Solution

Table 14.16 Worksheet for Calculating Chi-square

	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Below 25	45	50	–5	25	0.5
	55	50	5	25	0.5
25–50	30	25	5	25	1.0
	20	25	–5	25	1.0
Above 50	25	25	0	0	0
	25	25	0	0	0
					3.0

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = \mathbf{3}$$

H_0 : The model appeal is independent of the age group.

H_1 : It is not independent of the age group.

$df = (r-1)(c-1) = (2-1)(3-1) = 1 \times 2 = \mathbf{2}$

The critical value of χ^2 with 2 d.f. and at 5% level of significance is 5.991. As the calculated value of χ^2 < the critical value of χ^2, H_0 is accepted. The conclusion is that the model appeal is independent of the age group.

Example 14.17 Mr Ramana and Miss Lalita, researchers of management department, adopted different sampling techniques while investigating the sample of the students to know the number of students falling in different intelligence levels. The results are:

Researcher	*Below average*	*Average*	*Above average*	*Genius*
Ramana	129	90	66	15
Lalita	80	66	50	4

On the basis of the above data, can we say that the two researchers have adopted significantly different sampling techniques? Test the hypothesis at (a) 0.05 and (b) 0.01 level of significance.

Solution

Table 14.17 Worksheet for Calculating Chi-square

	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
R_1C_1	129	125.4	3.6	12.96	0.10
R_1C_2	90	93.6	–3.6	12.96	0.14
R_1C_3	66	69.6	–3.6	12.96	0.19
R_1C_4	15	11.4	3.6	12.96	1.14
R_2C_1	80	83.6	–3.6	12.96	0.16
R_2C_2	66	62.4	3.6	12.96	0.21
R_2C_3	50	46.4	3.6	12.96	0.28
R_2C_4	4	7.6	–3.6	12.96	1.71
					3.93

Degrees of freedom $(r-1)(c-1) = (2-1)(4-1) = 1 \times 3 = 3$

H_0 : The techniques adopted by the two researchers are not different.

H_1 : They adopted different techniques.

The critical value of $\chi^2_{0.05}$ for 3 *df* is 7.815.

The critical value of $\chi^2_{0.01}$ for 3 *df* is 11.345.

As the calculated value of $\chi^2 = 3.93$ is less than the critical value of $\chi^2_{0.05} = 7.815$ and $\chi^2_{0.01} = 11.345$, the null hypothesis H_0 is not rejected at both levels of significance. In other words, the two researchers have not adopted significantly different techniques.

Example 14.18 On the basis of the following data, ascertain whether there is an association between ownership of computers and income level.

	Own Computers	*Do not Own Computers*	*Total*
Low Income	0	250	250
Middle Income	50	100	150
High Income	80	20	100
	130	**370**	**500**

Solution

Table 14.18 Worksheet for Calculating Chi-square

	O	E	$O–E$	$(O–E)^2$	$(O–E)^2/E$
Own Computers					
L. Income	0	65	–65	4225	65.00
M. Income	50	39	11	121	3.10
H. Income	80	26	54	2916	112.15

(*Contd.*)

(*Contd.*)

Do not Own Computers					
L. Income	250	185	65	4225	22.84
M. Income	100	111	–11	121	1.09
H. Income	20	74	–54	2916	39.41
					χ^2 = **243.59**

Null hypothesis: There is no relationship between ownership of computers and income level.

$$df = (r - 1)(c - 1)$$
$$= (3 - 1)(2 - 1) = 2$$

Critical value of χ^2 @ 0.05 level of significance for 2 *df* is 5.991.

Hence, our null hypothesis that there is no association between ownership of computers and income level is rejected.

Example 14.19 A random sample of 150 college professors was asked to express their opinion on the most important basis for academic promotion. The survey produced the following results:

Teaching Field

Basis of Promotion	*Science*	*Professional*	*Arts*
Research	30	10	20
Teaching	10	20	20
Total performance	10	20	10

Use the chi-square test to verify the hypothesis that the opinion is the same for all faculty groups (at 5% level).

Solution

H_0 : Opinion on the basis for academic promotion is the same for all faculty groups.

H_1 : It is not so.

In order to calculate the expected frequencies, it is necessary to have the row totals and column totals. The table given above is reproduced with these totals.

Table 14.19 Table with Totals

Basis of Promotion	*Science*	*Professional*	*Arts*	*Total*
Research	30	10	20	60
Teaching	10	20	20	50
Total performance	10	20	10	40
Total	**50**	**50**	**50**	**150**

Table 14.20 Worksheet for Calculating Chi-square

	O	*E*	*O–E*	$(O–E)^2$	$(O–E)^2/E$
Science	30	20	10	100	5
	10	6.7	– 6.7	44.89	2.81
	10	13.3	– 3.3	10.89	0.82
Professional	10	20	– 10	100	5
	20	16.7	3.3	10.89	0.65
	20	13.3	6.7	44.89	3.38
Arts	20	20	0	0	0
	20	16.7	3.3	10.89	0.65
	10	13.3	– 3.3	10.89	0.82
				333.34	**19.13**

$$df = (c - 1)(r - 1)$$
$$= (3 - 1)(3 - 1)$$
$$= 2 \times 2 = \mathbf{4}$$

As calculated, $\chi^2 = 19.13$ exceeds the table value of $\chi^2 = 9.488$. Hence, the hypothesis that the opinion is the same for all faculty groups is rejected.

Example 14.20 The demand 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	Tue	Wed	Thu	Fri	Sat
Number of parts demanded	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of the week. Test this at 5% level.

Solution If number of parts demanded does not depend on the day of the week, then each day should have the same number of parts. A chi-square test is required.

Table 14.21 Worksheet for Calculating Chi-square

Day	*O*	*E*	*O–E*	$(O–E)^2$	$(O–E)^2/E$
Mon	1124	1120	4	16	0.014
Tue	1125	1120	5	25	0.022
Wed	1110	1120	– 10	100	0.089
Thu	1120	1120	0	0	0
Fri	1126	1120	6	36	0.032
Sat	1115	1120	– 5	25	0.022
	6720	**6720**			**0.179**

$$\chi^2 = 0.179$$

H_0 : Number of parts demanded does not depend on the day of the week.
H_1 : It is not so.

The critical value of χ^2 at 0.05 level of significance for $n - 1 = 5$ *df* is 11.070. As this is higher than the calculated value of $\chi^2 = 0.179$, H_0 is accepted. The number of parts demanded does not depend on the day of the week.

Example 14.21 The data on income of parents and admission of children in private and government schools is given below:

	Number of Children	
Income of Parents	*Private Schools*	*Govt. Schools*
Low income	494	506
High Income	162	438

Applying the appropriate test, find out whether high income families generally send their children to private schools.

Solution

H_0 : High income families send their children to private schools.
H_1 : It is not so.

Having formulated the hypotheses, we have now to use the chi-square test. In order to use this technique, we have to calculate expected frequencies. For this purpose, we have to provide row totals and column totals to the above table. This has been done in the following table.

Table 14.22 Table Admission of Children in Schools According to Income of Parents

	No. of children in		
Income of parents	*Private Schools*	*Government Schools*	*Total*
Low income	494	506	1000
High income	162	438	600
Total	**656**	**944**	**1600**

Table 14.23 Worksheet for Calculating Chi-square

	O	*E*	*O–E*	$(O–E)^2$	$(O–E)^2/E$
Low Income					
P	494	410	84	7056	17.21
G	506	590	– 84	7056	11.96
High Income					
P	162	246	– 84	7056	28.68
G	438	354	84	7056	19.93
					$\chi^2 \rightarrow$ **77.78**

$$df = (r - 1)\,(c - 1)$$
$$= (2 - 1)\,(2 - 1) = \mathbf{1}$$

The critical value of χ^2 at 0.05 level of significance for 1 degree of freedom is 3.841. As the calculated value of chi-square is greater than the critical value, we reject H_0 that high income families generally send their children to private schools.

Example 14.22 The following table gives the number of good and bad parts produced by each of the three shifts, day, evening and night.

Shift	*Good Parts*	*Bad Parts*	*Total Parts*
Day	900	130	1030
Evening	700	170	870
Night	400	200	600
Total	**2000**	**500**	**2500**

Test the hypothesis that quality of product is independent of shifts at 0.05 level of significance.

Solution

H_0 : Quality of product is independent of shifts.
H_1 : It is not so.

Table 14.24 Worksheet for Calculating Chi-square

	Observed frequency	*Expected frequency*	*O–E*	$(O–E)^2$	$(O–E)^2/E$
R_1C_1	900	824	76	5776	7.01
R_1C_2	130	206	–76	5776	28.04
R_2C_1	700	696	4	16	0.02
R_2C_2	170	174	– 4	16	0.09
R_3C_1	400	480	– 80	6400	13.33
R_3C_2	200	120	80	6400	53.33
					$\chi^2 = 101.82$

$$df = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

As the critical value of χ^2 at 0.05 level of significance for 2 *df* is 5.991 and the calculated $\chi^2 = 101.82$, H_0 is rejected.

Example 14.23 The table given below pertains to a survey conducted to ascertain whether the age of a driver (21 years and above) has any effect on the number of automobile accidents. Test the hypothesis that the number of accidents is independent of the age of the driver. Use the level of significance at (a) 0.05 and (b) 0.01.

	Age of Driver (years)				
Number of accidents	*21–30*	*31–40*	*41–50*	*51–60*	*61–70*
0	30	46	62	50	42
1	16	30	25	24	25
2	12	16	22	17	13
More than 2	10	14	18	16	12

Solution We set up the null and alternate hypotheses.

H_0 : The number of accidents is independent of the age of the driver.

H_1 : It is not so.

It can be seen that the above table does not show the totals. To calculate the expected frequencies, the totals are given below.

Table 14.25 Worksheet for Tables

Columns	*Totals*	*Rows*	*Totals*
1	68	1	230
2	106	2	120
3	127	3	80
4	107	4	70
5	92		**500**
	500		

Table 14.26 Worksheet for Calculating Chi-square

	O	*E*	*O–E*	$(O–E)^2$	$(O–E)^2/E$
R_1C_1	30	31	–1	1	0.03
R_2C_1	16	16	0	0	0
R_3C_1	12	11	1	1	0.09
R_4C_1	10	10	0	0	0
R_1C_2	46	49	–3	9	0.18
R_2C_2	30	25	5	25	1.00
R_3C_2	16	17	–1	1	0.06
R_4C_2	14	15	–1	1	0.07
R_1C_3	62	58	4	16	0.28
R_2C_3	25	30	–5	25	0.83
R_3C_3	22	28	–6	36	1.29
R_4C_3	18	18	0	0	0
R_1C_4	50	49	1	1	0.02
R_2C_4	24	26	–2	4	0.15
R_3C_4	17	17	0	0	0
R_4C_4	16	15	1	1	0.07
R_1C_5	42	42	0	0	0
R_2C_5	25	22	3	9	0.41
R_3C_5	13	15	–2	4	0.27
R_4C_5	12	13	–1	1	0.08
					4.83

On the basis of the above calculations, the value of chi-square = 4.83.

We have to decide the degrees of freedom.

$$df = (r - 1)(c - 1)$$
$$= (4 - 1)(5 - 1)$$
$$= 12$$

The critical value of χ^2 at 0.05 level of significance for 12 df is 21.026. At 0.01 level of significance, the critical value of χ^2 is 26.217.

As the calculated value of χ^2 is less than the critical value of χ^2 at 0.05 and 0.01 levels of significance, H_0 cannot be rejected at either level.

14.5 PRECAUTIONS ABOUT USING THE CHI-SQUARE TEST

In order to use a chi-square hypothesis test properly, one has to be extremely careful and keep in mind certain precautions. First, a sample size should be large enough. If the expected frequencies are too small, the value of χ^2 gets over-estimated. This will result in the rejection of the null hypothesis in several cases. To overcome this problem, we must ensure that the expected frequency in any cell of the contingency table should not be less than 5. In case the expected frequency is below 5 in more than 1 cell, we can combine these to obtain the expected frequency of at least 5. Another point to note is that the calculations must be made with actual numbers in each cell and not with proportions or per centages. If the proportions or percentages were used, then the theoretical distribution would not be applicable.

When the calculated value of χ^2 turns out to be more than the critical or theoretical value at a predetermined level of significance, we reject the null hypothesis. In contrast, when the χ^2 value is less than the critical or theoretical value, the null hypothesis is not rejected. However, when the χ^2 value turns out to be zero, we have to be extremely careful to confirm that there is no difference between the observed and the expected frequencies. Such a situation may sometimes arise on account of a faulty manner used in the collection of data.

In most of the cases, the problems of χ^2 involve simple calculations. However, for large sets of data the chi-square test involves very comprehensive calculations. In all such cases, computer should be used. Several computer Statistics packages contain routines for carrying out chi-square tests.

GLOSSARY

Chi-square distribution	A distribution, with degrees of freedom as the only parameter. It is skewed to the right for small degrees of freedom, but when degrees of freedom are large, it looks like a normal curve.
Chi-square test	A statistical technique used to test significance in the analysis of frequency distribution.
Contingency table	A table having rows and columns wherein each row corresponds to a level of one variable and each column to a level of another variable. The frequencies with which each variable combination has occurred are contained in the body of the table.
Degrees of freedom	The number of elements that can be chosen freely.
Expected frequencies	The frequencies for different categories of a multinomial experiment or for different cells of a contingency table, which are expected to occur on the assumption that the given hypothesis is true.
Goodness-of-fit test	A statistical test involving the chi-square statistic for determining whether some observed pattern of frequencies conforms to an expected pattern.

Observed frequencies	The frequencies actually obtained from the performance of an experiment.
Test of homogeneity	A statistical test involving the chi-square statistic, which verifies whether the proportions of elements belonging to different groups in two (or more) populations are similar or not.
Test of independence	A statistical test involving the chi-square statistic, which verifies whether the two attributes of a population are related or not.

LIST OF FORMULAE

1. Value of the test statistic χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \text{or} \quad \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i = observed frequency in the i^{th} category

E_i = expected frequency in the i^{th} category

k = number of categories

2. Expected frequency for a cell for an independence or homogeneity test.

$$E = \frac{RT \times CT}{N}$$

where E = expected frequency in a given cell

RT = the total of row in which that cell lies

CT = the total of column in which that cell lies

N = total number of observations

3. Number of degrees of freedom: $df = (r - 1)(c - 1)$, where r is the number of rows and c is the number of columns.

4. (a) Another formula for χ^2 where a, b, c and d are the totals of the four cells in a 2×2 contingency table.

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(c + d)(a + b)}$$

(b) Another formula for χ^2 for a 2×3 contingency table

$$\chi^2 = \frac{N}{a+b+c}\left(\frac{a^2}{a+d} + \frac{b^2}{b+e} + \frac{c^2}{c+f}\right) + \frac{N}{d+e+f}\left(\frac{d^2}{a+d} + \frac{e^2}{b+e} + \frac{f^2}{c+f}\right) - N$$

Both these formulas are used with observed frequencies alone, thus avoiding the computation of expected frequencies.

5. Yates' correction:

when results of continuous data are applied to discrete data, Yates' correction is applied by rewriting the chi-square.

$$\chi^2 \text{ (corrected)} = \frac{(|O_1 - E_1| - 0.5)^2}{E_1} + \frac{(|O_2 - E_2| - 0.5)^2}{E_2} + \cdots + \frac{(|O_k - E_k| - 0.5)^2}{E_k}$$

where $|O_k - E_k|$ means that the difference between observed frequency and expected frequency in 'K' category is to be taken ignoring '+' and '–' signs.

QUESTIONS

14.1 Given below are twelve statements. Indicate in each case whether the statement is true or false:

(a) A chi-square test can be used when more than two population proportions are involved in a problem.

(b) A 4×3 contingency table has four columns and three rows.

(c) The expected frequency for any cell in a contingency table should not be less than 5.

(d) A chi-square test can be used when the data are given in proportions or per centages.

(e) When the calculated value of chi-square is less than the theoretical value at a specified level of significance, the null hypothesis is rejected.

(f) On account of simple calculations involved, statisticians very frequently use chi-square tests.

(g) If the expected frequencies are too small, the value of chi-square gets underestimated.

(h) The total area covered by the curve of a chi-square distribution is always 1.

(i) The utility of a chi-square test depends largely on the quality of data used in the test.

(j) The number of degrees of freedom in a chi-square test depends on both the number of rows and the number of columns in a contingency table.

(k) To obtain the expected frequency for any observed frequency is to divide the product of row and column totals by the sample size n.

(l) The degree of freedom and the level of significance are the two factors on which the shape of the chi-square distribution depends.

Multiple Choice Questions (14.2 to 14.11)

14.2 A chi-square value can never be negative because

(a) The observed frequencies cannot be negative.

(b) Differences between observed and expected frequencies are squared.

(c) The sum of the differences is computed.

(d) None of these.

14.3 A contingency table for a chi-square test has 5 columns and 4 rows. How many degrees of freedom should be used?

(a) 20 (b) 9 (c) 15 (d) 12

14.4 In a chi-square contingency table, the expected frequency of a cell can be calculated from the expected proportion for that cell by

(a) Multiplying by the total sample size (b) Multiplying by that row's total

(c) Multiplying by that column's total (d) None of the above

14.5 While performing a chi-square hypothesis, we find that several cells have too small expected frequencies. What will be its impact?

(a) The degrees of freedom are considerably reduced.
(b) The rejection of null hypothesis will be more frequent.
(c) The value of chi-square will be over-estimated.
(d) None of these.
(e) (b) and (c)

14.6 For any given level of significance, the table value of chi-square
(a) declines as degree of freedom increases
(b) increases as degree of freedom increases
(c) increases as sample size increases
(d) declines as sample size increases

14.7 A contingency table should have frequencies in
(a) per centages (b) mean values
(c) proportions (d) frequencies

14.8 The degrees of freedom for a contingency table are on the basis of
(a) $n - 1$ (b) $c - 1$
(c) $r - 1$ (d) $(r - 1)(c - 1)$

14.9 To perform a chi-square test
(a) data conform to a normal distribution
(b) data be measured on a nominal scale
(c) each cell has an equal number of frequencies
(d) all of these

14.10 A contingency table
(a) always has 2 degrees of freedom (b) always has two variables
(c) always has two dependent variables (d) all of these

14.11 The acceptance of a null hypothesis indicates
(a) a correct decision (b) a Type II error
(c) a Type I error (d) either (a) or (c)

14.12 What is a chi-square test? How is it important in dealing with business problems? Give two examples where it can be used.

14.13 Describe the chi-square distribution. What are the parameters on which the shape of a chi-square distribution curve depends?

14.14 What are the properties of the chi-square test?

14.15 Find the value of χ^2 for 12 degrees of freedom and
(a) 0.025 area in the right-tail of the test
(b) 0.995 area in the right-tail of the test
(c) 0.05 area in the left-tail of the test.

14.16 What do you understand by the 'goodness-of-fit test'? Describe three situations where this test can be used appropriately.

14.17 Explain how the chi-square distribution can be used for judging the agreement between a hypothetical and an observed distribution. Show how the degrees of freedom are determined in different circumstances.

14.18 What is the chi-square test for independence? Describe three situations where this test can be used appropriately.

14.19 How do you decide the number of degrees of freedom to be used in a chi-square test? Illustrate your answer with a 3×4 contingency table.

14.20 What precautions would you keep in mind while using the chi-square test?

14.21 Write short notes on:

(i) Yates' corrections for continuity

(ii) Test of homogeneity

(iii) Expected frequencies in a contingency table

14.22 Differentiate between the test of homogeneity and the test of independence in the context of a chi-square test.

14.23 What do you understand by 'the additive property of a χ^2 test'? Explain this with an example.

14.24 In 80 tosses of a coin, 55 heads and 25 tails were observed. Test the hypothesis that the coin is fair using a significance level of (a) 0.05 (b) 0.01.

14.25 In a survey of 200 boys, of whom 75 were intelligent, 40 had skilled fathers; while 85 of the unintelligent boys had skilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys?

14.26 The librarian of a public library gives you the following average number of books borrowed by its members for each working day. The average is based on 10 weeks.

Mon	Tue	Wed	Thu	Fri	Sat
204	292	242	283	252	275

The librarian wants you to let him know whether more books are borrowed on any weekday than on any other. Use 0.05 level of significance.

14.27 In a certain city, a large number of persons suffered from a mysterious disease, which broke out in the form of an epidemic. A survey was undertaken and as a result the following data emerged:

Age Group (in years)	*Number of Patients*
Below 15	15
15–30	25
30–45	20
45–60	30
60 and above	10

Can it be said at 5 per cent level of significance that the incidence of this mysterious disease was equally distributed in all age groups?

14.28 For the data given here, use the chi-square test at 5 per cent level of significance to state whether the two attributes (i.e. condition of home and condition of the child) are independent:

Condition of Child	*Condition of Home*	
	Clean	*Dirty*
Clean	70	50
Fairly clean	80	20
Dirty	45	30

14.29 With a view to sell an article, a person gave an advertisement in three newspapers. In all, he received 54 replies with the break-up as 23, 16 and 15. Is there evidence to suggest that the first newspaper is the best newspaper to advertise in?

14.30 The data, given below, relate to the employment pattern of workers in three firms: A, B and C.

Firm	*A*	*B*	*C*
Unskilled	71	76	33
Skilled manual	50	42	28
Non–manual	79	82	39

Is there a significant difference in the proportions of three types of workers in the three firms?

14.31 A sample analysis of examination results of 200 MBAs was made. It was found that 46 students had failed, 68 secured third division, 62 secured second division and the rest were placed in the first division. Are these figures commensurate with the general examination result that is in the ratio of 2:3:3:2 for various categories respectively?

14.32 Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. Can we say that the sampling techniques adopted by the two researchers are significantly different? Results of the investigation are shown below:

Researcher	*Number of Students in Each Level*			
	Below Average	*Average*	*Above Average*	*Genius*
X	86	60	44	10
Y	40	33	25	2

14.33 The following table shows the break-up of an eight-hour shift into components produced in each of these slots. Use a chi-square test to find out whether a worker is prone to producing defective components throughout an eight-hour shift or not. Use 5 per cent level of significance.

Time Slot (hour)	*Observed Frequency*
8–10	8
10–12	11
12.30–14.30	16
14.30–16.30	15

14.34 A total of 1,600 families were selected at random in a city to test the belief that high income families usually send their children to public schools and low income families often send their children to government schools. The following results were obtained:

Income	*Schools*		
	Public	*Government*	*Total*
Low	494	506	1,000
High	162	438	600
Total	**656**	**944**	**1,600**

Test whether income and type of schooling are independent.

14.35 The theory predicts that the proportion 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 number in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory? Use chi-square test at 1 per cent level of significance.

14.36 The following table has been constructed on the basis of a sample of 220 companies classified by the changes in earnings and different stock exchanges. Examine whether they are independent.

Changes in Earnings	*Stock Exchanges*			
	S_1	S_2	S_3	*Total*
Large decrease	2 (6.65)	17 (14.66)	24 (21.69)	43
Moderate decrease	9 (5.10)	10 (11.25)	14 (16.65)	33
Small changes	11 (9.74)	11 (21.48)	41 (31.78)	63
Moderate increase	6 (7.10)	17 (15.68)	23 (23.22)	46
Large increase	6 (5.41)	20 (11.93)	9 (17.66)	35
Total	**34**	**75**	**111**	**220**

[Figures in the brackets indicate the estimated frequencies under the null hypothesis that the changes in earnings and stock exchanges are independent.]

14.37 A HRD manager is interested in assessing whether there is association between the commuting time of the workers and the level of stress-related problems observed on the job. The table below gives data on the frequencies of a random sample.

Use an appropriate statistical test at a significance level of 5 percent to assess whether there is association between the two factors.

Commuting Time (Min.)	*Stress*		
	High	*Moderate*	*Low*
0 to 30	5	10	10
30 to 60	14	13	17
60 to 90	22	15	11
Over 90	18	9	6

14.38 A certain drug is claimed to be effective in curing cold. In an experiment of 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 the data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills?

14.39 A survey was conducted to know the opinion on adopting presidential system of government by our country. Opinion given by various age–groups is given below:

	Age Groups				
	18–25	*26–35*	*36–45*	*46–55*	*> 55*
Should adopt	250	365	550	285	544
May adopt	155	254	247	321	125
Can't say	55	78	45	85	41
Need not adopt	125	89	101	215	63
Shouldn't adopt	45	25	85	98	101

At 5 per cent level of significance, can we infer that opinion about the system depends upon age?

14.40 The following data show the per centages of firms using computers in different aspects of their business:

Firm Size	*Computers Used in*			
	Admin. (per centage)	*Design (per centage)*	*Manufacture (per centage)*	*Total no. of Firms*
Small	60	24	20	450
Medium	65	30	28	140
Large	90	44	50	45

Test the hypothesis at 0.05 level of significance that there is no association between the size of firm and its use of computers.

14.41 A company uses five machines to manufacture a particular product. The output is then graded into four categories. These data are given below:

Grade	*Observed Values (Machines)*					
	1st	*2nd*	*3rd*	*4th*	*5th*	*Total*
1st	680	500	750	650	620	3,200
2nd	570	350	630	570	480	2,600
3rd	170	130	120	100	80	600
4th	60	50	60	90	40	300
Total	**1,480**	**1,030**	**1,560**	**1,410**	**1,220**	**6,700**

Do you find any difference in the quality of the product from the five machines? Test the hypothesis at **(a)** 0.01 level and **(b)** 0.05 level of significance.

14.42 Of a group of patients who complained that they were suffering from insomnia, some were given sleeping pills while others were given sugar pills (although they all thought they were getting sleeping pills). After two weeks, they were asked whether the pills helped them in getting adequate and sound sleep. The results of their responses are shown in the table given below. Do you think that sleeping pills were helpful in getting good sleep? Test the hypothesis at **(a)** 0.05 level and **(b)** 0.01 level of significance.

	Slept Well	*Did not Sleep Well*
Took sleeping pills	65	20
Took sugar pills	110	75

14.43 A company carried out, repeatedly, experiments that gave the following results:

χ^2	3.7	9.1	14.0	21.4
df	1	4	7	13

What conclusions can be made on the basis of this information?

14.44 Population samples from six small mining camps are given below:

	Camps					
Population	*1*	*2*	*3*	*4*	*5*	*6*
Males	22	12	32	18	46	28
Females	6	8	12	4	24	22

Compute the value of Chi-Square.

14.45 A company has categorised its 80 salesmen as having good and poor selling abilities. It has also given them a psychological test to measure their flexibility. The results are given below:

Selling ability	*Flexibility*	
	Poor	*Good*
Poor	32	8
Good	12	28

Use Chi-square test to identify if the selling ability is independent of flexibility at 5% level of significance.

14.46 What is χ^2 test? For the following data, use χ^2-test to find the superiority of the new treatment.

	Number of patients		*Total*
Treatment	*Favourable*	*Not favourable*	
New	140	30	170
Conventional	60	20	80
Total	**200**	**50**	**250**

14.47 200 digits were selected at random from a set of tables. The frequencies of digits were:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Use χ^2-test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which they were chosen.

14.48 In an experiment on immunization of cattle from tuberculosis, the following results were obtained:

	Affected	*Not affected*
Inoculated	12	26
Not inoculated	16	6

Calculate χ^2 and discuss the effect of vaccine in controlling susceptibility to tuberculosis (5% value of χ^2 for one degree of freedom = 3.84).

14.49 The research staff of the Credit and Collection Department of an Insurance Company wish to determine whether or not the type of printing used for collection reminders for annual premium of policies has a significant influence upon the member of immediate responses. They believe that personal attention may have an important influence upon the responses. To test their belief, 4000 reminders are mailed to a random sample of premium defaulters. Of these, 1600 are obviously mimeographed standard forms with no attempt to personalise the message, 600 are computer-printed reminders with the customer's name printed on standard form and the others are actual, originally typed personal letters. Those which generated immediate response included 600 of the mimeographed forms, 600 of the computer-printed reminders, and 400 of the individually typed letters.

What can the research staff conclude from these responses?

14.50 We wish to study if the preference of a customer for our brand of shampoo depends on his or her income level using a significance level of 0.05. Base your calculations on the data given below for a survey conducted on 350 consumers.

	Category of preference			
Income level	*Strongly prefer*	*Moderately prefer*	*Indifferent*	*Do not prefer*
High	15	35	21	27
Medium	48	18	20	22
Low	32	66	18	28

14.51 A researcher has collected data summarizing the number of respondents in his study under each combination of the level of income and the level of qualification, as shown in the following table. Check whether the income is independent of the qualification, while grouping the respondents, at significance levels of (a) 0.05 and (b) 0.01.

Level of income	*Level of qualification*		
	Diploma	*Undergraduation*	*Postgraduation*
Low	25	55	15
Medium	60	70	25
High	50	80	75

14.52 Weights in kg of 10 students are given below:

38 40 45 53 47 43 55 48 52 49

Can we say that the variance of distribution of weights of all students, from which the above sample of 10 students was drawn, is equal to 20 square kg?

14.53 The following table gives the classification of 100 workers according to sex and nature of work.

Test whether the nature of work is independent of the sex of the worker at $\alpha = 0.05\%$.

Sex	*Nature of work*	
	Stable	*Unstable*
Males	40	20
Females	10	30

14.54 The following contingency table presents the reactions of legislators to a tax plan, according to party affiliation. Test whether party affiliation influences the reaction to the tax plan.

Party	*Reaction*			
	In favour	*Neutral*	*Opposed*	*Total*
Party A	120	20	20	160
Party B	50	30	60	140
Party C	50	10	40	100
Total	**220**	**60**	**120**	**400**

14.55 4 coins were tossed 160 times and the following results were obtained:

Number of heads	0	1	2	3	4
Observed frequencies	17	52	54	31	6

Under the assumption that the coins are balanced, find the expected frequencies of setting 0, 1, 2, 3 and 4 heads and test the goodness of the fit.

The McGraw-Hill Companies

CHAPTER 15 ANALYSIS OF VARIANCE

Learning Objectives

By the end of your work on this chapter, you should be able to

- compare more than two population means using analysis of variance
- understand the assumptions involved in the use of analysis of variance technique.

Chapter Prerequisites

Before starting work on this chapter, make sure that you are quite familiar with

1. the procedure used in testing hypotheses
2. the computation of standard deviation and variance

15.1 INTRODUCTION

In Chapters 12 and 13, we studied inference methods (confidence intervals and hypothesis testing) concerning a single population mean. In Chapter 14, we discussed the comparison of two population means. In this chapter, we develop a method for comparing several population means at the same time. This method is known as *Analysis of Variance* though its abbreviated form 'ANOVA' is more frequently used.

Evolution of ANOVA

The analysis of variance was developed by R.A. Fisher. In the early stages of its evolution, it was mainly used in agricultural research. As such, the terminology used in analysis of variance is loaded with agricultural terms such as blocks (meaning land) and treatments (meaning varieties of seed, fertilizers or methods of cultivation). In course of time, the technique of analysis of variance began to be used in other spheres too. Today, it is widely used in many different fields. For example, in manufacturing industries, quality control engineers can use it to compare the production from different assembly lines or different plants.

The analysis of variance tests are performed using the F-distribution, which was discussed in Chapter 13.

15.2 ASSUMPTIONS OF ANALYSIS OF VARIANCE

Before we discuss the procedure of carrying out an analysis of variance, it is necessary to know the assumptions implicit in its use. There are three assumptions as stated on the next page.

1. The data are quantitative in nature and are normally distributed. In case of nominal or ordinal data where results are given in per centages or ranks, analysis of variance should not be carried out. If the data are not exactly normally distributed but are close to a normal distribution, analysis of variance still gives good results. One should ensure that data are not highly skewed.
2. The second assumption is that the samples are drawn from the population randomly and independently. In case of an experimental design, the treatments must be assigned to test units by means of some randomising device. In case this assumption does not hold, inferences drawn from ANOVA will not be valid.
3. The third assumption is homogeneity of variance, which means that the variances of the population from which samples have been drawn are equal. In order to combine the variances within the groups into a single within-group sources of variation (SSW), this assumption must be satisfied. However, non-compliance of this assumption may not seriously affect the inferences based on the *F*-distribution provided that the sample from each group is of equal size.

When the distribution is highly skewed and population variances are not approximately equal, then ANOVA should not be used. A non-parametric technique known as the *Kruskal-Wallis test*, can be used. This alternative technique is described in Chapter 20, which is devoted to non-parametric tests.

15.3 NOTATIONS AND BASIC CONCEPTS

Having looked into the assumptions of the analysis of variance, we now introduce the notations that are commonly used in problems involving its use. This can be done by giving a hypothetical example.

Example 15.1 Suppose there are three groups of students with marks (given out of 20) as shown in Table 15.1.

Table 15.1 Three Groups of Students Along with their Marks

Group 1	*Group 2*	*Group 3*
16	15	15
17	15	14
13	13	13
18	17	14
64	**60**	**56**

We have to show this table using symbols that are used in ANOVA as also to explain this transformation.

Solution In terms of symbols, the table can be written as follows:

Group 1	*Group 2*	*Group 3*
x_{11}	x_{21}	x_{31}
x_{12}	x_{22}	X_{32}
x_{13}	x_{23}	x_{33}
x_{14}	x_{24}	X_{34}
Σx_{1j}	Σx_{2j}	Σx_{3j}
$\bar{x}_{1.}$	$\bar{x}_{2.}$	$\bar{x}_{3.}$

In the above table, the data have been shown columnwise for the three groups. Alternatively, the same data can be shown in three rows as follows:

Group 1	x_{11}	x_{12}	x_{13}	x_{14}
Group 2	x_{21}	x_{22}	x_{23}	x_{24}
Group 3	x_{31}	x_{32}	x_{33}	x_{34}

In general, a value is shown by $x_{ij,}$ which indicates that it is the j^{th} value in the i^{th} group. The sum of the values of the first group is shown by

$$\sum_{j=1}^{n_1} x_{1j} = x_{11} + x_{12} + x_{13} + x_{14} = 64$$

where n_1 shows the size of the first group, which is 4. In the same way, the other two groups are shown:

$$\sum_{j=1}^{n_2} x_{2j} = x_{21} + x_{22} + x_{23} + x_{24} = 60$$

$$\sum_{j=1}^{n_3} x_{3j} = x_{31} + x_{32} + x_{33} + x_{34} = 56$$

The sum of all 12 values is shown by

$$\sum_{i=1}^{3} \sum_{j=1}^{n_i} x_{ij} = \sum_{j=1}^{n_1} x_{1j} + \sum_{j=1}^{n_2} x_{2j} + \sum_{j=1}^{n_3} x_{3j} = 64 + 60 + 56 = 180$$

The arithmetic mean of the groups are

$$\bar{x}_{1.} = 1/n_1 \sum_{j=1}^{n_1} x_{1j} = ¼(64) = 16$$

$$\bar{x}_{2.} = 1/n_2 \sum_{j=1}^{n_2} x_{2j} = ¼(60) = 15$$

$$\bar{x}_{3.} = 1/n_3 \sum_{j=1}^{n_3} x_{3j} = ¼(56) = 14$$

or in general

$$\bar{x}_{i.} = 1/n_i \sum_{j=1}^{n_i} x_{1j}$$

The arithmetic mean of all 12 values is

$$\bar{x} = 1/n \sum_{i=1}^{3} \sum_{j=1}^{4} x_{ij} = 1/12\ (64 + 60 + 56) = 180/12 = 15$$

$$n = \Sigma n_i = n_1 + n_2 + n_3 = 4 + 4 + 4 = 12$$

15.4 ONE-WAY CLASSIFICATION

The simplest form of analysis of variance is a one-way model, which we use with simple random samples in order to compare the effect of a single independent variable on the dependent variable. To illustrate one-way ANOVA, let us take an example. To simplify the discussion, we shall consider the case in which samples are of equal size.

Example 15.2 Let us take the same data as given in Table 15.1. Suppose there are three different methods of teaching English that are used on three groups of students. We are interested to know whether these different methods of teaching had an effect on the performance of students. Random samples of size 4 are taken from each group and the marks obtained by the sample students in each group are given in Table 15.2 (which is the same as Table 15.1).

Table 15.2 Marks Obtained by Students

Group A	*Group B*	*Group C*
16	15	15
17	15	14
13	13	13
18	17	14
64	**60**	**56**

Using ANOVA find out whether teaching methods had any effect on the students' performance.

Solution It is assumed that the marks obtained by students are distributed normally with means μ_1, μ_2 and μ_3 for the three groups A, B and C, respectively. Further, it is assumed that the standard deviations of the distribution of marks for groups A, B and C are equal and constant. This assumption implies that the mean marks of the groups may differ on account of using different methods of teaching, but they do not affect the dispersion of the marks. This means that the location of the distributions has merely shifted but the dispersion in each case remains the same.

Procedure for ANOVA

In order to test about the equality of means of three populations, we follow the five-step procedure as explained below:

Step 1 State the null and alternative hypotheses.

$H_0 : \mu_1 = \mu_2 = \mu_3$ (The means for three groups are equal.)
H_1 : At least one pair is unequal.

Note that the alternative hypothesis states that the mean marks of students of at least one of the three groups are different from the mean marks obtained by students of the other two groups.

Step 2 Select the test criterion to be used.

We have to decide which test criterion or distribution should be used. As our comparison involves means for three normally distributed populations, we should use the *F*-distribution to test the hypothesis.

Step 3 Determine the rejection and non-rejection regions.

We decide to use 0.05 level of significance. As a one-way ANOVA test is always right-tail, the area in the right-tail of the *F*-distribution curve is 0.05, which is the rejection region. Now, we need to know the degrees of freedom for the numerator and the denominator. Degrees of freedom for the numerator = $k - 1$, where k is the number of groups, suggesting there are different methods of teaching English. Degrees of freedom for the denominator = $n - k$, where n is the total number of observations. In our example, degrees of freedom for the numerator are $k - 1 = 3 - 1 = 2$ and for the denominator are $n - k = 12 - 3 = 9$.

Having ascertained the degrees of freedom (*df*), we now find the critical value of *F* for 2 *df* for the numerator and 9 *df* for the denominator, and 0.05 area in the right-tail of the *F*-distribution curve. The required value of *F* is 4.26 (Appendix Table 6). This implies that if the calculated value of *F* is less than 4.26, H_0 cannot be rejected. If the calculated value of *F* turns out to be higher than the critical value of *F* (4.26), we will reject H_0.

Step 4 Calculate the value of the test statistic.

Now, for the given data we have to calculate the test statistic *F*. Table 15.3 shows the calculations.

Applying ANOVA Technique

Having listed the four steps involved in the analysis of variance, we revert to the data given in Table 15.2.

Table 15.3 Worksheet for Calculating Variances

Group A			*Group B*			*Group C*		
x_{1j}	$x_{1j} - \bar{x}_1$	$(x_{1j} - \bar{x}_1)^2$	X_{2j}	$X_{2j} - \bar{x}_2$	$(x_{2j} - \bar{x}_2)^2$	X_{3j}	$X_{3j} - \bar{x}_3$	$(x_{3j} - \bar{x}_3)^2$
16	0	0	15	0	0	15	1	1
17	1	1	15	0	0	14	0	0
13	–3	9	13	–2	4	13	–1	1
18	2	4	17	2	4	14	0	0
64		**14**	**60**		**8**	**56**		**2**
Mean 16			***Mean 15***			***Mean 14***		

The sample variances for the groups are:

$$S_1^2 = 1/n_1 \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2 = ¼ (14) = 3.5$$

$$S_2^2 = 1/n_2 \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 = ¼ (8) = 2$$

$$S_3^2 = 1/n_3 \sum_{j=1}^{n_3} (x_{3j} - \bar{x}_3)^2 = ¼ (2) = 0.5$$

We can now estimate the variance by the pooled variance method as follows:

$$\hat{\sigma}^2 = \frac{\Sigma\Sigma(x_{ij} - x_i)^2}{n - 3}$$

The numerator can be written as

$$\sum_i \sum_j (\Sigma x_{ij} - \bar{x}_i)^2 = \sum_j (x_{1j} - \bar{x}_1)^2 + \sum_j (x_{2j} - \bar{x}_2)^2 + \sum_j (x_{3j} - \bar{x}_3)^2$$

The denominator can be written as

$$n_1 + n_2 + n_3 - 3$$

Applying the values in the above formulas, both the numerator and denominator

$$= (14 + 8 + 2)/(4 + 4 + 4 - 3) = 2.67$$

This is the variance within the samples.

Now, we can calculate variance between the samples by using the formula,

$$\hat{\sigma}^2 = 1/(3-1) \sum_{i=1}^{3} n_i (\bar{x}_i - \bar{x})^2$$

Since $n_1 = n_2 = n_3 = 4$, this could be written as

$$\frac{4[(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2 + (\bar{x}_3 - \bar{x})^2]}{3-1}$$

$$= \frac{4[(16-15)^2 + (15-15)^2 + (14-15)^2]}{3-1}$$

$$= \frac{(4)(1+0+1)}{3-1} = 8/2 = 4 \text{ (This is the variance between the samples.)}$$

Now, F is to be calculated. F = ratio of two variances

$$= \frac{\text{Estimate of } \sigma^2 \text{ between samples}}{\text{Estimate of } \sigma^2 \text{ within samples}} = \frac{4}{2.67} = 1.498$$

The foregoing calculations can be summarised in the form of an ANOVA table. Table 15.4 is an ANOVA table in general form and Table 15.5 is the ANOVA table giving the calculations done earlier.

Table 15.4 ANOVA Table in General Form

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*Variance Ratio F*
Between samples	*SSB*	$k-1$	MSB = SSB/($k-1$)	
Within samples	*SSW*	$n-k$	MSW = SSW/($n-k$)	F = MSB/MSW
Total	**SST**	$n-1$		

Table 15.5 ANOVA Table for Example 15.2

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*Variance Ratio F*
Between samples	8	3 – 1	8/2 = 4	
Within samples	24	12 – 3	24/9 = 2.67	4/2.67 = 1.498
Total	**32**	**12 – 1**	**32/11 = 2.9**	

Step 5 This is the final step wherein a decision is to be made. The critical value of F for 2 and 9 degrees of freedom at 5 per cent level of significance is 4.26 (Appendix Table 6). As the calculated value of $F = 1.498$ is less than the critical value of F (2, 9) for 0.05 level of significance being 4.26, it falls in the non-rejection region. Hence, the null hypothesis cannot be rejected. In other words, the difference in the three means obtained from the three samples could have arisen due to fluctuations of random sampling.

It will be better if a summary of the estimates is given here.

Summary of the Estimates

To summarise, we have the following three estimates:

1. Using the 'within'deviations $\hat{\sigma}^2 = 2.67$
2. Using the 'between' deviations $\hat{\sigma}^2 = 4$
3. Using the 'total' deviations $\hat{\sigma}^2 = 2.9$
4. We also see

   $\Sigma\Sigma (x_{ij} - \bar{x})^2 = 32$

   $\Sigma\Sigma (x_{ij} - \bar{x}_i)^2 = 24$

   $\Sigma n_i (x_i - \bar{x})^2 = 8$

   We can, therefore, write the fundamental identity as

   $\Sigma\Sigma (x_{ij} - \bar{x})^2 = \Sigma\Sigma (x_{ij} - \bar{x}_i)^2 + \Sigma n_i (x_i - \bar{x})^2$

   $32 = 24 + 8$

or Total variance = Variance within samples + Variance between samples. This means if any two values are known to us, the third value can be easily obtained.

15.5 TWO-WAY CLASSIFICATION

We now turn to the application of analysis of variance in a two-way classification. In such a classification, the data are classified according to two criteria or factors. There is a little difference in the procedure followed for analysis of variance in a two-way classification as compared to that used in a one-way classification. The analysis of variance table takes the form in a two-way classification as shown in Table 15.6.

Table 15.6 ANOVA Table in General Form (Two-way Analysis of Variance)

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*
Between columns	SSC	$(c - 1)$	MSC = SSC/$(c - 1)$
Within rows	SSR	$(r - 1)$	MSR = SSR/$(r - 1)$
Residual (error)	SSE	$(c - 1)(r - 1)$	MSE=SSE/$(c - 1)(r - 1)$
Total	**SST**	**cr – 1**	

The abbreviations used in the table are:

SSC = sum of squares between columns

SSR = sum of squares between rows

SST = total sum of squares

SSE = sum of squares of error (residual). It is obtained by subtracting SSR *and* SSC *from* SST

(c –1) indicates the number of degrees of freedom between columns

(r – 1) indicates the number of degrees of freedom between rows

(c – 1)(r – 1) indicates the number of degrees of freedom for residual
MSC = mean of sum of squares between columns
MSR = mean of sum of squares between rows
MSE = mean of sum of squares for residual
It may be noted that the total number of degrees of freedom are = (c – 1) + (r – 1) + (c – 1)(r – 1) = cr – 1 = N – 1.

Let us now take an example to illustrate the use of this table in the two-way analysis of variance.

Example 15.3 A company has appointed four salesmen, A, B, C and D, and observed their sales in three seasons—summer, winter and monsoon. The figures (in Rs lakh) are given in the following table:

	Salesmen				
Seasons	*A*	*B*	*C*	*D*	*Season Totals*
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Salesmen Totals	90	93	81	96	360

Using 5 per cent level of significance, perform an analysis of variance on the above data and interpret the results.

Solution We proceed with this problem step-by-step as follows:

Step 1 We set up the hypotheses in respect of seasons

H_0 : There is no difference in the mean sales in the three seasons, that is, $\mu_1 = \mu_2 = \mu_3$.

H_1 : There is difference in the mean sales in the three seasons.

We set up the hypotheses in respect of salesmen.

H_0 : There is no difference in the mean sales performance of A, B, C and D.

H_0 : There is difference in the mean sales performance of A, B, C and D.

Step 2 We use the *F*-distribution to test the above hypotheses.

Step 3 We need to determine the rejection and non-rejection region.

We decide the level of significance at 5 per cent. The degrees of freedom for rows are $(r - 1) = 2$ and for columns are $(c - 1) = 3$ and for residual $(r - 1)(c - 1) = 2 \times 3 = 6$. Thus, we have to compare the calculated value of *F* with the critical value of *F* for (a) 2 and 6 *df* at 5 per cent level of significance, and (b) 3 and 6 *df* at 5 per cent level of significance.

Step 4 Necessary calculations are to be done. These are shown in Table 15.7. In order to simplify calculations, it is preferable to reduce the magnitude of the figures given in the problem. We, therefore, subtract 30 from each figure. The table in the coded form is as follows:

Table 15.7 Coded Data for ANOVA

	Salesmen				
Seasons	*A*	*B*	*C*	*D*	*Season Totals*
Summer	6	6	–9	5	8
Winter	–2	–1	1	2	0
Monsoon	–4	–2	–1	–1	–8
Salesmen Totals	0	3	–9	6	0

Correction factor = $C = T^2/N = (0)^2/12 = 0$

Sum of squares between salesmen

$= 0^2/3 + 3^2/3 + (-9^2/3) + 6^2/3 = 0 + 3 + 27 + 12 = 42$

Sum of squares between seasons

$= 8^2/4 + 0^2/4 + (-8^2/4) = 16 + 0 + 16 = 32$

Total sum of squares

$= (6)^2 + (-2)^2 + (-4)^2 + (6)^2 + (-1)^2 + (-2)^2 + (-9)^2 + (1)^2 + (-1)^2 + (5)^2 + (2)^2 + (-1)^2$

$= 36 + 4 + 16 + 36 + 1 + 4 + 81 + 1 + 1 + 25 + 4 + 1$

$= 210$

We may now show these calculations in analysis of variance table.

Table 15.8 Analysis of Variance Table

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*
Between columns (salesmen)	42	4 – 1 = 3	14.00
Within rows (seasons)	32	3 – 1 = 2	16.00
Residual (error)	136	3 × 2 = 6	22.67
Total	**210**	**12 – 1 = 11**	

We now test the hypothesis **(i)** that there is no difference in the sales performance among the four salesmen and **(ii)** there is no difference in the mean sales in the three seasons. For this, we have to first compare the salesmen variance estimate with the residual estimate. This is shown below:

F_A 14/ 22.67 = 0.62

In the same manner, we have to compare the season variance estimate with the residual variance estimate. This is shown below:

F_B 16/22.67 = 0.71

Step 5 It may be noted that the critical value of *F* for 3 and 6 degrees of freedom at 5 per cent level of significance is 4.76 (Appendix Table 6). Since the calculated value of F_A is 0.62 is less than the critical value we can conclude that there is no significant difference among salesmen. Again the critical value of *F* for 2 and 6 degrees of freedom at 5 per cent level of significance is 5.14. Since the calculated value of F_B = 0.71 is less than the critical value of 5.14, we can conclude that there is no difference in the sales in different seasons. The overall conclusion is that the salesmen and the seasons are alike in respect of sales.

Additional Examples

Example 15.4 Consider the following ANOVA table, based on information obtained for three randomly selected samples from three independent populations, which are normally distributed with equal variances.

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*Value of the Test Statistic*
Between samples			20	
Within samples	60			*F* =
Total		**14**		

(a) Complete the ANOVA table by filling in missing values.
(b) Test the null hypothesis that the means of the three populations are all equal, using 0.01 level of significance.

Solution

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*Value of the Test Statistic*
Between samples	(40)	(2)	20	
Within samples	60	(12)	(5)	F = 20/5 = (4)
Total	**(100)**	**14**		

(a) The missing values are shown in the table within circles. As there are three randomly selected samples, *df* between samples will be 3 – 1 = 2.
(b) The critical value of F at 0.01 level of significance is 6.93. As the calculated value of F is less than the critical value, the null hypothesis that the means of the three populations are equal, is not rejected.

Example 15.5 Consider H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

(a) Complete the following ANOVA table.

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*F*
Between samples	132			
Within samples				
Total	**532**	**44**		

(b) Test the hypothesis at (i) 5 per cent level of significance (ii) 1 per cent level of significance and interpret the results.

Solution

ANOVA Table

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*F*
Between samples	132	(4)	33	
Within samples	(400)	(40)	(10)	33/10 = (3.3)
Total	**532**	**44**		

(a) Missing values are shown in the table within circles. It may be noted that as there are 5 samples, *df* between samples will be 5 – 1 = 4.
(b) The critical value of F at 5 per cent level of significance is 2.61. As the calculated value of F is greater than the critical value, hypothesis H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ is rejected.

The critical value of F at 1 per cent level of significance is 3.83. As the calculated value of F is less than the critical value, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ is accepted.

Example 15.6 A company has devised three training methods to train its workers. It is keen to know which of these three training methods would lead to greatest productivity after training. Given below are productivity measures for individual workers trained by each method.

Method 1	30	40	45	38	48	55	52
Method 2	55	46	37	43	52	42	40
Method 3	42	38	49	40	55	36	41

Find out whether the three training methods lead to different levels of productivity at the 0.05 level of significance.

Solution

H_0 : The three training methods lead to the same level of productivity.

H_1 : The three training methods lead to different levels of productivity.

We now set up a worksheet to carry out calculations.

Worksheet

Method 1			*Method 2*			*Method 3*		
X_1	$X_1 - \bar{X}$	$(X_1 - \bar{X})^2$	X_2	$X_2 - \bar{X}$	$(X_2 - \bar{X})^2$	X_3	$X_3 - \bar{X}$	$(X_3 - \bar{X})^2$
30	–14	196	55	10	100	42	–1	1
40	–4	16	46	1	1	38	–5	25
45	1	1	37	–8	64	49	6	36
38	–6	36	43	–2	4	40	–3	9
48	4	16	52	7	49	55	12	144
55	11	121	42	–3	9	36	–7	49
52	8	64	40	–5	25	41	–2	4
308		450	315		252	301		268

$$\bar{X}_1 = 308/7 = 44 \quad \bar{X}_2 = 315/7 = 45 \quad \bar{X}_3 = 301/7 = 43$$

The sample variances for the three methods are:

$$S_1^2 = \frac{1}{n_1}\sum_{j=1}^{n_1}(X_{1j} - \bar{X}_1)^2 = \frac{1}{7}(450) = 64.29$$

$$S_2^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2 = \frac{1}{7}(252) = 36$$

$$S_3^2 = \frac{1}{n_3}\sum_{j=1}^{n_3}(X_{3j} - \bar{X}_3)^2 = \frac{1}{7}(268) = 38.29$$

We now estimate the variance by the pooled variance method as follows:

$$\hat{\sigma}^2 = \frac{\Sigma\Sigma(X_{ij} - X_i)^2}{n-3} = \frac{450 + 252 + 268}{7 + 7 + 7 - 3} = \frac{970}{18} = 53.89$$

This is the variance within the samples.

Now, we calculate variance between the samples by using the formula,

$$\hat{\sigma}^2 = \frac{1}{(3-1)} \sum_{i=1}^{3} n_i (\bar{X}_i - \bar{X})^2$$

$$= \frac{7[(44-44)^2 + (45-44)^2 + (43-44)^2]}{3-1}$$

$$= \frac{(7)(0+1+1)}{2} = \frac{14}{2} = 7 \quad \text{Variance between the samples.}$$

Now, F ratio is to be calculated.

$$F = \frac{\text{Estimate of } \sigma^2 \text{ between samples}}{\text{Estimate of } \sigma^2 \text{ within samples}}$$

$$= \frac{7}{53.89} = 0.13$$

df between samples $3 - 1 = 2$

df within samples $21 - 3 = 18$

The critical value of $F(2, 18)$ at 0.05 level of significance is 3.55. As the calculated value of F is less than the critical value of F, the null hypothesis is accepted. Hence, we conclude that the three methods of training lead to the same level of productivity.

Example 15.7 The following represent the number of units of production per day turned out by four different workers using five different types of machines:

	Machine Type					
Worker	*A*	*B*	*C*	*D*	*E*	*Total*
1	4	5	3	7	6	25
2	5	7	7	4	5	28
3	7	6	7	8	8	36
4	3	5	4	8	2	22
Total	**19**	**23**	**21**	**27**	**21**	**111**

On the basis of this information, can it be concluded that **(i)** the mean productivity is the same for different machines, **(ii)** the workers don't differ with regard to productivity?

Solution

(i) H_0 : The mean productivity is the same for different machines, i.e. $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

(ii) H_0 : The workers don't differ with regard to productivity.

First we take correction factor:

$$\text{Correction factor} = \frac{T^2}{n} = \frac{(111)^2}{20} = 616.05$$

Sum of squares between machines:

$$= \frac{(19)^2}{4} + \frac{(23)^2}{4} + \frac{(21)^2}{4} + \frac{(27)^2}{4} + \frac{(21)^2}{4} - \frac{T^2}{n}$$

$$= 90.25 + 132.25 + 110.25 + 182.25 + 110.25 - 616.05$$

$$= 9.2$$

Degrees of freedom, $df = (c - 1) = (5 - 1) = 4$

Sum of squares between workers:

$$= \frac{(25)^2}{5} + \frac{(28)^2}{5} + \frac{(36)^2}{5} + \frac{(22)^2}{5} - \frac{T^2}{n}$$

$$= 125 + 156.8 + 259.2 + 96.8 - 616.05$$

$$= 21.75$$

Degrees of freedom, $df = (r - 1) = (4 - 1) = 3$

Total sum of squares

$$= (4)^2 + (5)^2 + (7)^2 + (3)^2 + (5)^2 + (7)^2 + (6)^2 + (5)^2 + (3)^2 + (7)^2 + (7)^2 + (4)^2 + (7)^2 + (4)^2 + (8)^2 + (8)^2 + (6)^2 + (5)^2 + (8)^2 + (2)^2 - 616.05$$

$$= 16 + 25 + 49 + 9 + 25 + 49 + 36 + 25 + 9 + 49 + 49 + 16 + 49 + 16 + 64 + 64 + 36 + 25 + 64 + 4 - 616.05$$

$$= 62.95$$

Residual = Total sum of square – Sum of squares between machines – Sum of squares betweens workers

$$= 62.95 - 9.2 - 21.75 \quad = 32$$

Degrees of freedom, $df = (c - 1)(r - 1)$

$$= (5 - 1)(4 - 1) = 4 \times 3 = 12$$

Now, we set up the ANOVA table.

ANOVA Table

Sources of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Squares*	*Variance ratio 'F'*
Between machines	9.2	4	2.3	2.3/2.67 = 0.86
Between workers	21.75	3	7.25	7.25/2.67 = 2.72
Residual	32	12	2.67	
Total	62.95	19		

$F_{0.05}$ for 4 and 12 degrees of freedom = 3.26. As the calculated value of F is less than the critical value of F, null hypothesis is accepted. This means that there is no significant difference in the mean productivity of five different machines.

$F_{0.05}$ for 3 and 12 degrees of freedom = 3.49. Again, the calculated value of F is less than the critical value of F, hence the null hypothesis is accepted. The conclusion is that there is no significant difference in the mean productivity of four different workers.

Example 15.8 There are four classes using different methods of programmed learning of Business Statistics. All the four classes were given an identical test and the students were graded on a 10-point basis. Samples of size 5 were drawn from each class. The data are as follows:

Methods	*Grades*				
I	3	4	3	2	3
II	3	6	6	7	3
III	5	7	8	8	7
IV	9	8	9	9	10

Determine whether there is a significant difference in the results of the different methods.

Solution

The null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

The alternative hypothesis is $H_1 : \mu_i$ not all equal

A_1	*Square*	A_2	*Square*	A_3	*Square*	A_4	*Square*
3	9	3	9	5	25	9	81
4	16	6	36	7	49	8	64
3	9	6	36	8	64	9	81
2	4	7	49	8	64	9	81
3	9	3	9	7	49	10	100
15	225	25	625	35	1225	45	2025

$$C = \frac{T^2}{n} = \frac{(15 + 25 + 35 + 45)^2}{20} = \frac{14400}{20} = 720$$

Total sum of squares = $\Sigma\Sigma x_{ij}^2 - C = 844 - 720 = 124$

Sum of squares between columns

$$\frac{\Sigma x_i^2}{n_i} - C = \frac{1}{5}(225 + 625 + 1225 + 2025) - 720$$

$$= \frac{1}{5}(4100) - 720$$

$$= 820 - 720 = 100$$

Sum of squares between rows SSR

SST– SSC = 124 – 100 = 24

ANOVA Table

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Between columns	100	4 – 1 = 3	100/3 = 33.3
Between rows	24	20 – 4 = 16	24/16 = 1.5
Total	**124**	***n* – 1 = 20 – 1 = 19**	

$$F = \frac{33.3}{1.5} = 22.2$$

The critical value of $F(3, 16)$ at 0.05 level of significance is 2.24. As the calculated value of F is greater than the critical value of F, we reject H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ and accept H_1 : μ_i are not all equal. This means there is a significant difference in the results of different methods.

Example 15.9 Four experiments determine the moisture content of samples of a powder, each man taking a sample from each of the six consignments. Their assessments are given as follows:

	Consignments					
Observers	*1*	*2*	*3*	*4*	*5*	*6*
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Analyse the data and discuss whether there is any significant difference between consignments or between observers.

Solution

	Consignments						
Observers	*1*	*2*	*3*	*4*	*5*	*6*	*Total*
1	9	10	9	10	11	11	60
2	12	11	9	11	10	10	63
3	11	10	10	12	11	10	64
4	12	13	11	14	12	10	72
Total	**44**	**44**	**39**	**47**	**44**	**41**	**259**

Correction factor: $\frac{T^2}{n} = \frac{(259)^2}{24} = 2795$

Sum of squares between consignments

$$= \frac{(44)^2}{4} + \frac{(44)^2}{4} + \frac{(39)^2}{4} + \frac{(47)^2}{4} + \frac{(44)^2}{4} + \frac{(41)^2}{4} - \frac{T^2}{n}$$
$$= 484 + 484 + 380.25 + 552.25 + 484 + 420.25 - 2795$$
$$= 9.75$$

Degrees of freedom: $(c - 1) = (6 - 1) = 5$

Sum of squares between observers

$$= \frac{(60)^2}{6} + \frac{(63)^2}{6} + \frac{(64)^2}{6} + \frac{(72)^2}{6} - \frac{T^2}{n}$$
$$= 600 + 661.5 + 682.67 + 864 - 2795$$
$$= 13.17$$

Degrees of freedom: $(r - 1) = (4 - 1) = 3$

Total sum of square

$$= (9)^2 + (12)^2 + (11)^2 + (12)^2 + (10)^2 + (11)^2 + (10)^2 + (13)^2 + (9)^2 + (9)^2 + (10)^2 + (11)^2 + (10)^2 + (11)^2 + (12)^2 + (14)^2 + (11)^2 + (10)^2 + (11)^2 + (12)^2 + (11)^2 + (10)^2 + (10)^2 + (10)^2$$

$= 81 + 144 + 121 + 144 + 100 + 121 + 100 + 169 + 81 + 81 + 100 + 121 + 100 + 121 + 144 + 196 + 121 + 100 + 121 + 144 + 121 + 100 + 100 + 100 - 2795$

$= 36$

Residual = Total sum of squares – Sum of squares between consignments – Sum of squares between observers

$= 2795 - 9.75 - 13.17$

$= 2772.08$

Degrees of freedom $= (c - 1)(r - 1) = (6 - 1)(4 - 1)$

$= 5 \times 3 = 15$

ANOVA Table

Source of Variation	*Sum of Squares*	*df*	*Mean Squares*	*Variance ratio (F)*
Between columns	9.75	5	1.95	1.95/0.872 = 2.24
Between rows	13.17	3	4.39	4.39/0.872 = 5.03
Residual	13.08	15	0.872	
Total	**36**	**23**		

H_0 : There is no significant difference between consignments.

H_1 : There is significant difference between consignments.

$F(5, 15)$ at 0.05 level of significance = 2.9

As the calculated value of F is less than the critical value, H_0 is accepted. Hence, there is no significant difference between consignments.

H_0 : There is no significant difference between observers.

H_1 : There is significant difference between observers.

$F(3, 15)$ at 0.05 level of significance = 3.29

As the calculated value is greater than the critical value, H_0 is rejected. Hence there is significant difference between observers.

Example 15.10 A manufacturing company has purchased three new machines for producing a commodity. Five hourly production from these three machines are observed at random. The following sum of squares have been worked out from the data:

Sum of squares of variation between samples = 250

Sum of squares of variation within samples = 200

Draw an analysis of variance table and test whether one of the machines is faster in production. (Given that at 5% level of significance, $F_{2, 12} = 3.89$).

Solution

ANOVA Table

Source of variation	*Sum of Squares*	*Df*	*Mean Squares*	*MS Variance Ratio F*
Between samples	250	3–1	125	
Within samples	200	15–3	16.67	125/16.67 = 7.49
Total	**450**	**14**	**450/14 = 32.14**	

As the calculated $F = 7.49$ is > the critical value of $F_{2, 12} = 3.89$, it falls in the rejection region. This means that none of the machines is faster in production.

GLOSSARY

Analysis of variance (ANOVA)	A statistical technique used to test the equality of three or more sample means.
Between-column variance	An estimate of the population variance derived from the variance among the sample means.
F-distribution	A continuous distribution that has two parameters (*df* for the numerator and *df* for the denominator). It is mainly used to test hypotheses concerning variances.
F-ratio	In ANOVA, it is the ratio of between-column variance to within-column variance.
Grand mean	In ANOVA, it is the mean of all observations across all treatment groups.
Mean Square between Samples (MSB)	A measure of the variation among means of samples taken from different populations.
Mean Square within Samples (MSW)	A measure of the variation within data of all samples taken from different populations.
One-way ANOVA	The analysis of variance technique that analyses one variable only.
SSB	The sum of squares between samples. Also called the sum of squares of the factor or treatment.
SSE	The sum of squares of error (or residual).
SST	The total sum of squares given by the sum of SSB and SSW.
SSW	The sum of squares within the samples.
Two-way ANOVA	The analysis of variance technique that involves two-factor experiments.
Within-column variance	An estimate of the population variance based on the variances within the k samples, using a weighted average of the k sample variances.

LIST OF FORMULAE

Let k = the number of different samples (or treatments).

n_i = the size of sample i.

T_i = the sum of the values in sample i.

n = the number of values in all samples.

$= n_1 + n_2 + n_3 + \ldots$

Σx = the sum of the values in all samples

$= T_1 + T_2 + T_3 + \ldots$

Σx^2 = the sum of the squares of values in all samples.

1. Between-samples sum of squares

$$SSB = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \ldots \right) - \frac{(\Sigma x)^2}{n}$$

2. Within-samples sum of squares

$$SSW = \Sigma x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \ldots \right)$$

3. Total sum of square

$$SST = SSB + SSW = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

4. $SSE = SST - (SSC + SSR)$
5. Variance between samples
   $MSB = SSB/k - 1$
6. Variance within samples
   $MSW = SSW/n - k$
7. Test statistic F for a one-way ANOVA test

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{MSB}{MSW}$$

8. F statistic for a two-way ANOVA test
   F_c = MSC/MSE and F_r = MSR/MSE
   (for these notations see pages 413 and 414]

QUESTIONS

15.1 Given below are ten statements. Indicate in each case whether the statement is true or false:

(a) The lower the value of F-statistic, the less we tend to believe there is a difference among the various samples.

(b) In analysis of variance, it is not necessary that samples should be of the same size.

(c) Any test of hypothesis involving the F statistic must be an upper-tail test.

(d) If the calculated value of F at a specified level of significance is less than the critical value of F, then the null hypothesis should be rejected.

(e) If all the data values are increased by 3 in a series (where analysis of variance has been done), then our answer will be different from the first one.

(f) The terminology used in analysis of variance is loaded with agricultural terms.

(g) In case of nominal or ordinal data the results are given in per centages or ranks, analysis of variance should not be carried out.

(h) There is no difference in the procedure followed for analysis of variance in a two-way classification as compared to that used in a one-way classification.

(i) F ratio can be calculated by dividing variance within the samples by variance between the samples.

(j) It is possible to complete a partially filled analysis of variance table if three figures are given: (a) sum of squares between samples (b) total sum of squares and (c) total degrees of freedom.

Mutiple Choice Questions (15.2 to 15.10)

15.2 In a one-way analysis of variance, the sum of squares is obtained by

(a) SST – SSB (b) SST + SSB (c) SSB – SST (d) none of these

15.3 The F ratio shows:

(a) Two estimates of the population mean
(b) One estimate of the population mean and one estimate of the population variance
(c) Two estimates of the population variance
(d) Both (a) and (c)
(e) None of the above

15.4 A step in performing ANOVA is to determine

(a) an estimate of the population variance from among the sample means.
(b) an estimate of the population variance from within the samples.
(c) the difference between expected and observed frequency for each class.
(d) (a) and (b)
(e) (b) and (c)

15.5 Given $n_1 = 20$, $n_2 = 15$ and $a = 0.10$, a two-tail test in ANOVA is to be performed. Which of the following represents the upper value to which s_1^2/s_2^2 should be compared?

(a) $\dfrac{1}{F(19, 14, 0.05)}$ (b) $\dfrac{1}{F(19, 14, 0.95)}$

(c) $F(9, 14, 0.05)$ (d) $F(14, 19, 0.05)$

15.6 Any difference among the population means in ANOVA will result into an increased expected value of

(a) MSB (b) MSW (c) MSE (d) all of these

15.7 The degrees of freedom between samples for k samples of size n will be

(a) $n - 1$ (b) $nk - 1$ (c) $k - 1$ (d) none of these

15.8 The error sum of squares can be obtained from the equation:

(a) SSE = SSR + SSC – SST (b) SSE = SST + SSR + SSC
(c) SSE = SST – SSR – SSC (d) none of these

15.9 If the data are displayed in rows and columns in a two-way classification, then the degree of freedom will be

(a) $c - 1$ (b) $r - 1$
(c) $(r - 1)(c - 1)$ (d) $c(r - 1)$
(e) none of these

15.10 Which is the assumption of analysis of variance?

(a) The samples drawn from the population are random.
(b) The data are quantitative and are normally distributed.
(c) The variances of the population from which samples have been drawn are equal.
(d) (a) and (b) but not (c)
(e) (a), (b) and (c)

15.11 What is 'ANOVA'? What is it used for?

15.12 Describe the procedure involved in the analysis of variance.

15.13 How is the analysis of variance technique useful to business?

15.14 What assumptions are made while using the analysis of variance technique?

15.15 What is an *F*-test? Discuss its application in testing whether the two variances are homogeneous.

15.16 For what kinds of problems is analysis of variance used?

15.17 State a management problem where ANOVA would be relevant to facilitate the decision-making process. State ANOVA model for two-way classified data with one observation per cell.

15.18 There are some missing values in the following ANOVA table. The data relate to three different types of packaging.

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*F*
Between samples	40			
Within samples		12		
Total	**184**	**14**		

(a) Find the missing values and complete the ANOVA table.

(b) Test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ at 5 per cent level of significance.

15.19 Consider $H_0 : m_1 = m_2 = m_3 = m_4 = m_5$

(a) Complete the following ANOVA table.

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Squares MS*	*F*
Between samples	132			
Within samples				
Total	**532**	**44**		

(b) Test the hypothesis at **(i)** 5 per cent level of significance **(ii)** 1 per cent level of significance and interpret the results.

15.20 Three different methods of teaching Statistics are used on three groups of students. Random samples of size 5 are taken from each group and the results are shown below. The grades are on a 10-point scale.

Group A	*Group B*	*Group C*
7	3	4
6	6	7
7	5	7
7	4	4
8	7	8

Determine on the basis of the above data whether there is a difference in the teaching methods.

15.21 The following table gives the data on the performance of three different detergents at three different water temperatures. The performance was obtained on the 'whiteness' readings based on specially designed equipment for nine loads of washing:

	Detergent A	*Detergent B*	*Detergent C*
Cold water	45	43	55
Warm water	37	40	56
Hot water	42	44	46

Perform a two-way analysis of variance, using the level of significance $\alpha = 0.05$.

15.22 The following data represent the number of units of production per day turned out by five different workmen using different types of machines:

	Machine Types			
Workman	*A*	*B*	*C*	*D*
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

(a) Test whether the mean productivity is the same for the four different machine types.

(b) Test whether five men differ with respect to mean productivity.

15.23 Three training methods were compared to see if they led to greater productivity after training. The productivity measures for individuals trained by different methods are as below:

Method 1	36	26	31	20	34	25
Method 2	40	29	38	32	39	34
Method 3	32	18	23	21	33	27

At the 0.05 level of significance, do the three training methods lead to different levels of productivity?

15.24 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			
Workman	*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.

15.25 Suppose that you are to test the hypothesis that the means of populations A, B and C are the same with a 5 per cent level of significance. Consider the following sample data:

A	*B*	*C*
0	5	17
2	8	20
8	13	25
2	9	21
3	9	23
3	10	20

Using these data, the following ANOVA tables is obtained:

Source of Variation	*Sum of Squares*	*Degrees of Freedom df*	*Mean Squares MS*	*F*
Between samples	1008	2	504	70
Within samples	108	15	7.2	
Total	**1116**	**17**		

Suppose that 2 is added to each data value. The following data are obtained:

A	*B*	*C*
2	7	19
4	10	22
10	15	27
4	11	23
5	11	25
5	12	22

(a) Construct an ANOVA table for these data and compare with the previous ANOVA table.
(b) Explain why such a relationship holds between the two ANOVA tables.

15.26 Three varieties of wheat, A, B and C, were treated with four different fertilizers, 1, 2, 3 and 4. The yields of wheat per acre were as follows:

	Varieties of Wheat			
Fertilizers	*A*	*B*	*C*	*Total*
1	55	72	47	174
2	64	66	53	183
3	58	57	74	189
4	59	57	58	174
Total	**236**	**252**	**232**	**720**

Perform an analysis of variance on the above data and interpret the results.

15.27 The price/earning ratios of three types of companies, viz. Banking, Financial Services and Insurance Companies are given below:

	Price/Earning Ratios		
	Banking	*Financial Services*	*Insurance*
1	15	19	16
2	14	25	12
3	20	14	16
4	16	12	13
5	23	17	19
6	17	15	21
7	22	25	33
8	32	27	19
9	31	33	33
10	25	27	30

Can we infer that the average price/earning ratio is same for all the three types of companies? Use 1per cent level of significance.

15.28 Four different drugs have been developed for a certain disease. These drugs are used in three different hospitals and the results given below show the number of cases of recovery from the disease per 100 people who have taken the drugs.

	A_1	A_2	A_3	A_4
B_1	19	8	23	8
B_2	10	9	12	6
B_3	11	13	13	10

What conclusions can you draw?

15.29 Apply the technique of Analysis of Variance to the following data, relating to yields of four varieties of wheat in three blocks:

	Blocks		
Varieties	*1*	*2*	*3*
I	10	9	8
II	7	7	6
III	8	5	4
IV	5	4	4

15.30 A certain company had four salesmen—A, B, C and D, each of whom was sent for a month to three types of areas—countryside K, outskirts of a city O and shopping centre of city S. The sales in hundreds of rupees per month are shown below:

Districts	*Salesmen*			
	A	*B*	*C*	*D*
K	30	70	30	30
O	80	50	40	70
S	100	60	80	80

Carry out an analysis of variance and interpret the results.

15.31 Set up an ANOVA table for the following per hectare yield for three varieties of wheat each grown on four plots. (Use 5 per cent level of significance.)

Plot of Land	*Variety of Wheat*		
	A_1	A_2	A_3
1	16	15	15
2	17	15	14
3	13	13	13
4	18	17	14

15.32 Experiments were carried out to test the effect of four chemical agents on the tensile strength of a particular type of cloth. Because there may be variability from one bolt to another, each chemical was tested on each of the five bolts randomly selected. The results of the experiments on the tensile strengths are shown below.

Use an appropriate statistical test at 5 per cent level to assess whether the tensile strengths differ between chemicals and between bolts.

Chemicals	*Bolts*				
	1	*2*	*3*	*4*	*5*
1	73	68	74	71	67
2	74	67	75	72	70
3	75	68	78	73	68
4	73	71	75	73	69

15.33 The following table gives the number of units of production, per day, turned out by four different types of machines.

Employees	*Types of Machines*			
	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

Perform two-way ANOVA, with $\alpha = 0.05$.

15.34 Three varieties of potatoes are being compared for yield. The experiment is carried out by assigning each variety, at random, to four of the twelve equal sized plots, one in each of the four locations. The following are the yield results in tonnes.

Location	*Variety of Potatoes*		
	A	*B*	*C*
1	18	13	12
2	20	23	21
3	14	12	9
4	11	17	10

Test whether there are any difference in the yields of the potatoes.

15.35 The following data are the outputs, per day, from three machines, when operated by four mechanics.

	Machines		
Mechanics	*A*	*B*	*C*
1	44	48	38
2	37	40	36
3	45	38	32
4	40	44	44

Perform the analysis of variance, with $\alpha = 0.05$.

15.36 A test was given to students, chosen at random, from the M. Com. classes of three universities in Rajasthan. Five students from each university were selected. Their scores were as follows:

University	*Scores*				
A	90	70	60	50	80
B	70	40	50	40	50
C	60	50	60	70	60

Perform analysis of variance and show if there is any significant difference between the scores of students in the three universities (Given F, 5% = 3.44).

15.37 Yields of four varieties of wheat in three blocks are given below:

	Blocks		
Varieties of Wheat	*1*	*2*	*3*
A	10	9	8
B	7	7	6
C	8	5	4
D	5	4	4

Is the difference between varieties significant @ 5% level of significance?

15.38 In a feeding experiment on pigs, three types of feed, R_1, R_2 and R_3 were tried. The animals were classified into three breeds, according to breed and body weight. The following table gives the gains in body weight in kg.

	Gain in Body Weight (kg)		
Types of Feed	*Breed 1*	*Breed 2*	*Breed 3*
R_1	4	16	10
R_2	14	18	19
R_3	3	14	7

Analyse the data and draw your inferences (5% level).

15.39 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.

Technician I	*Technician II*	*Technician III*	*Technician IV*
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

The McGraw-Hill Companies

CHAPTER 16 REGRESSION ANALYSIS

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand and use the least squares method to calculate the equations of a regression line for a given set of data
- use alternative methods to obtain a regression line
- use the regression coefficients sensibly to make forecasts
- understand the assumptions under which the regression analysis is carried out
- evaluate as to how good the regression is.

Chapter Prerequisites

Before starting work on this chapter, you should ensure that you are conversant with

1. the equation of a straight line
2. the basic ideas of graph plotting
3. the ANOVA table

16.1 INTRODUCTION

In this chapter, we study the association between variables, in two respects. *First*, we learn how to build statistical models of relationships to have a better understanding of their features. *Second*, we extend the models to consider their use in forecasting. At the end, we emphasise the need for caution in using the regression analysis and sources of potential error in its application.

In business, several times it becomes necessary to have some forecast so that it can take a decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets. Thus, to determine the nature and extent of relationship between these two variables becomes important for the company.

In Chapter 14, we discussed chi-square tests of independence to determine whether a statistical relationship between two variables existed. But these tests do not indicate the nature and extent of relationship. For this purpose, we have to use another technique—*regression analysis*—which forms the subject-matter of this chapter.

What is Regression?

Let us first understand the term regression. It was Sir Francis Galton who first used the term regression as a statistical concept in 1877. He made a statistical study that showed that the height of children born to tall parents tends to 'regress' towards the mean height of population. Galton used the term regression as a statistical technique to predict one variable (the height of children) from another variable (the height of parents). This is called 'regression' or 'simple regression' confined to bivariate data. Subsequently, statisticians coined another terms 'multiple regression' indicating the process of predicting one variable from two or more variables instead of only one. This chapter focuses on simple regression while Chapter 18 will deal with multiple regression.

The variable that forms the basis for predicting another variable is known as the *independent* or *predictor variable* and the variable that is predicted, is known as the *dependent variable*. In the study done by Galton, mentioned above, the height of the parents was the independent variable and the height of children was the dependent variable. Likewise, when we are trying to predict the demand for television sets on the basis of population, we are using the demand for television sets as the dependent variable and the population as the independent or predictor variable.

16.2 REGRESSION MODEL*

A statistical model is a set of mathematical formulas and assumptions which describe a real world situation. In this sense, simple linear regression as also multiple regression are statistical models. A statistical model tries to capture the systematic behaviour of the given data, leaving out those factors that cannot be foreseen or predicted. These factors are the errors. Despite our best efforts, it is highly unlikely that a model will reveal a perfect real world situation. However, this seldom happens on account of inherent uncertainity because multiple factors operate in a real world situation.

A good statistical model is one which provides as large a systematic component as possible, minimising errors. The errors are denoted by ε and constitute the random component in the model. This means a statistical model provides a break-up of data into two components—a non-random, systematic component, which can be described by a formula, and a purely random component that reflects errors. A major question arises here: How do we deal with the errors.

Errors in Random Component

These errors are on account of a number of factors that we are unable to identify. We assume that the random errors ε are normally distributed. In case we are able to construct a good model, then the average of observed errors will be zero. These errors should also be independent of one another. Although it is not absolutely necessary to use the assumption of normality, it is used to enable ourselves to perform statistical hypothesis tests using the F and t distributions. The only necessary assumptions are that errors ε have a mean zero with a constant variance σ and that they should be independent of each other.

* The discussion here is on the lines of Amir D. Aczel and J. Sounderpandian: *Complete Business Statistics,* Tata McGraw-Hill Edition, New Delhi, 2002, pp. 436–437.

As a first step, we choose a particular model, say a linear regression model, for describing the relationship between the two variables. As a second step, we work out the estimates of the model parameters on the basis of random sample data. The third step is to consider the errors that are called residuals, arising on the fit of the model to the data. When we are convinced that the residuals contain only pure randomness, we consider our model quite appropriate for its intended purpose, which invariably happens to make predictions.

A common example in the business world is the relationship between advertising and sales. When a linear regression model involving these two variables is appropriate for prediction, we may use it for predicting sales for a given level of advertising expenditure. It may be noted that the level of advertising should be within the range of expenditure on advertising covered in the study.

16.3 ESTIMATION USING THE REGRESSION LINE

When we have to study the relationship between two variables, X and Y, the very first step we should take is to plot the given data on the graph. For each pair of X and Y values, there will be a point on the graph. Such a graph or chart is known as a scatter diagram (Figure 16.1).

A scatter diagram can give us a broad idea of the type of relationship (or even absence of any relationship) between the two variables under study. In the next chapter on Correlation, we shall discuss scatter diagrams in some detail. Here, we shall show how to calculate the regression line by using an equation that relates the two variables. It may be noted that we are concerned here with the linear relationship between two variables only.

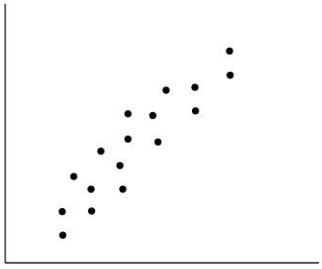


Fig. 16.1 **Scatter Diagram**

The equation for a straight line is

$$Y = a + bX$$

where Y is the dependent variable, X is the independent variable, a is the Y-intercept, which is the point at which the regression line crosses the Y-axis (the vertical axis) and b is the slope of the regression line. It should be noted that the values of both a and b will remain constant for any given straight line.

On the basis of this equation, we can find the value of Y for any value of X if values of a and b are known to us. Suppose that a is 5 and b is 3 and we have to find the value of Y if X is 10.

$$Y = a + bX$$
$$= 5 + (3 \times 10) = 35$$

The question now is: how to determine the values of two constants a and b? Let us illustrate this with the help of Fig. 16.2. From this figure, we can clearly make out that the Y-intercept a is 2 as the straight line touches Y-axis at point 2. As regards the slope of the straight line, that is, b, we can determine it only when we know how the dependent variable Y changes as the independent variable X changes. A straight line can be drawn when at least two points are known. As such, we can locate two points in

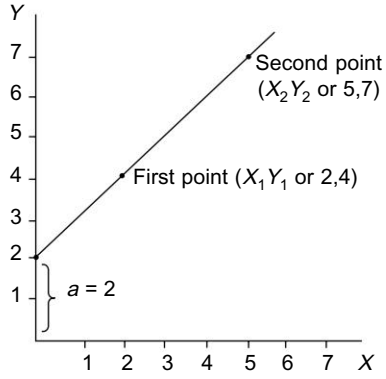


Fig. 16.2 **Calculating the Straight Line (Positive Slope)**

Fig. 16.2. In other words, we must find the value of X and Y (the coordinates) of both points. We take the coordinates of our first point X_1, Y_1 as 2 and 4. Likewise, we take the coordinates of our second point X_2, Y_2 as 5 and 7. By joining these two points, we get a straight line and extend it to meet Y-axis at point 2.

The slope of the straight line given in Fig. 16.2 can be numerically calculated by the following formula:

$$b = (Y_2 - Y_1) / (X_2 - X_1)$$

Applying the values of $X_1 = 2$, $X_2 = 5$, $Y_1 = 4$ and $Y_2 = 7$ as we have just used in the above formula, $b = (7 - 4)/(5 - 2) = 1$. In this particular case, the slope b is 1. We may now write the equation $Y = a + bX$ with numerical values as $Y = 2 + 1X$.

On the basis of this equation, we can now find the values of the dependent variable Y for varying values of X. Suppose, we want to find what will be the value of Y if X is 50. Since the value of b is 1 and the value of X will remain the same, that is, 50, the corresponding value of Y will be

$$Y = 2 + (1 \times 50) = 52$$

It is also to be noted that as X increases, Y too increases, which leads to the value of b as positive. This means that X and Y have a positive relationship.

We may now take another example to show the negative slope.

$$b = (Y_2 - Y_1)/(X_2 - X_1) = (4 - 7)/(6 - 1) = -0.6$$

It is clear from Fig. 16.3 that the slope b is negative. When the value of X is small (1), the corresponding value of Y is large (7). Again, when X is large, that is, 6, the corresponding value of Y tends to be small—4. By applying the formula as given earlier, we find that the slope of this straight line is –0.6. When b is negative, X and Y have an inverse relationship as shown in Fig. 16.3. When b is positive, X and Y have a positive relationship as shown in Fig. 16.2.

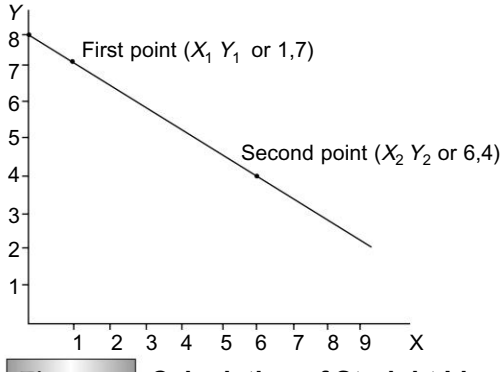


Fig. 16.3 **Calculation of Straight Line (Negative Slope)**

Suppose that, we have to find the value of Y when X is 20. As can be seen from Fig. 16.3, the Y-intercept is 8, that is, the point at which the straight line meets the Y-axis. This is the value of a. The value of b is –0.6, which we just calculated. Now, we substitute the values of a and b in the estimating equation $Y = a + bX$

$$Y = 8 + (-0.6)\,X = 8 - (0.6 \times 20) = -4$$

Similarly, we can calculate the different values of Y for the varying values of the independent variable X.

16.4 THE METHOD OF LEAST SQUARES

Having discussed the Y-intercept and the slope of a straight line, we now introduce the method of least squares. In a scatter diagram, if a straight line is drawn, it obviously does not touch several points. In fact, it may not pass through any of the scattered points. Some points may be above the straight line while the remaining points may be below it. Surely, there must be a technique to draw a line that is

'best fit'. Such a straight line will minimise the error between the estimated point that lies on the line and the actual point based on the data given. Figure 16.4 shows this. It will be seen that there are, in all, 10 points and all of them do not lie on the estimating line. Only three points are on it, four points are above it and the remaining three points are below it. The technique used for this purpose is known as the *method of least squares*.

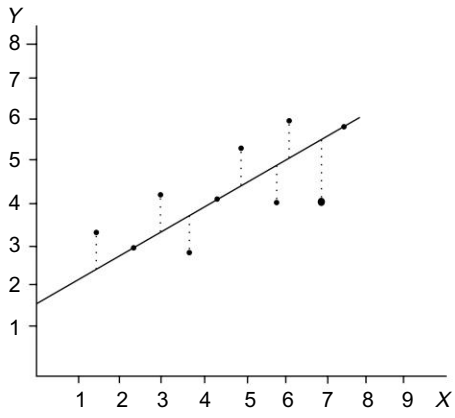


Fig. 16.4 **Scatter Diagram with an Estimating Line**

In order to explain the method of least squares, it is necessary to introduce a new symbol. We have used Y so far to represent the individual value of the observed points on the Y-axis. A new symbol $\hat{Y}$ (computed or estimated value of Y) is used to represent individual values of the estimated points, that is, those points that actually lie on the estimating line. In view of this, the equation for the estimating line becomes $\hat{Y} = a + bX$. We again refer to Fig. 16.4. The difference between the actual point and the estimated point is shown by a vertical line in respect of each of the seven points either above or below it.

We may raise a pertinent issue here: which would give a better fit—one or two points which are far away from the observed points in an estimating equation or a few points showing small differences from the original or observed points? We can visualise that the second alternative smaller differences between the estimated points and the observed points would give a 'better fit'. This leads us to square the individual differences or errors before summing them up.

This process of squaring serves two purposes. *First*, it magnifies the larger difference or error. *Second*, the effects of positive and negative differences get cancelled. This is because the square of negative difference or error is a positive figure. This process is known as the *least square method* as it minimises the sum of the squares of the error in the estimating line.

With the help of this method of least squares, we can find out whether one estimating line is better than the other. But, as a very large number of estimating lines can be drawn on the basis of the same set of data, it is extremely difficult for us to say that the estimating line that we have obtained is the best-fitting line. To overcome this problem, a set of two equations known as the *normal equations* is used to determine both the Y-intercept and the slope of the best-fitting regression line.

The two normal equations are:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

where

ΣY = the total of Y series

n = number of observations

ΣX = the total of X series

ΣXY = the sum of XY column

ΣX^2 = the total of squares of individual items in X series

a and b are the Y-intercept and the slope of the regression line, respectively.

We now take up an example to illustrate the use of the two normal equations.

Example 16.1 Given the following data, find the regression equation of Y on X.

X	2	3	4	5	6
Y	7	9	10	14	15

Solution We have now to set up a worksheet to get the values of the terms shown earlier.

Table 16.1 Worksheet for Computing Regression

X	Y	XY	X^2
2	7	14	4
3	9	27	9
4	10	40	16
5	14	70	25
6	15	90	36
$\Sigma X = 20$	$\Sigma Y = 55$	$\Sigma XY = 241$	$\Sigma X^2 = 90$

Substituting these values in the normal equations given above

$$55 = 5a + 20b \quad \text{(i)}$$
$$241 = 20a + 90b \quad \text{(ii)}$$

Multiplying (i) by 4, and from this equation (iii) subtracting (ii), we get

$$220 = 20a + 80b \quad \text{(iii)}$$
$$241 = 20a + 90b \quad \text{(ii)}$$
$$-21 = -10b$$

Therefore, $b = -21/-10 = 2.1$

Substituting the value of $b = 2.1$ in (i) above,

$$55 = 5a + (20 \times 2.1)$$

or $$55 = 5a + 42$$

or $$a = 13/5 = 2.6$$

Therefore, the regression equation of Y on X is

$$Y = 2.6 + 2.1X$$

16.5 ALTERNATIVE APPROACH

We can use an alternative approach, which involves the use of two formulae—one to calculate the Y-intercept and the other to calculate the slope.

The formula for calculating the slope is

$$b = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}$$

In order to apply the above formula, we should know the values of $\bar{X}$ and $\bar{Y}$ in addition to those of ΣXY and Σx^2.

$$\bar{X} = \Sigma X/n = 20/5 = 4$$

$$\bar{Y} = \Sigma Y/n = 55/5 = 11$$

$$b = \frac{241 - 5(4 \times 11)}{90 - 5(4)^2}$$

$$= \frac{241 - 220}{90 - 80} = 2.1$$

The formula for calculating the Y-intercept of the line is

$$a = \bar{Y} - b\,\bar{X}$$

where a is Y-intercept.

Applying this formula to calculate the Y-intercept, we get

$$a = 11 - (2.1 \times 4)$$
$$= 11 - 8.4 = 2.6$$

Hence, the regression equation is $Y = 2.6 + 2.1X$ (same as was obtained earlier by applying the normal equations).

On the basis of this regression, we can find the value of Y for any value of X. Suppose that we have to ascertain the value of Y where X is 14. Applying this value of $X = 14$ in the above equation,

$$Y = 2.6 + (2.1 \times 14) = 2.6 + 29.4 = 32$$

We are now clear as to how the regression line is obtained. The question is how to check the accuracy of our results.

Checking Accuracy of Regression Line Results One method is to draw a scatter diagram with original data pertaining to X and Y series and then to fit a straight line. This graph will give a visual idea about the suitability of the straight line fitted.

A more refined and, therefore, better approach is based on the mathematical properties of a line fitted by the method of least squares. This means that the positive and the negative errors (i.e. differences between the original data points and the calculated points) must be equal so that when all individual errors are added together, the result is zero. This is illustrated below.

Table 16.2 Worksheet for Obtaining Values of $\hat{Y}$

X	Y	$\hat{Y}$	$Y - \hat{Y}$
2	7	$2.6 + (2.1 \times 2) = 6.8$	0.2
3	9	$2.6 + (2.1 \times 3) = 8.9$	0.1
4	10	$2.6 + (2.1 \times 4) = 11.0$	–1.0
5	14	$2.6 + (2.1 \times 5) = 13.1$	0.9
6	15	$2.6 + (2.1 \times 6) = 15.2$	–0.2
		Total	**0**

Here, the calculated value of Y is shown as $\hat{Y}$. We find that the sum of positive errors $Y - \hat{Y}$ is equal to 1.2. The same is true for negative errors. Thus, the sum of the column $Y - \hat{Y}$ comes to zero. In view of this result, we can be reasonably sure that we have correctly done our regression problem.

16.6 USE OF DEVIATIONS FROM MEANS OF *X* AND *Y*

So far we have used absolute values of X and Y in calculating the Y-intercept and the slope of the regression line. Instead of using the absolute values, we may use an alternative approach wherein deviations from arithmetic means of X and Y are taken. If calculations are made manually, they can be simplified by this method. In such a case, the regression equation of Y on X will be

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

where $r\frac{\sigma_y}{\sigma_x}$

is the regression coefficient of Y on X. When deviations are taken from actual means, the regression coefficient of Y on X can be obtained as follows:

$$r = \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2}$$

Let us take our earlier example.

Table 16.3 Worksheet using Deviations from Arithmetic Means

X	Y	$X - \bar{X}$ x	x^2	$Y - \bar{Y}$ y	y^2	xy
2	7	–2	4	–4	16	8
3	9	–1	1	–2	4	2
4	10	0	0	–1	1	0
5	14	1	1	3	9	3
6	15	2	4	4	16	8
20	55		$\Sigma x^2 = 10$			$\Sigma xy = 21$

$$\bar{X} = 20/5 = 4$$
$$\bar{Y} = 55/5 = 11$$

Regression coefficient

$$r \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{21}{10} = 2.1$$

The regression equation of Y on X, as given earlier, is

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

or $Y - 11 = 2.1(X - 4)$

or $Y = 11 + 2.1X - 8.4$

or $Y = 2.6 + 2.1X$

Thus, we find that the regression line is the same as was obtained earlier when original values were used.

16.7 USE OF DEVIATIONS FROM THE ASSUMED MEANS

Sometimes we find that actual means of the two series X and Y are in fractions. In such cases, the calculation of regression coefficients becomes tedious. To overcome this problem, assumed means are used. Deviations are taken from the assumed means. The regression equation of Y on X needs to be modified as given below.

$$Y - \bar{Y} = \frac{\Sigma dxdy - \frac{\Sigma dx \times \Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}} X - \bar{X}$$

Let us take our earlier example to show the use of this approach. We take the assumed mean 3 for X series and 9 for Y series.

Table 16.4 Worksheet using Deviations from Assumed Means

X	Y	$X-3$ dx	dx^2	$Y-9$ dy	dy^2	$dx.dy$
2	7	–1	1	–2	4	2
3	9	0	0	0	0	0
4	10	1	1	1	1	1
5	14	2	4	5	25	10
6	15	3	9	6	36	18
		$\Sigma dx = 5$	$\Sigma dx^2 = 15$	$\Sigma dy = 10$	$\Sigma dy^2 = 66$	$\Sigma dx.dy = 31$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma dxdy - \frac{\Sigma dx \times \Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$$

$$= \frac{31 - \frac{5 \times 10}{5}}{15 - \frac{(5)^2}{5}}$$

$$= \frac{31 - 10}{15 - 5} = 2.1$$

Hence, the regression of Y on X is

$Y - \bar{Y} = 2.1\ (X - \bar{X})$

or $Y = 11 + 2.1X - 8.4$

or $Y = 2.6 + 2.1X$—same as obtained earlier.

16.8 REGRESSION IN CASE OF BIVARIATE GROUPED FREQUENCY DISTRIBUTIONS

Sometimes we come across bivariate frequency distributions where we have to fit in the regression function. In such cases, we have to first prepare a correlation table. After having prepared such a table, we have to apply the following formula to calculate regression coefficients.

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma fdxdy - \frac{\Sigma fdx \times \Sigma fdy}{N}}{\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}} \times \frac{i_x}{i_y}$$

where f is the frequency

i_x = *class-interval of X series*

i_y = *class-interval of Y series*

It will be seen that each of the terms in the earlier formula (except N) has been multiplied by the frequencies as the grouped data contain frequencies also. In addition, adjustment has been made in respect of the class-intervals of X and Y series. We take an example to explain the use of this formula.

Example 16.2 The following table gives the security prices and the amount of annual dividend. Calculate the regression of Y on X.

Annual Dividend (Rs) (X) / *Security Prices (Rs) (Y)*	*4–8*	*8–12*	*12–16*	*16–20*	*Total*
120–140	—	—	2	4	6
100–120	—	1	2	3	6
80–100	—	2	3	—	5
60–80	2	2	2	—	6
40–60	4	2	1	—	7
Total	**6**	**7**	**10**	**7**	**30**

Solution Table 16.5 gives different values required for calculating coefficient of correlation.

The calculations on the basis of the values obtained in Table 16.5 are as follows:

Substituting the values in the formula given earlier:

$$r\frac{\sigma_y}{\sigma_x} = \frac{68 - \frac{18 \times -4}{30}}{44 - \frac{(18)^2}{30}} = \frac{68 - \frac{72}{30}}{44 - \frac{324}{30}} = \frac{\frac{1968}{30}}{\frac{996}{30}}$$

$$= \frac{1968}{996} = 1.98$$

Regression equation

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

Table 16.5 Calculation of Regression of Y on X

Y \ X		4–8	8–12	12–16	16–20		$\frac{Y-90}{10}$			
	MV	6	10	14	18	*f*	*dy*	*fdy*	fd^2y	*fdxdy*
120–140	130	—	—	2 (8)	4 (32)	6	4	24	96	40
100–120	110	—	1 (0)	2 (4)	3 (12)	6	2	12	24	16
80–100	90	—	2 (0)	3 (0)	—	5	0	0	0	0
60–80	70	2 (4)	2 (0)	2 (–4)	—	6	–2	–12	24	0
40–60	50	4 (16)	2 (0)	1 (–4)	—	7	–4	–28	112	12
	f	6	7	10	7	30	0	–4	256	68
dx = *X*–10/4		–1	0	1	2					
fdx		–6	0	10	14	18				
fd^2x		6	0	10	28	44				
fdxdy		20	0	4	44	68				

Check

We have to calculate $\bar{X}$ and $\bar{Y}$.

$$\bar{X} = A + \frac{\Sigma fdx}{N} \times i_x$$

$$= 10 + \frac{18}{30} \times 4$$

$$= 10 + \frac{72}{30} = 10 + 2.4 = 12.4$$

$$\bar{Y} = A + \frac{\Sigma fdy}{N} \times i_y$$

$$= 90 + \frac{-4}{30} \times 10$$

$$= 90 - \frac{40}{30} = 88.67$$

Regression equation

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

or $Y - 88.67 = 1.98\,(X - 12.4)$

or $Y = 88.67 + 1.98X - 24.55$

or $Y = 64.12 + 1.98X$

16.9 REGRESSION COEFFICIENT

So far our discussion on regression analysis related to finding the regression of Y on X. It is just possible that we may think of X as dependent variable and Y as an independent one. In that case, we may have to use $X = a + bY$ as an estimating equation. Then, the normal equations will be

$$\Sigma X = na + b\Sigma Y$$
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Here, we will have to get the values of ΣX, ΣY, ΣXY, ΣY^2 and n. Once these values are known, we may enter them in the two normal equations. The equations then can be solved in the same manner as in the case of regression of Y on X.

(i) Regression equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

The term $r\frac{\sigma_x}{\sigma_y}$

can be written as $\Sigma xy/\Sigma y^2$.

(ii) Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

The term $r\frac{\sigma_y}{\sigma_x}$

can be written as $\Sigma xy/\Sigma x^2$.

It may be noted that the square root of the product of two regression coefficients is the value of the coefficient of correlation. We may write

$$bxy \text{ or } r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2}$$

$$byx \text{ or } r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2}$$

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

Another point to note is that x and y are the deviations in X and Y series from their arithmetic means. Let us take an example to illustrate the use of this method in the calculation of the coefficient of correlation.

Table 16.6 Calculation of Correlation Coefficient

X	Y	$X - \bar{X}$ x	$Y - \bar{Y}$ y	x^2	y^2	xy
2	7	–2	–4	4	16	8
3	9	–1	–2	1	4	2
4	10	0	–1	0	1	0
5	14	1	3	1	9	3
6	15	2	4	4	16	8
20	55	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 10$	$\Sigma y^2 = 46$	$\Sigma xy = 21$

$$\bar{X} = 20/5 = 4$$
$$\bar{Y} = 55/5 = 11$$

Regression equation of X on Y:

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{21}{46} = 0.46$$

Hence, $X - 4 = 0.46\,(Y - 11)$

or $X = 4 + 0.46Y - 5.06$

or $X = -1.06 + 0.46Y$

Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{21}{10} = 2.1$$

Hence, $Y - 11 = 2.1\,(X - 4)$

or $Y = 11 + 2.1X - 8.4$

or $Y = 2.6 + 2.1X$

Since $r = \sqrt{b_{xy} \cdot b_{yx}}$

$$= \sqrt{\frac{\Sigma xy}{\Sigma y^2} \times \frac{\Sigma xy}{\Sigma x^2}}$$

$$= \sqrt{\frac{21}{46} \times \frac{21}{10}}$$

$$= \sqrt{0.9587} = 0.98$$

Applying the values of r and σ_x and σ_y in the formula

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2}\, s$$

$$= (0.98)\frac{\sqrt{\frac{10}{5}}}{\sqrt{\frac{46}{5}}} = 0.98\frac{\sqrt{2}}{\sqrt{9.2}}$$

$$= \frac{0.98 \times 1.41}{3.03} = 0.46 \text{ (same as obtained earlier)}$$

Let us take a few examples where we can apply the above formulae.

Example 16.3 Given the following information:

	X	Y
Mean (Rs)	6	8
Standard deviation (Rs)	5	40/3

$r = 8/15$. Find (i) the regression coefficient of X on Y (ii) the regression coefficient of Y on X (iii) the most likely value of Y when $X =$ Rs 100.

Solution

(i) Regression equation of X on Y is as follows:

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

$$X - 6 = \frac{8}{15} \times \frac{5}{40/3} (Y - 8)$$

$$= 6 + \frac{1}{5} (Y - 8) = 6 + \frac{1}{5} Y - \frac{8}{5} = 4.4 + 0.2Y$$

(ii) Regression equation of Y on X is given by

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

or $$Y - 8 = \frac{8}{15} \times \frac{40}{3} \times \frac{1}{5} (X - 6)$$

or $Y = 8 + 1.422 (X - 6)$

or $Y = 8 + 1.422 X - 8.532$

or $Y = -0.532 + 1.422 X$

(iii) When $X = 100$, then Y will be

$Y = -0.532 + (1.422 \times 100) =$ Rs 141.67

Example 16.4 Compute the two regression equations on the basis of the following information:

	X	Y
Mean	40	45
Standard deviation	10	9

Karl Pearson's correlation coefficient = 0.5

Also estimate the value of Y for $X = 48$, using the appropriate regression equation.

Solution Regression equation of Y on X is

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

or $Y = 45 + 0.5 \times (9/10) (X - 40)$

$Y = 45 + 0.45 (X - 40)$

$Y = 45 + 0.45X - 18$

$Y = 27 + 0.45X$

Regression equation of X on Y is

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

or $X = 40 + 0.5 (10/9) (Y - 45)$

or $X = 40 + 0.556 (Y - 45)$

or $X = 40 + 0.556Y - 25.02$

or $X = 14.98 + 0.556Y$

In order to estimate the value of Y for $X = 48$, we have to use the regression equation of Y on X

$Y = 27 + 0.45X$

when $X = 48$

$Y = 27 + (0.45 \times 48)$

or $Y = 27 + 21.6$

or $Y = 48.6$

Properties of Regression Coefficients

At this stage, we should know the properties of the regression coefficients, which are:

1. The coefficient of correlation is the geometric mean of the two regression coefficients: Symbolically, $r = \sqrt{bxy \cdot byx}$.
2. As the coefficient of correlation cannot exceed 1, in case one of the regression coefficients is greater than 1, then the other must be less than 1.
3. Both the regression coefficients will have the same sign, either positive or negative. If one regression coefficient is positive, then the other will also be positive.
4. The coefficient of correlation and the regression coefficients will have the same sign. If the regression coefficients are positive, then the correlation coefficient will also be positive and vice versa.
5. The average of the two regression coefficients will always be greater than the correlation coefficient.

Symbolically,

$$(bxy) + (byx) / 2 > r \text{ since } (bxy) + (byx) / 2 > \sqrt{bxy} \times \sqrt{byx}$$

This should be obvious as r happens to be the square root of the two regression coefficients.

6. Finally, regression coefficients are not affected by change of origin. But this is not the case in respect of scale.

16.10 THE STANDARD ERROR OF ESTIMATE

We now turn to the concept of the standard error of estimate. It is the measure of the spread of observed values from the estimated ones, expressed by regression equation. This concept is similar to the standard deviation, which measures the variation of individual items about the arithmetic mean. As in the case of standard deviation, we can find the standard error of an estimate by calculating the mean of squares of deviations between the actual or observed values and the estimated values based on the regression equation.

It can be written symbolically

$$Syx = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}}$$

where $\hat{Y}$ *is the calculated or estimated value of Y*

Y is the actual or observed value of variable

n is the total number of observations

Syx is the standard deviation of regression of Y values on X values

It may be noted that the sum of the squared deviations is divided by n–2 and not by n. This is because we have lost 2 degrees of freedom in estimating the regression line. One can reason out that while obtaining values of a and b from the sample data, 2 degrees of freedom have been lost while estimating the regression line from these points.

There is a short-cut method for finding the standard error of estimate. The formula is

$$Syx = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2}}$$

where X = values of the independent variable
Y = values of dependent variable
a = Y-intercept
b = slope of the estimating equation
n = number of observations

It should be obvious that this formula gives a short-cut method. When we estimate the regression equation, all the values, which we need, are already determined. We illustrate both the methods with the help of an example.

Example 16.5 Suppose that we have been given the following data pertaining to two series X and Y:

X	40	30	20	50	60	40	20	60
Y	100	80	60	120	150	90	70	130

X series indicates advertising expenditure in thousand rupees and Y series relates to sales in units. We are told the regression equation is $\hat{Y} = 24.444 + 1.889X$. We are asked to calculate the standard error of estimate.

Solution We now set up a worksheet as follows:

Table 16.7 Worksheet

X	Y	$\hat{Y}$	$Y_i - \hat{Y}$	$(Y_i - \hat{Y})^2$
40	100	100	0	0
30	80	81	–1	1
20	60	62	–2	4
50	120	119	1	1
60	150	138	12	144
40	90	100	–10	100
20	70	62	8	64
60	130	138	–8	64
			Total	**378**

$$S_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n-2}}$$

$$= \sqrt{\frac{378}{8-2}} = \sqrt{\frac{378}{6}} = \sqrt{63} = 7.9 \text{ (standard error)}$$

In order to use the short-cut method, we rewrite the series X and Y.

Table 16.8 Worksheet

X	Y	Y^2	XY
40	100	10000	4000
30	80	6400	2400
20	60	3600	1200
50	120	14400	6000
60	150	22500	9000
40	90	8100	3600
20	70	4900	1400
60	130	16900	7800
	800	**86,800**	**35,400**

$$Syx = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2}}$$

$$Syx = \sqrt{\frac{86{,}800 - (24.444 \times 800) - (1.889 \times 35{,}400)}{8-2}}$$

$$= \sqrt{\frac{86{,}800 - 86{,}425}{6}} = 7.9 \text{ (same as obtained earlier)}$$

Having learnt the calculation of the standard error of estimate by the two methods, it is now necessary to know the interpretation of such an estimate.

Interpreting Standard Error of Estimate Higher the magnitude of the standard error of estimate, the greater is the dispersion or variability of points around the regression line. In contrast, if the standard error of estimate is zero, then we may take it that the estimate in equation is the best estimator of the dependent variable. In such a case, all the points would lie on the regression line. As such, there would be no point scattered around the regression line.

As in the case of standard deviation, we can expect that 68 per cent of the points lie within ±1 standard error, 95.5 per cent of the points within ±2 standard error, and 99.7 per cent of the points within ±3 standard error. This implies that the standard error can be used as a tool of statistical analysis in the same manner as is used in case of the standard deviation.

The standard error of estimate can be used as a statistical tool for determining prediction interval around an estimated value of $\hat{Y}$ within which the actual value of Y lies. As we have mentioned in the preceding paragraph that 95.5 per cent of the points lie within ±2 standard errors, since these intervals are around the estimated $\hat{Y}$ we may call these intervals as approximate prediction intervals. These intervals serve the same function as the confidence intervals as was discussed earlier. Let us take our previous example of advertising expenditure (X) and sale (Y) of a certain product. The regression equation of X and Y series given to us was

$$\hat{Y} = 24.444 + 1.889X$$

Suppose we want to know how much our sales would increase if we spend Rs. 70,000 on advertising. Multiplying the b coefficient by 70, we get

$$\hat{Y} = 24.444 + (1.889 \times 70)$$
$$= 24.444 + 132.23 = 157 \text{ approx. expected sales of the product}$$

It may be recalled that the standard error of the estimate based on the same X and Y series of data calculated earlier was 7.9. It is now possible to combine these two pieces of information. We can say that we are broadly 68 per cent confident that the actual sale of the given product will be within ±1 standard error of estimate from $\hat{Y}$. Both the upper and the lower limits of this prediction interval can be calculated as follows:

$$\hat{Y} + 1\text{Se} = \text{Rs } 157 + (1)\ (7.9)$$
$$= \text{Rs } 164.9 \text{ (Upper limit)}$$
$$\hat{Y} - 1\text{Se} = \text{Rs } 157 - (1)\ (7.9)$$
$$= \text{Rs } 149.1 \text{ (Lower limit)}$$

If we want to be 95.5 per cent confident that the actual sales of the product will be within ±2 standard errors of estimate from Y series, then our calculation for upper and lower limits would be as follows:

$$\hat{Y} + 2\text{Se} = \text{Rs } 157 + (2)\ (7.9)$$
$$= \text{Rs } 172.8 \text{ (Upper limit)}$$
$$\hat{Y} - 2\text{Se} = \text{Rs } 157 - (2)\ (7.9)$$
$$= \text{Rs } 141.2 \text{ (Lower limit)}$$

It should be borne in mind that these prediction intervals are based on the assumption that the distribution is normal. Since our sample size is very small, n being 8, our calculations may turn out to be inaccurate. All the same, the principle involved in calculating prediction levels remains the same.

In order to avoid this inaccuracy, we may have to use the t-distribution table given in Appendix Table 2. Suppose we want to be 90 per cent certain that the actual sales of the product would be within the production levels. We may have to use the values given in column .95. Since our n is equal to 8, we lose 2 degrees of freedom. As such we have to find the value of t against 6 degrees of freedom in the column .95. This value is 1.943. We can now calculate the upper and lower limits of the prediction interval as follows:

$$\hat{Y} + t\,(\text{Se}) = \text{Rs } 157 + (1.943)\ (7.9)$$
$$= \text{Rs } 172.35 \text{ (Upper limit)}$$
$$\hat{Y} - t\,(\text{Se}) = \text{Rs } 157 - (1.943)\ (7.9)$$
$$= \text{Rs}141.65 \text{ (Lower limit)}$$

We can be 90 per cent certain that the sales of the product will be between Rs 141.59 and Rs 172.41.

At the end, it may be reiterated that these prediction levels are only *approximate*.

16.11 HYPOTHESIS TESTS ABOUT REGRESSION RELATIONSHIP

Our discussion in this chapter has so far confined to the relationship of two variables on the basis of sample information. Since the sample data represent only part of the population, we may like to know whether our sample regression line can be regarded as a true (but unknown) population regression line $Y = A + BX$.

In order to determine whether a significant relationship exists between X and Y variables, a hypothesis test is performed. The null and alternative hypotheses are stated as follows:

H_0: B = 0 (There is no linear relationship.)

H_1: B ≠ 0 (There is a linear relationship.)

The formula for test statistic t is as follows:

$$t = \frac{b - B}{S_b}$$

where

$$S_b = \frac{Se}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}}$$

where ΣX^2, $(\Sigma X)^2$ and N are population values.

This process will be better understood if we take an example.

Example 16.6 Let us take Example 16.5 where the regression line is

$$Y = 24.444 + 1.889X$$

Solution Now, to perform the t test, we have to calculate S_b. The example gives Se = 7.9 but we have to calculate ΣX^2 and $\bar{X}$. Each X value as given in the problem is to be squared and then the total of squared values is to be obtained. This comes to 14,600 and $\bar{X}$ comes to 320/8 = 40. Applying these values in the formula given above, we get

$$S_b = \frac{7.9}{\sqrt{14600 - \frac{(320)^2}{8}}}$$

$$= \frac{7.9}{\sqrt{14600 - 12800}}$$

$$= \frac{7.9}{\sqrt{1800}}$$

$$= \frac{7.9}{42.43} = 0.186$$

Now we apply the t-test.

$$t = \frac{b - B}{0.186} = \frac{1.889 - 0}{0.186} = 10.156$$

Our null hypothesis is

H_0 : There is no linear relationship.

H_1 : There is a linear relationship.

The critical value of t for $n - 2 = 8 - 2 = 6$ df at 0.05 level of significance is 2.447. As the calculated value is more than the critical value, we reject the null hypothesis and conclude that B is not O, i.e. there is a linear relationship.

We can further take this exercise by using this test for a hypothesized value of B. Suppose we hope that population regression coefficient B is 2, not much different from b. Taking $B = 2$, we apply the t-test:

$$t = \frac{1.889 - 2}{0.186}$$

$$= -\frac{0.111}{0.186}$$

$$= -0.597$$

This calculated value of t is less than the critical value of t (+ 2.447) indicates that the hypothesized population regression coefficient is not significantly different from sample regression coefficient.

Apart from hypothesis testing, we can also construct a *confidence interval* for the value B. Let us first explain the need for having confidence interval.

Interval Estimate of B

We recall that $\hat{Y} = a + bx$ really is a sample regression line and, as such, is only one of several possible sample regression lines. The population regression line is $\hat{Y} = A + BX$ where A equals the population equivalent of the sample a. Similarly, B is the parameter analogous to b, which is the slope of the sample regression line.

Our point estimate of B is b, which in Example 16.5 is 1.889. With a different sample, obviously b will vary. In order to construct an interval estimate of B, it is necessary for us to know something about the sampling distribution of the regression coefficient. As the mean of the sampling distribution of b is B, as such b is an unbiased estimator of B. The standard error of this sampling distribution of b is S_b. We have worked out earlier the value of S_b as 0.186. Since $n = 8$ in our example, the t distribution is applicable.

In order to determine the interval estimate of B, the formula is

$$b \quad t\,S_b$$

The value of b is 1.889. The critical value of t for $n - 2$, i.e. 6 degrees of freedom at 0.05 level of significance is 2.447. Applying these values in $b \pm t\,S_b$, we get

$$1.889 \pm 2.447\ (0.186)$$

$$= 1.889 \pm 0.455$$

$$= 1.434 - 2.344$$

On this basis, we can say that we are 95 per cent confident that the true value of B lies between 1.434 and 2.344. That is, the range of change in Y for each unit of X is between 1.434 and 2.344. On an average, the change in Y with each unit of change in X is 1.889. It may be noted that if the sample is large, then instead of using t in the formula $b \pm t\,S_b$, we have to use Z.

16.12 HOW GOOD IS THE REGRESSION?

Once we have fitted a regression model to the given data, we would like to know how good the fit is. For this we use the ANOVA table. In Chapter 15, while discussing analysis of variance, a detailed discussion along with some examples was given pertaining to ANOVA. The procedure is the same

although the terms used in the context of regression model will be different. An example of ANOVA table for regression is given below.

Table 16.9 ANOVA Table for Regression

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Ratio*
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F(1,n-2) = \frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

There are three sources of variation in regression—the explained variation, error—the unexplained variation, and their sum, i.e. the total variation. We already know how to obtain sums of squares and the degrees of freedom, and based on these the mean squares. When we divide the mean square regression by the mean square error, we obtain another measure of the accuracy of our regression model. The resultant is an *F*-distribution with 1 and *n*–2 degrees of freedom. This *F* test is used to examine whether there is a linear relationship between the two variables *X* and *Y*.

$H_0 : \beta_1 = 0$ There is no linear relationship between *X* and *Y* (The two variables are independent.)

$H_1 : \beta_1 \neq 0$ There is linear relationship between *X* and *Y*.

If the calculated value of *F* exceeds the critical value of $(1, n-2)$ at 5 per cent level of significance, then H_0 is to be rejected and H_1 is to be accepted.

Strength of the Association It may be noted that the above test only confirms the linearity or otherwise of the association between the two variables. We do not know how strong this association is. In other words, how well *X* predicts *Y*. For this purpose, we have to calculate the coefficient of determination, i.e. r^2:

$r^2 = \frac{SSR}{SST}$, which shows variation in *Y* explained by regression compared to total variation. It should be obvious that greater is r^2, higher is the degree of association. The range of r^2 is 0 to 1 while *r* varies from –1 to +1.

Having explained the possible analysis of regression model, we now take a numerical example which will further clarify the concepts and the procedure involved in the tests.

Example 16.7 Let us use our earlier Example 16.1 wherein the data were as follows:

X	2	3	4	5	6
Y	7	9	10	14	15

Using these data, we arrived at the regression equation

$$Y = 2.6 + 2.1\,X$$

On the basics of this regression equation, the value of $\hat{Y}$ for corresponding values of *Y* were obtained. Table 16.2 is reproduced below:

X	Y	$\hat{Y}$	$Y-\hat{Y}$
2	7	6.8	0.2
3	9	8.9	0.1
4	10	11.0	–1.0
5	14	13.1	0.9
6	15	15.2	–0.2
			0

We have now to obtain sources of variation by calculating SST, SSR and SSE. This is shown below:

$$\text{SST} = (Y - \bar{Y})^2 \qquad \left[\bar{Y} = \frac{\Sigma Y}{n} = \frac{55}{5} = 11\right]$$
$$= (7-11)^2 + (9-11)^2 + (10-11)^2 + (14-11)^2 + (15-11)^2$$
$$= 16 + 4 + 1 + 9 + 16 = 46$$
$$\text{SSR} = (\hat{Y} - \bar{Y})^2$$
$$= (6.8-11)^2 + (8.9-11)^2 + (11.0-11)^2 + (13.1-11)^2 + (15.2-11)^2$$
$$= 17.64 + 4.41 + 0 + 4.41 + 17.64 = 44.1$$
$$\text{SSE} = (Y - \hat{Y})^2$$
$$= (0.2)^2 + (0.1)^2 + (-1.0)^2 + (0.9)^2 + (-0.2)^2$$
$$= 0.04 + 0.01 + 1 + 0.81 + 0.04$$
$$= 1.9$$

On the basis of above calculations, we set up the ANOVA table:

ANOVA Table

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Squares*	*F ratio*
Due to regression	1	44.1	44.1	44.1/0.24 = 183.75
Due to error	8	1.9	0.24	
Total	9	46		

The next step is to set up the two hypotheses:

H_0 : There is no linear relationship between X and Y, i.e. $\beta = 0$.

H_1 : Y is linearly related to X, i.e. $\beta \neq 0$.

The critical value of $F(1,8)$ at 5 per cent level of significance is 5.32. As the calculated value of F is far greater than the critical value of F, H_0 is rejected. The conclusion is that Y is a linear function of X with a confidence level of 95 per cent.

Strength of Association Coefficient of determination r^2

$$= \frac{\text{SSR}}{\text{SST}}$$
$$= 1 - \text{SSE/SST}$$
$$= 1 - \frac{1.9}{46} = 1 - 0.041 = 0.959$$

This implies that 95.9 per cent of the variation in Y is explained by the regression and only 4.1 per cent of the variation is explained by error. Hence, the positive association between the two variables is very strong to enable X_i to predict Y

$$r = \sqrt{r^2}$$
$$= \sqrt{0.959} = 0.979.$$

Additional Examples

Example 16.8 A company wants to assess the impact of R&D expenditure on annual profits. The following table gives the information for 8 years:

Year	*R&D Expenditure ('000 Rs)*	*Annual Profit ('000 Rs)*
1993	9	45
1994	7	42
1995	5	41
1996	10	60
1997	4	30
1998	5	34
1999	3	25
2000	2	20

Estimate the regression equation and predict the annual profit for the year 2002 for an allocated sum of Rs 15,000 as R&D expenditure.

Solution

	R&D Exp. ('000)		*A. profit ('000 Rs.)*	
Year	X	X^2	Y	XY
1993	9	81	45	405
94	7	49	42	294
95	5	25	41	205
96	10	100	60	600
97	4	16	30	120
98	5	25	34	170
99	3	9	25	75
2000	2	4	20	40
	45	309	297	1909

$$\Sigma X = 45\ \Sigma X^2 = 309\ \Sigma Y = 297\ \Sigma XY = 1909$$

Normal equations for regression are

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\ \Sigma X^2$$

Substituting the above values in the normal equations, we get

$$297 = 8a + 45b \quad (1)$$
$$1909 = 45a + 309b \quad (2)$$

Multiplying (1) by 45 and (2) by 8, we get

$$13365 = 360a + 2025b \quad (3)$$

$$15272 = 360a + 2472b \quad (4)$$

Subtracting (3) from (4), we get

$$1907 = 447b$$

$$\therefore \quad b = \frac{1907}{447} = 4.27$$

Substituting the value of $b = 4.27$ in (1) above, we get

$$297 = 8a + (45)(4.27)$$

or $$297 = 8a + 192.15$$

or $$8a = 297 - 192.15$$

$$\therefore \quad a = \frac{104.85}{8} = 13.11$$

Regression equation is $Y = 13.11 + 4.27X$

For an allocated sum of Rs 15000 on R&D, the annual profits would be

$$Y = 13.11 + (4.27)(15)$$
$$= 13.11 + 64.05$$
$$= 77.16, \text{ i.e., Rs } 77160$$

Example 16.9 Given the following values, find the expected value of X when $Y = 12$:

	X series	*Y series*
Average	25	22
S.D.	4	5
$r = 0.42$		

Solution In order to find the expected value of X when $Y = 12$, we have to use the regression equation of X on Y.

Regression equation of X on Y is:

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

Substituting the values given in the above equation:

$$X - 25 = 0.42 \frac{4}{5} (Y - 22)$$

or $$X = 25 + 0.336 (Y - 22)$$

or $$X = 25 + 0.336Y - 7.392$$

or $$X = 17.608 + 0.336Y$$

When Y is 12:

$$X = 17.608 + (0.336 \times 12)$$
$$= 17.608 + 4.032$$
$$= 21.64$$

Example 16.10 You are given below the following data about the sale and advertising expenditure of a firm:

	Sales (Rs crore)	*Adv. exp. (Rs crore)*
Arithmetic mean	50	10
Standard deviation	10	2
Coefficient of correlation + 0.9		

(a) Calculate the two regression equations.
(b) Calculate the likely sales for a proposed advertisement expenditure of Rs 13.5 crore.
(c) What should be the advertising budget if the company wants to achieve a sales target of Rs 70 crore?

Solution Advertising expenditure—X Sales—Y

(a) *The regression equation of Y on X*

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

or $Y - 50 = 0.9\frac{10}{2}(X - 10)$

or $Y = 50 + 4.5(X - 10)$

or $Y = 50 + 4.5X - 45$

or $Y = 5 + 4.5X$

The regression equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

or $X - 10 = 0.9\frac{2}{10}(Y - 50)$

or $X = 10 + 0.18\,(Y - 50)$

or $X = 10 + 0.18Y - 9$

or $X = 1 + 0.18Y$

(b) $Y = 5 + (4.5 \times 13.5)$
$= 5 + 60.75$
$=$ Rs 65.75 crore

(c) $X = 1 + (0.18 \times 70)$
$= 1 + 12.6$
$=$ Rs 13.6 crore

Example 16.11 The following table shows the marks obtained in two tests by 10 students:

Marks in Ist Test (X)	6	5	8	8	7	6	10	4	9	7
Marks in 2nd Test (Y)	8	7	7	10	5	8	10	6	8	6

(a) Find the least square regression line of Y on X.
(b) Test the hypothesis that marks in the two tests are linearly related.

Solution **(a)**

Worksheet for Computing Regression

X	Y	XY	X^2	Yi	$Y - Yi$	$(Y - Yi)^2$
6	8	48	36	7	1.0	1
5	7	35	25	6.5	0.5	0.25
8	7	56	64	8	–1.0	1
8	10	80	64	8	2.0	4
7	5	35	49	7.5	–2.5	6.25
6	8	48	36	7	1.0	1
10	10	100	100	9	1.0	1
4	6	24	16	6	0	0
9	8	72	81	8.5	–0.5	0.25
7	6	42	49	7.5	–1.5	2.25
70	75	540	520			17.00

Normal equations:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the above values in the normal equations,

$$75 = 10a + 70b \quad (1)$$

$$540 = 70a + 520b \quad (2)$$

Multiplying equation (1) by 7 and subtracting it from equation (2), we get

$$540 = 70a + 520b \quad (2)$$

$$525 = 70a + 490b \quad (3)$$

$$15 = 30b$$

$$\therefore \quad b = 15/30 = 0.5$$

Substituting the value of $b = 0.5$ in (1) above, we get

$$75 = 10a + (70 \times 0.5)$$

or $10a = 75 - 35$

$$\therefore \quad a = 40/10 = 4$$

Hence, the regression line of Y on X is

$$Y = 4 + 0.5X$$

(b) In order to test the hypothesis about linear relationship between the two variables, we have to first calculate *Se* for which the following formula is used:

$$Se = \sqrt{\frac{\Sigma(Y - Y_i)^2}{n - 2}}$$

$$= \sqrt{\frac{17}{10 - 2}} = \sqrt{\frac{17}{8}} = 1.458$$

Value of another term S_b is also to be obtained.

$$S_b = \frac{Se}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}}$$

$$= \frac{1.458}{\sqrt{520 - \frac{(70)^2}{10}}}$$

$$= \frac{1.458}{\sqrt{520 - 490}}$$

$$= \frac{1.458}{\sqrt{30}} = \frac{1.458}{5.477} = 0.266$$

The two hypotheses are:

H_0 : There is no linear relationship.

H_1 : There is a linear relationship.

Now, we have to apply the *t* test:

$$t = \frac{b - B}{S_b}$$

$$= \frac{0.5 - 0}{0.266}$$

$$= 1.88$$

The table (critical) value of *t* for $n - 2 = 8$ degrees of freedom at 10% level of significance is 1.860. As the calculated value of *t* is greater than the table value of *t*, the null hypothesis is rejected. The conclusion is that the marks in the two tests obtained by 10 students are linearly associated.

Example 16.12 On the basics of the following data, determine regression of *Y* on *X*:

$\Sigma X = 231 \quad \Sigma Y = 498 \quad \Sigma X^2 = 4903$

$\Sigma Y^2 = 21582 \quad \Sigma XY = 10211 \quad n = 12$

Also estimate the confidence interval of the regression coefficient.

Solution The normal equations for regression are:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$$498 = 12a + 231b \quad (1)$$

$$10211 = 231a + 4903b \quad (2)$$

Multiplying (1) by 77 and (2) by 4, we get

$$38346 = 924a + 17787b \quad (3)$$

$$40844 = 924a + 19612b \quad (4)$$

Subtracting (3) from (4), we get

$$2498 = 1825b$$

$$\therefore \quad b = \frac{2498}{1825} = 1.37$$

Substituting the value of $b = 1.37$ in (1) above, we get

$$498 = 12a + (231)(1.37)$$

or $$12a = 498 - 316.47$$

or $$a = \frac{181.53}{12} = 15.13$$

Hence, the regression equation is

$$Y = 15.13 + 1.37\,X$$

Now, we can determine the standard error, *Se* by using the following formula:

$$Se = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2}}$$

$$= \sqrt{\frac{21582 - (15.13)(498) - (1.37)(10211)}{12-2}}$$

$$= \sqrt{\frac{21582 - 7534.74 - 13989.07}{10}}$$

$$= \sqrt{\frac{58.19}{10}} = 2.41$$

The formula for calculating the standard error of the regression coefficient is

$$S_b = \frac{Se}{\sqrt{\Sigma X^2 - n\bar{X}^2}}$$

$$= \frac{2.41}{\sqrt{4903 - 12\left(\frac{231}{12}\right)^2}}$$

$$= \frac{2.41}{\sqrt{456.25}} = \frac{2.41}{21.36} = 0.11$$

For 95 per cent confidence and 10 degrees of freedom, the table value of t is 2.228. Applying this information, the confidence intervals of the regression coefficient are:

$$b + t\,(S_b) = 1.37 + (2.228)(0.11)$$
$$= 1.37 + 0.245$$
$$= 1.615 \text{ (upper limit)}$$
$$b - t\,(S_b) = 1.37 - (2.228)(0.11)$$
$$= 1.37 - 0.245$$
$$= 1.125 \text{ (lower limit)}$$

Hence, confidence interval of B is 1.125 – 1.615.

Example 16.13 For 10 observations of price (P) and supply (Y), the following data were obtained.

$\Sigma P = 130 \quad \Sigma Y = 220 \quad \Sigma P^2 = 2288$

$\Sigma Y^2 = 5506$ and $\Sigma PY = 3467$

Find the regression (supply function) of Y on P and estimate the supply when the price is 25 units.

Solution

It may be noted that instead of the usual notation X, notation P is used, as the problem pertains to price and supply. The normal equations with the changed notation would be

$$\Sigma Y = na + b\,\Sigma P$$
$$\Sigma PY = a\,\Sigma P + b\,\Sigma P^2$$

Substituting the given values in the above normal equations, we get

$$220 = 10a + 130b \quad (1)$$
$$3467 = 130a + 2288b \quad (2)$$

Multiplying equation (1) by 13, we get

$$2860 = 130a + 1690b \quad (3)$$

Subtracting (3) from (2), we get

$$607 = 598b$$
$$\therefore \quad b = 607/598$$
$$= 1.015$$

Substituting the value of $b = 1.015$ in (1), we get

$$220 = 10a + (130 \times 1.015)$$
or $$220 = 10a + 131.95$$
or $$10a = 220 - 131.95$$
or $$a = 88.05/10$$
$$= 8.805$$

Hence, the regression of Y on P is

$$Y = 8.805 + 1.015P$$

In order to estimate the supply when price is 25 units, we use the above regression equation:

$$Y = 8.805 + (1.015 \times 25)$$
$$= 8.805 + 25.375$$
$$= 34.18$$

Example 16.14 Given the following information

	Price (Rs)	*Amount demanded (units in '000)*
Average	10	35
S.D.	2	5

Correlation coefficient = 0.8

Obtain the regression equation of the amount demanded on the given price, and estimate the likely demand when the price is Rs 12.50.

Solution

Let the price be X and the amount demanded be Y.

Hence, we have to obtain the regression equation of Y on X, which is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

or $Y - 35 = 0.8\frac{5}{2}(X - 10)$

or $Y = 35 + 0.8 \times [2.5(X - 10)]$

or $Y = 35 + 0.8(2.5 \times -25)$

or $Y = 35 + 2X - 20$

or $Y = 15 + 2X$

When x = Rs 12.5, the likely demand will be

$Y = 15 + (2 \times 12.5)$

or $Y = 15 + 25 = 40$

Example 16.15 Estimate the regression of Y on X, given that

$\Sigma X = 15 \quad \Sigma Y = 15 \quad \Sigma X^2 = 49$

$\Sigma Y^2 = 49, \quad \Sigma XY = 44, \quad \Sigma n = 5$

Solution

The normal equations used for estimating regression are

$$\Sigma Y = na + b\Sigma x$$

$$\Sigma XY = a\Sigma x + b\Sigma x^2$$

Substituting the given values in the two normal equations, we get

$$15 = 5a + 15b \quad (1)$$

$$44 = 15a + 49b \quad (2)$$

Multiplying equation (1) by 3 and subtracting it from equation (2), we get

$$\begin{array}{r} 44 = 15a + 49b \\ 45 = 15a + 45b \\ \hline -1 = -4b \end{array} \quad (3)$$

$$\therefore \quad b = \frac{1}{4} = 0.25$$

Substituting the value of b in equation (1), we get

$15 = 5a + 3.75$

or $5a = 15 - 3.75$

or $5a = 11.25$

$$\therefore \quad a = \frac{11.25}{5} = 2.25$$

Hence, the regression of Y on X is

$Y = 2.258 + 0.25X$

Example 16.16 From the following data, obtain the regression of Y on X and X on Y.

X:	1	4	2	3	5
Y:	3	1	2	5	4

Solution

Worksheet

X	Y	X^2	Y^2	XY
1	3	1	9	3
4	1	16	1	4
2	2	4	4	4
3	5	9	25	15
5	4	25	16	20
15	**15**	**55**	**55**	**46**

$$\bar{X} = \frac{\Sigma X}{n} = \frac{15}{5} = 3$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{15}{5} = 3$$

$$\Sigma X^2 = 55$$

$$\Sigma Y^2 = 55$$

$$\Sigma XY = 46$$

For calculating regression of Y on X, the two normal equations used are

$$\Sigma y = na + b\,\Sigma X$$

$$\Sigma XY = a\,\Sigma x + b\,\Sigma X^2$$

Applying the above values in the two equations, we get

$$15 = 5a + 15b \quad (1)$$

$$46 = 15a + 55b \quad (2)$$

Multiplying (1) by 3, and subtracting (2) from it, we get

$$45 = 15a + 45b \quad (3)$$

$$\underline{-46 = -15a - 55b}$$

$$-1 = -10b$$

or $10b = 1$

Hence, $b = \frac{1}{10} = 0.1$

Substituting the value of $b = 0.1$ in (1), we get

$$15 = 5a + 1.5$$

or $5a = 15 - 1.5$

or $5a = 13.5$

Hence, $a = \frac{13.5}{5} = 2.7$

Hence, the regression of Y on X is

$$Y = 2.7 + 0.1X$$

For calculating regression of X on Y,

We write the two normal equations,

$$\Sigma x = na + b\Sigma y \quad (1)$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2 \quad (2)$$

the values from the the worksheet, we get

$$15 = 5a + 15b \quad (3)$$

$$46 = 15a + 55b \quad (4)$$

Multiply equation (3) by 3, and subtracting (4) from it, we get

$$45 = 15a + 45b \quad (5)$$

$$\underline{\_46 = \_15a + \_55b} \quad (4)$$

$$-1 = -10b$$

or $$b = \frac{1}{10} = 0.1$$

Now, substituting the value of $b = 0.1$ in equation (3)

$$15 = 5a + (15 \times 0.1)$$

or $$5a = 15 - 1.5$$

$\therefore$ $$a = 13.5/5 = 2.7$$

Hence, the regression of X on Y is

$$X = 2.7 + 0.1Y$$

Example 16.17 A child specialist observed 10 school students for their average calorie intake (x) and body weight (y) (in kg). The data analyst offered the following summation quantities based on the basic data on the two variables: $\Sigma x = 166$, $\Sigma y = 577$, $\Sigma xy = 9840$, $\Sigma x^2 = 2892$ and $\Sigma y^2 = 33927$. Using these quantities, find (a) absolute increase in weight per unit of calorie intake, (b) the minimum weight y-intercept, (c) the most likely weight against a calorie intake of 25, and (d) the standard error of estimate.

Solution

The two normal equations are

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Substituting the values given in the question in the above equations, we get

$$577 = 10a + 166b \quad (1)$$

$$9840 = 166a + 2892b \quad (2)$$

Multiplying (1) by 166 and (2) by 10, we get

$$95782 = 1660a + 27556b \quad (3)$$

$$\underline{\_98400 = \_1660a + \_28920b} \quad (4)$$

$$-2618 = -1364b$$

$\therefore$ $$b = -2618/-1364$$

$$= 1.919 \quad \text{or} \quad 1.92 \text{ approx.}$$

Substituting the value of $b = 1.92$ in (1), we get

$$577 = 10a + 318.72$$

or $$10a = 577 - 318.72$$

or $$a = 258.28/10$$

$$= 25.828 \quad \text{or} \quad 25.83 \text{ approx.}$$

Substituting the values of a and b in the regression equation, we get

$$Y = a + bx$$

or $$Y = 25.83 + 1.92x$$

(a) Average absolute increase in weight per unit of calorie intake is indicated by coefficient b, which is in kg.

(b) The minimum weight of intercept

$$Y = 25.83 + (1.92 \times 0) = 25.83 \text{ kg}$$

(c) If x is 25, then

$$Y = 25.83 + (1.92 \times 25)$$
$$= 25.83 + 48 = 73.83 \text{ kg}$$

(d) Standard error of estimate

$$\sqrt{\frac{\Sigma y^2 - a\Sigma y + b\Sigma xy}{n-2}}$$

$$= \sqrt{\frac{33927 - (25.83 \times 577) + (1.92 \times 9840)}{10-2}}$$

$$= \sqrt{\frac{33927 - 14903.91 + 18892.8}{8}}$$

$$= \sqrt{\frac{37915.89}{8}}$$

$$= \sqrt{4739.48625}$$

$$= 68.84 \text{ approx.}$$

Example 16.18 The Personnel Manager of a large industrial unit is interested in finding a measure that can be used to fix the yearly wages of skilled workers. On an experimental basis, he compiled data on the length of service and the respective yearly wages (in Rs '000) of a group of ten randomly selected workers.

Length of service (in years)	11	7	9	5	8	6	10	12	3	4
Yearly wages (in Rs '000)	14	11	10	9	13	10	14	16	6	7

Obtain the regression equation of wages on length of the service.

Solution

Worksheet

Length of service (X)	X^2	*Yearly wages ('000 Rs) Y*	*XY*
11	121	14	1114
7	49	11	77

(*Contd.*)

(*Contd.*)

9	81	10	90
5	25	9	45
8	64	13	104
6	36	10	60
10	100	14	140
12	144	16	192
3	9	6	18
4	16	7	28
$\Sigma X = 75$	**$\Sigma X^2 = 645$**	**$\Sigma Y = 82$**	**$\Sigma XY = 1868$**

The two normal equations are

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the given values in the normal equations, we get

$$82 = 10a + 75b \quad (1)$$

$$1868 = 75a + 645b \quad (2)$$

Multiplying (1) by 75 and (2) by 10 and subtracting equation (4) from equation (3), we get

$$6150 = 750a + 5625b \quad (3)$$

$$\underline{18680 = 750a + 6450b} \quad (4)$$

$$-12530 = -825b$$

$$\therefore \quad b = 12530/825$$

$$= 15.19$$

Substituting the value of $b = 15.19$ in (1), we get

$$82 = 10a + (15.19 \times 75)$$

$$82 = 10a + 1139.25$$

or $10a = 82 - 1139.25$

or $a = -1057.25/10$

or $a = -105.73$

Hence, the regression equation of wages on the length of service is $Y = -105.73 + 15.19X$.

Example 16.19 Given the following regression of Y on X and X on Y, find the correlation coefficient between X and Y and their means, $3X + 4Y = 7$ and $4X + Y = 5$.

Solution

Since $r = \sqrt{b_{xy} \times b_{yx}}$, we have to first find the values of b_{xy} and b_{yx}.

Now, $4X + Y = 5$

or $4X = 5 - Y$

or $X = \frac{5}{4} - \frac{1}{4} Y$ (The second term is the regression coefficient.)

or $b_{xy} = \frac{1}{4}$

$$3X + 4Y = 7$$

or $$4Y = 7 - 3X$$

or $$Y = \frac{7}{4} - \frac{3}{4}X$$ (The second term is the regression coefficient.)

or $$b_{yx} = \frac{3}{4}$$

Since $$r = \sqrt{b_{xy} \times b_{yx}}$$

$\therefore$ $$r = \sqrt{\frac{1}{4} \times \frac{3}{4}} = \sqrt{\frac{3}{16}} = 0.43$$

Now, to find the values of $\bar{X}$ and $\bar{Y}$.

Since the regression lines always intersect at a point $(\bar{X}, \bar{Y})$, representing the mean values of X and Y, the two equations given in the question are to be solved as shown below:

$$3X + 4Y = 7 \quad (1)$$

$$4X + Y = 5 \quad (2)$$

Multiplying (2) by (4) and subtracting it from (1), we get

$$\begin{array}{r} 3X + 4Y = 7 \\ -16X + -4Y = -20 \\ \hline -13X = -13 \end{array}$$

$\therefore$ $X = 1$ or $\bar{X} = 1$

Substituting the value of $X = 1$ in equation (1) above, we get

$$(3 \times 1) + 4Y = 7$$

or $$4Y = 7 - 3$$

or $$Y = 4/4 = 1 \quad \text{or} \quad \bar{Y} = 1$$

Example 16.20 The results of a survey on the sale of product (Y) as a function of time period (X) are summarized below:

	X	Y
Arithmetic Mean	40	125
Standard deviation	25	16
Correlation coefficient (r)	0.85	

(a) Fit the regression line of Y on X and estimate the value of Y when X is 45.
(b) Fit the regression line of X on Y and estimate the value of X when Y is 135.

Solution

(a) Regression line of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

or $$Y - 125 = 0.85\frac{16}{25}(X - 40)$$

or $Y = 125 + 0.544(X - 40)$

or $Y = 125 + 0.544X - 21.76$

or $Y = 103.24 + 0.544X$

When X is 45, then Y will be

$Y = 103.24 + (0.544 \times 45)$

or $Y = 103.24 + 24.48$

or $Y = 127.72$ or 128 approx.

(b) Regression line of X on Y is

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

or $X - 40 = 0.85\ \frac{25}{16}\ (Y - 125)$

or $X = 40 + 1.328\ (Y - 125)$

or $X = 40 + 1.328Y - 166$

or $X = -126 + 1.328Y$

When Y is 135, X will be $-126 + (1.328 \times 135) = 53.28$ or 53 approx.

Example 16.21 Given $X = 4Y + 7$ and $Y = 9X + 5$, as regression of X on Y and Y on X, respectively, examine whether there is any inconsistency in these regressions.

Solution

Since $X = 4Y + 7$ (1)

$Y = 9X + 5$ (2)

Multiplying (1) by 9, we get

$9X = 36Y + 63$ (3)

$9X = Y - 5$ (from the 2nd equation) (4)

Subtracting (4) from (3), we get

$35Y = -68$

$\therefore$ $Y = -68/35 = -1.94$

Substituting this value of Y in (1), we get

$X = [4 \times (-1.94)] + 7$

or $X = -7.76 + 7$

or $X = -0.76$

Applying the value of $X = -0.76$ in equation (2), we get

$Y = (9 \times -0.76) + 5$

$= -6.84 + 5$

$= -1.84$

As we can see, the value of Y differs from equation (1) (−1.94) to equation (2) (–1.84). Hence, it becomes obvious that there is inconsistency in the regressions.

Cautions in the Use of Regression Analysis

While regression analysis is an extremely useful technique for making predictions and as such it is frequently used, at the same time one should be careful in avoiding errors that may arise on account of wrong application of regression analysis. It is, therefore, necessary to know how errors can arise while using regression analysis.

1. The inclusion of one or two extreme items can completely change a given relationship between the variable. As such, extreme values should be excluded from the data.
2. It is advisable to first draw a scatter diagram so that one can have an idea of the possible relationship between X and Y. In the absence of a scatter diagram, one may attempt a linear regression model but the given set of data may actually show a non-linear relationship.
3. When predictions based on regression analysis are made, one should be sure that the nature and extent of relationship between X and Y will remain the same. This assumption at times is completely overlooked that may lead to errors in prediction.
4. In many cases the regression line computed is a sample regression line This implies that the constant a and the regression coefficient b are for the sample. It is advisable to make some refinement for providing an interval within which the true population regression line lies.

GLOSSARY

Dependent variable	The variable, which we are trying to predict with the help of an estimating equation.
Estimated or predicted value of Y	The value of the dependent variable denoted by $\hat{Y}$, which is obtained for a given value of X by using the estimated regression model.
Estimating equation	An equation used in regression analysis, which relates the unknown variable to the known variable(s).
Independent variable	The known variable or variables in a regression analysis.
Linear regression model	A regression model that gives a straight line relationship between two variables.
Multiple regression	It is concerned with relationship involving more than one independent variable.
Non-linear regression model	A regression model that does not give a straight line relationship between two variables.
Regression coefficient	The slope 'b' of the regression equation $Y = a + bX$.
Regression line	A line of best fit, which can always be found for a scatter diagram by using the method of least squares.
Regression of X on Y	The equation taking the form $X = a + bY$, which minimises the scatter in the X variable or horizontal direction.
Regression of Y on X	The equation taking the form $Y = a + bX$, which minimises the scatter in the Y variable or vertical direction.
Regression	A method that uses past data to estimate the relationship between two variables.
Scatter diagram	A plot of the paired observations of X and Y.
Slope (b)	It is the coefficient of the X term in the equation $Y = a + bX$. It shows how much the dependent variable changes for one unit change in the independent variable. When positive, it gives the increase in Y per unit increase in X.
Standard error of estimate	A measure to determine the extent to which observed values differ from their predicted values on the regression line.

Standard error of the regression coefficient	A measure, which shows the variability of sample regression from the population regression coefficient.
The least squares method	A method that minimises the average squared deviations. It shows how close the line fits the given data.
Y-intercept	It is the value of Y where the line $Y = a + bX$ cuts the Y-axis. It is the constant term in the estimating equation.

LIST OF FORMULAE

1. Simple linear regression model:

$$Y = a + bX$$

This is the equation of a straight line where the variable Y depends on the variable X, which is an independent variable.

2. Slope of the regression line:

$$b = (Y_2 - Y_1)/(X_2 - X_1)$$

'b' is the slope of the line. To calculate its numerical value, the values of coordinates X and Y for two points are to be found. The first point's coordinates are X_1 and Y_1 and those of the second point are X_2 and Y_2.

3. Slope of the best-fitting regression line:

$$b = \frac{\sum XY - n\,\bar{X}\,\bar{Y}}{\sum X^2 - n\,\bar{X}^2}$$

This formula is used to calculate the slope of the best-fitting regression line pertaining to a given two-variable data set.

4. Regression equation of Y on X:

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

5. Regression equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

6. Regression coefficient of Y on X:

$$r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} = b_{yx}$$

7. Regression coefficient of X on Y:

$$r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2} = b_{xy}$$

8. Coefficient of correlation based on the two regression coefficients:

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

9. Standard error of estimate, *Se*, measuring the variability of the observed values around the regression line.

$$Se = \sqrt{\frac{\Sigma(Y_i - \hat{Y})^2}{n-2}}$$

where Y_i is the original value and $\hat{Y}$ is the corresponding calculated value of Y variable.

10. Simplified method of calculating standard error of estimate:

$$\sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2}}$$

11. Confidence interval for B

$b \pm tS_b$ where $S_b = \dfrac{Se}{\sqrt{\Sigma X^2 - n\bar{X}^2}}$

S_b is the standard error of b.

12. Test statistic for a test of hypothesis about B

$$t = (b - B)/S_b$$

QUESTIONS

16.1 Given below are ten statements. Indicate in each case whether the statement is true or false:

(a) Given that the equation of a line is $Y = 30 - 11X$, it can be said that the relationship between X and Y is direct and linear.

(b) Both regression and correlation analyses are used to ascertain cause-and-effect relationships.

(c) In the equation $Y = a + bX$ (where Y and X are dependent and independent variables, respectively), 'a' is the Y-intercept.

(d) Regression analysis is used to determine how well an estimating equation describes the relationship between the two variables.

(e) An estimating equation $Y = 10 - 15X$ shows that there is an inverse relationship between X and Y.

(f) In a scatter diagram, if a straight line is drawn, there will be some points on either side of the line.

(g) The least square method always minimises the sum of the squares of the error (SSE) in the estimating line.

(h) The square root of the product of two regression coefficients does not give the value of the coefficient of correlation.

(i) Higher the magnitude of the standard error of estimate, the greater is the variability of points around the regression line.

(j) The regression line is not derived from the entire population, it is derived from a sample.

Multiple Choice Questions (16.2 to 16.13)

16.2 While solving a problem, we get an estimating equation $\hat{Y} = 10 - 4x$. Which of the following indicates this situation?
(a) The slope of the line is negative
(b) The Y-intercept of the line is 4
(c) There is an inverse relationship
(d) (a) and (c)
(e) (a) and (b)

16.3 When the regression line of Y on X and the regression line of X on Y form a 90° angle
(a) $r = \pm 1$ (b) $r = 0$ (c) $r = \pm 0.5$ (d) none of these

16.4 Assuming that after solving a problem, you have obtained the regression equation which calculated a is 6 and b is 3. The problem relates to one dependent and one independent variable. If the independent variable has a value of 5, what should be the value of the dependent variable?
(a) 15 (b) 18 (c) 10 (d) 21

16.5 Which of the following shows the most correct variation of the Y values around the regression line?
(a) $\Sigma(Y - \bar{Y})^2$ (b) $\Sigma(Y + \bar{Y})^2$ (c) $\Sigma(Y + \hat{Y})^2$ (d) $\Sigma(Y - \hat{Y})^2$

16.6 When we divide S $(Y - \hat{Y})^2$ by S $(Y - \bar{Y})^2$, the resultant figure shows:
(a) The fraction of total variation in Y that is explained.
(b) The fraction of total variation in Y on account of changes in X.
(c) The fraction of total variation in Y that remains unexplained.
(d) None of these.

16.7 What will be the range of r when we find that the dependent variable increases as the independent variable increases?
(a) 0 to –0.05 (b) 1 to 2 (c) 0.1 to 1 (d) none of these

16.8 When the two regression lines coincide, then r is
(a) 0 (b) –1 (c) 1 (d) 0.5
(e) none of these

16.9 Which of the following is true if the estimating equation has to be a perfect estimator of the dependent variable?
(a) The coefficient of determination is –1.
(b) All the data points are on the regression line.
(c) The standard error of the estimate is zero.
(d) (b) and (c)
(e) (a), (b) and (c)

16.10 While testing regression relationship we formulate $H_0 : B = 0$ and $H_1 : B \neq 0$. Which of the following has to be calculated first to verify H_0?
(a) S_p (b) S_b
(c) S_e (d) No particular order is required

16.11 Coefficient of determination, r^2 can be written as
(a) SST/SSR (b) 1 – (SSE/SST)
(c) SSR/SST (d) (b) and (c)
(e) none of these

16.12 Assuming Y as advertising expenditure and X as sales, which is the correct expression of regression of Y on X?

(a) $(X - \bar{X}) = r \frac{\sigma_y}{\sigma_x} (Y - \bar{Y})$ (b) $(Y - \bar{Y}) = r (X - \bar{X})$

(c) $(Y - \bar{Y}) = \frac{\sigma_y}{\sigma_x} (X - \bar{X})$ (d) $(Y - \bar{Y}) = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$

(e) None of these

16.13 Which of the following is denoted by the standard error of estimate S_{yx}?

(a) linearity (b) closeness (c) variability (d) none of these

16.14 Explain the concept of 'regression'. How is it important in economic analysis?

16.15 Differentiate between **(a)** linear and curvilinear relationships, and **(b)** direct and inverse relationships.

16.16 What are regression lines? Why should there be, in general, two lines of regression for each bivariate distribution?

16.17 Suppose that you wish to fit a regression line without an intercept, that is, $Y = bX$. Will the regression line pass through the points $\bar{Y}$, $\bar{X}$?

16.18 What is a scatter diagram? Illustrate your answer with a hypothetical example.

16.19 Explain the term 'standard error of the estimate'. What is its relevance in regression analysis?

16.20 What is an estimating equation in regression analysis?

16.21 What do you understand by the Method of Least Squares and the Least Squares Regression Line? Why are they called by these names?

16.22 State some of the important properties of regression coefficients. How are these helpful in analysing the regression lines?

16.23 Point out the role of regression analysis in business and industry. What are the properties of regression coefficients?

16.24 What are the assumptions involved in the use of regression analysis?

16.25 Distinguish between correlation and regression. Why are there two regression lines? Do they cut each other? If so, where?

16.26 The following data relate to training and performance of salesmen employed in a company.

Salesman	*Hours of Training (X)*	*Performance (Average weekly sales in '000 Rs) (Y)*
1	20	44
2	5	22
3	10	25
4	13	32
5	12	27

Fit a regression line to the above data. Find out the weekly sales that are likely to be attained by a salesman who is given 16 hours of training.

16.27 An investigation into the demand for television sets in 7 towns has resulted in the following data:

Town	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Population (lakh) (*X*)	11	14	14	17	17	21	25
No. of TV sets demanded ('000) (*Y*)	15	27	27	30	34	38	46

Fit a linear regression of *Y* on *X* and estimate the demand for TV sets for a city with a population of **(a)** 20 lakh and **(b)** 32 lakh.

16.28 Fit a linear regression of *Y* on *X* for the following data:

Output ('000 Rs)	*X*	5	*7*	9	11	13	15
Profit per unit of output (Rs)	*Y*	2	3	3	4	4	5

16.29 Construct a scatter diagram for the data given below:

X	3	4	6	7
Y	5	10	9	12

Determine the least-squares regression equation for estimating *Y* when *X* is known. Fit the regression line on the scatter diagram.

16.30 For 10 observations on price (*X*) and supply (*Y*), the following data were obtained:

$\Sigma X = 130$ $\Sigma Y = 220$ $\Sigma X^2 = 2288$

$\Sigma Y^2 = 5506$ $\Sigma XY = 3467$

Fit a linear regression of *Y* on *X* and estimate the supply when the price is 25 units.

16.31 Find the regression of *Y* on *X* from the following data and estimate the probable sales of a salesman having a test score of **(a)** 75; and **(b)** 100.

Salesman	*Test Scores (X)*	*Sales ('000 Rs) (Y)*
1	40	2.5
2	70	6.0
3	50	4.5
4	60	5.0
5	80	4.5
6	50	2.0
7	90	5.5
8	40	3.0
9	60	4.5
10	60	3.0

16.32 Fit a regression line of consumption (*C*) and income (*Y*) on the basis of the following data:

C ('000 Rs)	90	95	100	105	106
Y ('000 Rs)	100	110	120	130	140

Estimate consumption when income is Rs 150000.

16.33 Fit a regression line of yield (Y) on the fertilizer (X) on the basis of the following data:

Amount of Fertilizer Used (kg) (X)	*Yield (Y)*
2	7
3	9
4	10
5	14
6	15

Estimate the yield when the amount of fertilizer is 7 kg.

16.34 A company manufactures different types of electrical appliances. It has been using radio for advertising its products. The following table shows the amounts of radio time (X, in minutes) and the number of electrical appliances sold (Y) over the last seven weeks.

X	25	18	32	21	35	28	30
Y	16	11	20	15	26	32	20

(a) Fit a regression of Y on X.

(b) What will be the value of Y when X is 27?

16.35 The editor-in-chief of a major metropolitan newspaper has been trying to convince the paper's owner to improve the working conditions in the press room. He is convinced that the noise level, when the presses are running, creates unhealthy levels of tension and anxiety. He recently had a psychologist conduct a test during which pressmen were placed in rooms with varying levels of noise and then given a test to measure mood and anxiety levels. The following table shows the index of their degrees of nervousness and the level of noise to which they were exposed. (5.0 is low and 10.0 is high.)

Noise Level	*Degree of Nervousness*
7.0	23
6.5	38
5.5	45
6.0	36
8.0	16
8.5	18
6.0	39
6.5	41

(a) Develop an estimating equation, using the method of least squares.

(b) Predict the degrees of nervousness that we might expect when the noise level is 7.5.

16.36 From the following data obtain the two regression equations:

Sales	*Purchase*
91	71
97	75
108	69
121	97

(*Contd.*)

(*Contd.*)

67	70
124	91
51	39
73	61
111	80
57	47

16.37 A survey was conducted to study the relationship between the sales and advertising expenditure. Estimate (i) the sales for advertising expenditure of Rs 90 lakh, (ii) the advertising expenditure for a sales target of Rs 25 crore, and (iii) their correlation. Use regression technique to solve the problem.

Sales (Rs crore)	10	11	13	15	16	19	14
Ad. Exp. (Rs lakh)	60	62	65	70	73	75	71

16.38 To investigate the relationship between the level of advertising in local newspapers and the level of sales, the marketing manager of a national firm in the consumer-products field applies different amounts of advertising in 10 randomly selected geographic areas. The following table indicates the level of advertising and the level of sales for the relevant time period in the 10 areas.

Area	*Level of Advertising ('00 Rs)*	*Level of Sales ('000 Rs)*
1	10.5	17.3
2	6.0	14.0
3	8.7	19.1
4	9.3	14.5
5	11.8	20.0
6	7.5	16.3
7	15.0	23.8
8	6.3	14.0
9	8.5	17.3
10	5.4	13.3

(i) Determine regression equation for estimating the level of sales based on the level of advertising.

(ii) Estimate the level of sales for an area in which Rs 1000 is spent on advertising, as a point estimate.

16.39 For the following data:

	Adv. Exp. (Rs crore)	*Sales (Rs crore)*
Mean	20	120
Standard deviation	5	25
Coefficient of correlation = 0.8		

(i) Obtain the two regression equations.

(ii) Find the likely sales when advertising expenditure is Rs 25 crore.

(iii) What should be the advertising budget to achieve the sales target of Rs 150 crore?

16.40 A firm that sells office supplies wants to expand. The head of the firm wants to know what sales volume can be expected in various market areas. Regression analysis with sales as the dependent variable, is suggested. It is decided that effective buying income would be the best independent variable. A sample of 15 trade areas in which the firm now does business gives the following results in lakh of rupees.

Sum of $X = 1385$ Sum of $Y = 83.6$

Sum of $X \times Y = 10917.6$ Sum of X squares = 179661

Sum of Y squares = 681.32

(i) Develop the equation that best describes the relationship between effective buying income and sales.

(ii) For a trade area with an effective buying income of 115, what is the estimated amount of sales?

(iii) What is the correlation coefficient for these data? Is it an appropriate measure that enables the manager to determine the proportion of variability in sales explained by the effective buying income? Explain.

16.41 As a part of a study on transportation safety, a state government collected data on number of fatal accidents per 1000 licences and percentage of licensed drivers under the age of 21 in 15 cities. The data are as follows:

Fatal accidents per 1000 licences	2.962	0.708	0.885	1.652	2.091	2.627	3.830	0.368
Per cent of licensed drivers	13	12	8	12	11	17	18	8
Fatal accidents per 1000 licences	1.142	0.645	1.028	2.801	1.405	1.433	0.039	
Percent of licenced drivers	13	8	9	16	12	9	10	

Fit a linear regression to the above data to forecast fatal accidents per 100 licences. Is the fit good?

16.42 Obtain the two lines of regression for the following bivariate frequency distribution:

Sales Revenue	*Adv. Exp. ('000 Rs)*				
('000 Rs)	*5–15*	*15–25*	*25–35*	*35–45*	*Total*
75–125	4	1	—	—	5
125–175	7	6	2	1	16
175–225	1	3	4	2	10
225–275	1	1	3	4	9
Total	**13**	**11**	**9**	**7**	**40**

16.43 Two efficiency tests were conducted by an operations manager on 300 workers of a manufacturing system, which produced the following results:

	Test A	*Test B*
Mean efficiency score	75	70
Standard deviation	6	8
Coefficient of correlation between the efficiency scores of Test A and Test B = + 0.72		

Obtain the regression equation of Test B on Test A and predict the efficiency score of a worker in Test B whose efficiency score in Test A is 65.

16.44 The following data are given for marks in English and Mathematics at a certain examination:

	English	*Mathematics*
Mean marks	39.5	47.5
S.D. marks	10.8	16.8
$r = 0.8$		

Find the two regression equations. Using these regressions, estimate the value of Y for $X = 50$ and the value of X for $Y = 30$.

16.45 Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Question 16.26.

16.46 Calculate the sample coefficient of determination and the sample coefficient of correlation for the data in Question 16.32.

16.47 A study was conducted to find out whether there is an association between consumer's perceptions of a TV commercial (measured on a special scale) and their interest in purchasing the product (measured on a scale). The number of respondents was $n = 65$ and the correlation coefficient was $r = 0.37$. Do you think that there is statistical evidence of a linear correlation between the two variables?

16.48 The following data relate to price-earning ratios and growth per cent per year:

X	P/E	10	12	15	17	12	22	18
Y	% Growth	15	20	35	25	30	20	40

Determine the regression of Y on X and conduct the hypothesis test at 0.01 level of significance for the existence of a linear relationship between the two variables.

16.49 The following data relate to hours of training and performance of five salesmen. Determine the regression of performance on hours of training. Give 95% confidence intervals for the regression slope and the regression intercept parameters.

Saleman	*Hours of training* X	*Performance* Y
1	20	44
2	5	22
3	10	25
4	13	32
5	12	27

16.50 Responses measured in 10-point interval scale about the preference of a product based on package design from 10 consumers show the following result:

Consumer	1	2	3	4	5	6	7	8	9	10
Preference	9	3	2	7	3	2	8	1	2	6
Package design	8	4	1	9	3	3	7	1	4	3

(i) Fit a linear regression model of preference on package design.
(ii) Test the validity of the equation statistically.
(iii) Comment on the strength of association.

16.51 Using the data given in Q. 16.32, determine by applying an appropriate test whether the two variables are linearly associated. Also estimate the confidence interval of the regression co-efficient.

16.52 Conduct the F test for the existence of a linear relationship between the two variables in Q. 16.29.

16.53 In a simple linear regression analysis, it is found that regression coefficient b is 2.5 and the standard error of the slope is 4.2. The sample size is 15. Conduct an F test for the existence of a linear relationship between the two variables.

16.54 The advertisement cost and effected sales are given in the following table. Calculate the line of regression of sales on advertisement expenses (cost) and estimate the sales when the advertisement cost is Rs 1,00,000.

Advt. cost (Rs '000)	39	65	62	90	82	75	25	98	36	78
Sales (Rs lakhs)	47	53	58	86	62	68	60	91	51	84

Calculate the coefficient of correlation between advertisement cost and sales.

16.55 Given the two regression lines Y on X and X on Y as $Y = 67.72 + 0.48X$ and $X = -41.1375 + 0.9075Y$, find $\bar{X}$, $\bar{Y}$, σx and σy.

16.56 Obtain the equations of the two lines of regression for the data given below:

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

16.57 The following data relate to the sales of a department store for 7 years. Using the regression line, find the projected sales for 2002.

Year	1995	1996	1997	1998	1999	2000	2001
Sales (Rs lakhs)	80	90	92	83	94	99	92

16.58 Obtain the two regression equations from the following data:

X	27	27	27	28	28	29	29	29	30	31
Y	18	18	19	20	21	21	22	23	24	25

16.59 From the following data, obtain the two regression equations

X	6	2	10	4	8
Y	9	11	5	8	7

16.60 Given the following bivariate data, find the regression equation.

X	–1	5	3	2	1	1	7	3
Y	–6	1	0	0	1	2	1	5

Predict the value of Y where $X = 4$.
Predict X when $Y = 2.5$.

16.61 Construct two regression equations for the following data:

A	23	43	53	63	73	83
B	5	6	7	8	9	10

16.62 The following data are related to the sales and advertising expenditure of a firm

	Sales (Rs Crore)	*Ad expenditure (Rs Crore)*
Mean	40	6
S.D.	10	1.5
r	0.9	

(i) Estimate the likely sales for proposed expenditure of Rs 10 crore on advertising.
(ii) What should be the advertising expenditure if the firm proposes sales target of Rs 60 crore?

16.63 A student was given two regression lines, as follows:

$$3X + 2Y = 26 \quad \text{and} \quad 6X + Y = 31$$

He obtained the mean values, $\bar{X} = 7$, and $\bar{Y} = 4$ and the coefficient of correlation, + 0.5. Do you agree with him? If not, indicate your results.

The McGraw-Hill Companies

CHAPTER 17 CORRELATION

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the importance, as also the limitations, of correlation analysis
- distinguish between (a) linear and non-linear correlation, (b) positive and negative correlation, and (c) simple, partial and multiple correlation
- recognise when a scatter diagram suggests relationship between two variables
- calculate and interpret coefficient of correlation for individual observations as well as for bivariate grouped data.

Chapter Prerequisites

Before starting work on this chapter, make sure you are quite conversant with

1. calculation and interpretation of mean and standard deviation
2. the basic idea of graph plotting.

17.1 CONCEPT AND IMPORTANCE OF CORRELATION

Many times we come across problems or situations where two variables seem to move in the same direction such as both are increasing or decreasing. At times an increase in one variable is accompanied by a decline in another. Such changes in variables suggest that there is a certain relationship between them. For example, take the case of monthly earnings of households as one variable and monthly expenses on entertainment as another. If we collect data for a couple of years, we may find that both are associated. Let us take another example of two variables—price of a commodity and the demand for it. We may find that when price increases, there is a decline in its demand. Here, the two variables, price and demand, are found to be moving in opposite directions. Such associations are studied in correlation analysis.

Correlation analysis is used as a statistical tool to ascertain the association between two variables. The problem in analysing the association between two variables can be broken down into three steps.

1. We try to know whether the two variables are related or independent of each other.

2. If we find that there is a relationship between the two variables, we try to know its nature and strength. This means whether these variables have a positive or a negative relationship and how close that relationship is.
3. We may like to know if there is a causal relationship between them. This means that the variation in one variable causes variation in another.

It may be noted that correlation analysis is one of the most widely used statistical techniques adopted by applied statisticians. In the early days, its use was confined to biological problems. Subsequently, it has been used extensively in agriculture, economics, business and several other fields. In this chapter, we shall consider correlation analysis for two variables and in Chapter 18 it will be extended to cases where more than two variables are involved.

There are several *reasons*, which show clearly the importance of correlation analysis.

1. In several cases, when we deal with statistics, we find that the variables are related to each other. For example, take the case of income and consumption expenditure. With the help of correlation analysis, we can be very specific by measuring the degree of relationship between the concerned variables.
2. In business, correlation analysis can be extremely helpful. It can enable the management to estimate costs, sales, prices and other variables on the basis of some other series that is closely related to these.
3. When a specific and reliable relationship has been established between two variables, we can find the value of a variable given the value of another. In fact, this was done with the help of regression analysis that was discussed in the preceding chapter. This also shows that the two concepts, correlation analysis and regression analysis, are closely involved.

17.2 CORRELATION AND CAUSATION

When we use correlation analysis and establish a relationship between two variables, then we confront a major question: does this relationship indicate the existence of cause and effect relationship? It may be noted that there may be a very high degree of relationship between two variables, but they may just show similar movements and causal relationship is non-existent. Let us see when such a situation arises.

1. The correlation may be due to chance particularly when the data pertain to a small sample. A small sample bivariate series may show the relationship but such a relationship may not exist in the universe.
2. It is possible that both the variables are influenced by one or more other variables. For example, expenditure on food and entertainment for a given number of households show a positive relationship because both have increased over time. But, this is due to rise in family incomes over the same period. In other words, the two variables have been influenced by another variable— increase in family incomes.
3. There may be another situation where both the variables may be influencing each other so that we cannot say which is the cause and which is the effect. For example, take the case of price and demand. The rise in price of a commodity may lead to decline in the demand for it. Here, price is the cause and the demand is the effect. In yet another situation, an increase in demand may lead to rise in its price. Here, the demand is the cause while price is the effect, which is just the reverse of the earlier situation. In such a situation, it is difficult to identify which variable is causing the effect on which variable, as both are influencing each other.

The foregoing discussion clearly shows that correlation does not indicate any causation or functional relationship. Even when there is no cause-and-effect relationship in bivariate series and one interprets the relationship as causal, such a correlation is called *spurious* or *non-sense correlation*. Obviously, this will be misleading. As such, one has to be very careful in correlation exercises and look into other relevant factors before concluding a cause-and-effect relationship.

17.3 TYPES OF CORRELATION

Correlation may be of different types. Some of the most important types are:

(i) Positive and negative
(ii) Linear and non-linear
(iii) Simple, partial and multiple.

We briefly describe these types of correlation.

(i) Positive and Negative Correlation Positive correlation indicates that the movement of the two variables is in the same direction, that is, both the variables are either increasing or decreasing. In contrast, if the movement of the two variables is in the opposite direction, that is, one variable is increasing and the other is decreasing, then the correlation is negative.

(ii) Linear and Non-linear Correlation If the extent of change in one variable tends to have a constant ratio in the extent of change in another variable, then the correlation is said to be linear. This will be clear from the following example.

X	5	10	15	20	25
Y	30	60	90	120	150

In this case, we find that the ratio of change from one figure to another in the two series is the same. Thus, it will give a linear correlation. In contrast, in a non–linear correlation, this consistency of ratio of change will not exist. If a couple of figures in either series X or Y are changed, it may give a non–linear correlation.

(iii) Simple, Partial and Multiple Correlation The distinction amongst these three types of correlation depends upon the number of variables involved in a study. If only two variables are involved in a study, then the correlation is said to be simple correlation. When three or more variables are involved in a study, then it is a problem of either partial or multiple correlation. In multiple correlation, three or more variables are studied simultaneously. But in partial correlation we consider only two variables influencing each other while the effect of other variable is held constant.

Suppose we have a problem comprising three variables X_1, X_2 and Y. X_1 is the number of hours studied, X_2 is I.Q. and Y is the number of marks obtained in the examination. In a multiple correlation, we will study the relationship between the marks obtained (Y) and the two variables, number of hours studied (X_1) and I.Q. (X_2). In contrast, when we study the relationship between X_1 and Y, keeping an average I.Q. as constant, it is said to be a study involving partial correlation.

The discussion given below is confined to linear correlation. The different methods of correlation can be depicted as follows:

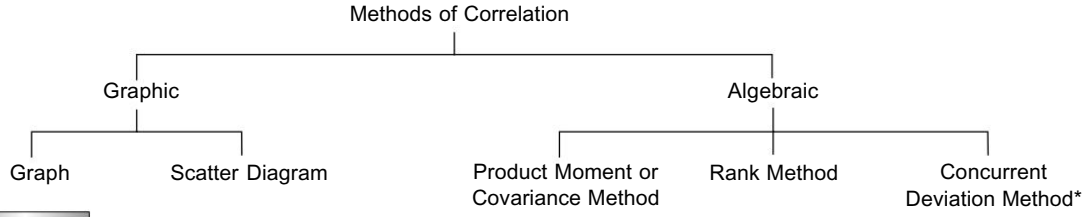


Fig. 17.1 **Different Methods of Correlation**

17.4 GRAPHIC METHOD OF CORRELATION

Graph

This method is very simple. The data pertaining to two series X and Y are plotted on a graph sheet. A visual observation of the graph will give a broad idea of the direction of movement and closeness of the two curves. If the curves move in the same direction, then the correlation is positive. If they move in the opposite direction, then the correlation is negative. If the graph does not show any definite pattern on account of erratic fluctuations in the curves, then it shows an absence of correlation.

Let us take a hypothetical example.

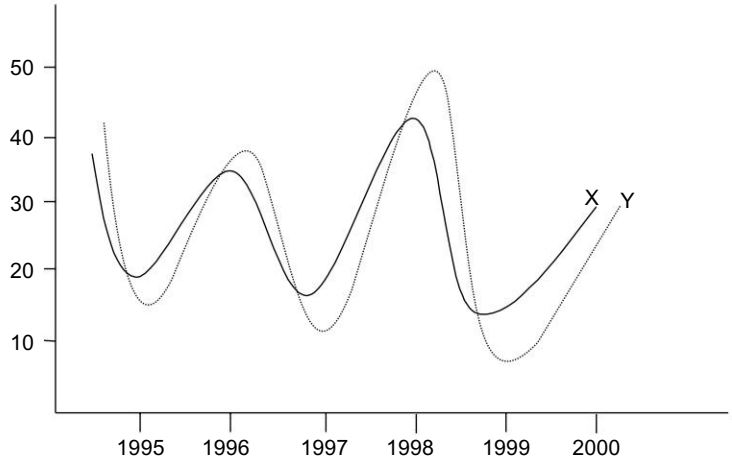


Fig. 17.2 **An Example of Graphic Correlation**

Figure 17.2 shows that the two curves move in the same direction and, moreover, they are very close to each other, suggesting a close relationship between the two variables X and Y. However, such a graph is unable to quantify the relationship between them.

Scatter Diagram

The other graphic method is scatter diagram. The independent variable is shown on the X-axis while the dependent variable on the Y-axis. A scatter diagram reveals whether the movements in one series

* This method is hardly used. As such this is not discussed here.

are associated with those in the other series. In case of a perfect correlation, the points will form on a straight line in a diagonal form. If this straight line is rising on the right, the correlation is positive and if it is falling, the correlation is negative.

Figure 17.3 shows different scatter diagrams. Each point in a diagram shows a pair of values (*X* and *Y*). It may be noted that scatter diagrams A and B show a high degree of positive and negative relationship between *X* and *Y*, respectively. Scatter diagram C shows that there is a non-linear relationship between the two variables. Scatter diagram D shows that there is no relationship between them.

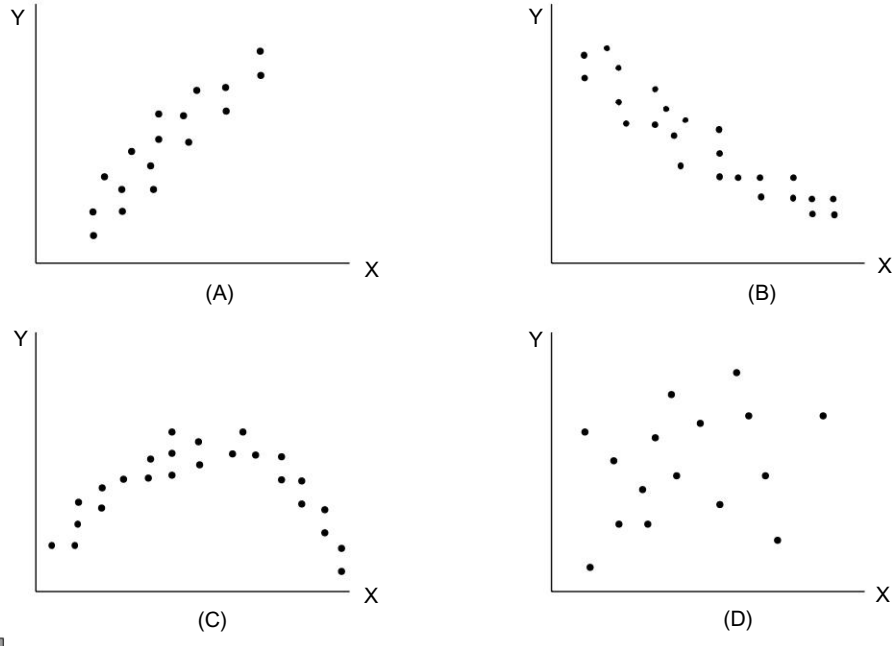


Fig. 17.3 Scatter Diagrams

Figure 17.4 shows three sets of scatter diagrams. Each set shows two regression lines-regression of *X* on *Y* and regression of *Y* on *X*. A careful look at the scatter diagrams gives us an intutive notion of weak and strong correlation as well as a notion of a bivariate relationship.

It can be seen that when the correlation is strong ($r = \pm 0.9$), points lie close to the regression lines. When r is ± 0.5 points do not lie that close to the regression lines. When r is ± 0.1 points lie wide apart from the regression lines. Further, one finds that the angle between the two regression lines is smaller when correlation is strong. In contrast, when correlation is weak, the angle between the two regression lines is larger.

The correlation coefficient may be thought of as the geometric mean between the two slopes, viz. *bxy* and *byx*. Thus,

$$r = \sqrt{bxy \cdot byx}$$

where $bxy = \dfrac{\Sigma xy}{\Sigma y^2}$ and $byx = \dfrac{\Sigma yx}{\Sigma x^2}$

Regression of *x* on *y* Regression of *y* on *x*

In case the two slopes coincide with each other, then $r = 1$. When one of the slopes is zero, then $r = 0$. It may be noted that this formula does not indicate whether r is positive or negative, it takes the sign of r from bxy and byx. However, the above formula is of interest as it is convenient in connection with partial correlation, which will be considered in the next chapter.

We now come to algebraic methods.

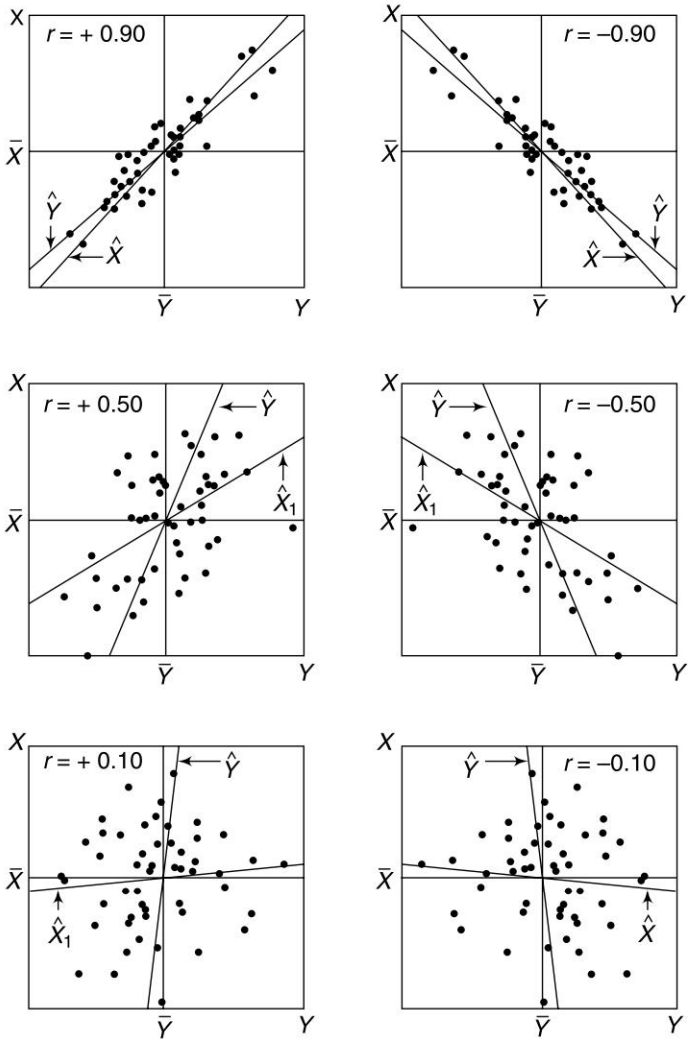


Fig. 17.4 Scatter Diagrams showing Varying Degrees of Correlation

Source: Adapted from Chart 15.1 given in Frederick E. Croxton, Dudley J. Cowden and Ben W. Bolch: *Practical Business Statistics* (4th ed.) Prentice Hall of India Private Limited, New Delhi, 1974.
Original Source: E. C. Fieller, T. Lewis and E. S. Pearson: *Correlated Random Normal Deviates,* Tracts for Computers No. XXVI (Cambridge: Cambridge University Press, 1955).

17.5 ALGEBRAIC METHODS OF CORRELATION

Karl Pearson's Method

Karl Pearson's method of calculating coefficient of correlation is based on the covariance of the two variables in a series. In fact, there are two methods to calculate the coefficient of correlation. These are:

(i) Direct method
(ii) Short-cut method

Direct Method In this method, the coefficient of correlation, denoted by the symbol '*r*', is calculated as the ratio of the covariance of the two variables to the product of their standard deviations.

Symbolically,

$$r = \frac{\Sigma xy}{N\sigma_x \cdot \sigma_y}$$

where $x = (X - \overline{X})$, $y = (Y - \overline{Y})$, Σxy = sum of the product of deviations in X and Y series from their arithmetic means, σ_x = standard deviation of the series X, σ_y = standard deviation of series y and N = total number of pair of observations.

A simple version of this formula is

$$\frac{\Sigma xy}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}}$$

This has been derived as follows:

$$r = \frac{\Sigma xy}{N\sigma_x \cdot \sigma_y} \quad \sigma_x = \sqrt{\frac{\Sigma x^2}{N}} \quad \sigma_y = \sqrt{\frac{\Sigma y^2}{N}}$$

$$= \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N}} \times \sqrt{\frac{\Sigma y^2}{N}}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}}$$

It may be noted that this formula can be applied only when the deviations are taken from the actual arithmetic means of the two series. It is sometimes known as the *covariance method* or the *product-moment method*.

We shall now take an example to show the process of calculation of coefficient of correlation. But, before attempting the calculation of coefficient of correlation, we should know the steps involved in its calculation.

Calculating Coefficient of Correlation by Direct Method The steps of this calculation are:

1. Calculate the means of the two series, $\overline{X}$ and $\overline{Y}$.
2. Take deviations in the two series from their respective means, indicated as x and y. The deviation should be taken in each case as the value of the individual item minus (–) the arithmetic mean.
3. Square the deviations in both the series and obtain the sum of the deviation-squared columns. This would give Σx^2 and Σy^2.
4. Take the product of the deviations, that is, Σxy. This means individual deviations are to be multiplied by the corresponding deviations in the other series and then their sum is obtained.

5. The values thus obtained in the preceding steps Σxy, Σx^2 and Σy^2 are to be used in the formula for correlation, given earlier.

We take an example to illustrate the use of this formula.

Example 17.1 Calculate the coefficient of correlation for the following data:

X	2	3	4	5	6
Y	7	9	10	14	15

Solution We set up Table 17.1 for carrying out the necessary calculations.

Table 17.1 Calculation of Correlation Coefficient

X	Y	$X-\bar{X}$ x	x^2	$Y-\bar{Y}$ y	y^2	xy
2	7	–2	4	–4	16	8
3	9	–1	1	–2	4	2
4	10	0	0	–1	1	0
5	14	1	1	3	9	3
6	15	2	4	4	16	8
20	**55**		**$\Sigma x^2 = 10$**		**$\Sigma y^2 = 46$**	**$\Sigma xy = 21$**

$$\bar{X} = \Sigma X/N = 20/5 = 4$$
$$\bar{Y} = \Sigma Y/N = 55/5 = 11$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}}$$

$$= 21/\left(\sqrt{10} \times \sqrt{46}\right)$$
$$= 21/(3.16 \times 6.78)$$
$$= 0.98$$

The value of $r = 0.98$ shows that the two series X and Y have almost perfect positive correlation. It may be noted at this stage that the coefficient of correlation always ranges between –1 and +1. If r is 0, then it is clear that the two series are not at all associated with each other. If r is –1, it shows that there is a perfect negative correlation. If r is +1, then it is a case of perfect positive correlation between the two series.

Short-cut Method In the preceding example, we have used arithmetic means of the two series for calculating the deviations of individual observations from them. A number of times the arithmetic mean may be a fractional value; it then becomes difficult to use it for calculating deviations. In such cases, deviations are taken from assumed mean. This method is known as the short-cut method.

Calculating Coefficient of Correlation by Short-cut Method Following steps are involved in calculating coefficient of correlation by this method:

1. Choose convenient values as assumed means of the two series, X and Y.
2. Deviations (now dx and dy instead of x and y) are obtained from the assumed means in the same manner as in the earlier example.
3. Obtain the sum of the dx and dy columns, that is, Σdx and Σdy.
4. Deviations dx and dy are squared up and their totals Σdx^2 and Σdy^2 are obtained.
5. Finally, obtain $\Sigma dxdy$, which is the sum of the products of deviations taken from the assumed means in the two series.

Now, a numerical example follows.

Example 17.2 A company manufactures different types of electrical appliances. It has been using radio for advertising its products. The following table shows amounts of radio time (X, in minutes) and the number of electrical appliances sold (Y) over the last six weeks.

X	25	18	32	21	35	29
Y	16	11	20	15	26	28

Calculate the coefficient of correlation between the two series.

Solution

Table 17.2 Calculation of Correlation Coefficient

X	dx (A.M.=25)	dx^2	Y	dy (A.M.=20)	dy^2	$dx.dy$
25	0	0	16	–4	16	0
18	–7	49	11	–9	81	63
32	7	49	20	0	0	0
21	–4	16	15	–5	25	20
35	10	100	26	6	36	60
29	4	16	28	8	64	32
$\Sigma X = 160$	**$\Sigma dx = 10$**	**$\Sigma dx^2 = 230$**	**$\Sigma Y = 116$**	**$\Sigma dy = -4$**	**$\Sigma dy^2 = 222$**	**$\Sigma dxdy = 175$**

Note: 'A.M.' stands for arbitrary or assumed mean.

The formula for calculating r *by the short-cut method is*

$$r = \frac{\Sigma dxdy - \dfrac{\Sigma dx \times \Sigma dy}{N}}{\sqrt{\left(\Sigma dx^2 - \dfrac{(\Sigma dx)^2}{N}\right)\left(\Sigma dy^2 - \dfrac{(\Sigma dy)^2}{N}\right)}}$$

Applying the above values in this formula,

$$r = \frac{175 - \dfrac{10 \times -4}{6}}{\sqrt{\left(230 - \dfrac{(10)^2}{6}\right)\left(222 - \dfrac{(-4)^2}{6}\right)}}$$

$$= \frac{175 - \left(\frac{-40}{6}\right)}{\sqrt{\left(230 - \frac{(100)}{6}\right)\left(222 - \frac{(16)}{6}\right)}}$$

$$= \frac{\frac{1090}{6}}{\sqrt{\frac{1280}{6} \times \frac{1316}{6}}} = \frac{1090 \times 6}{6\sqrt{1280 \times 1316}}$$

$$= \frac{1090}{\sqrt{16,84,480}} = \frac{1090}{1298} = 0.84$$

This shows a high degree of correlation between amounts of radio time (X) and the number of electrical appliances sold (Y).

If we examine carefully the formula used for the short-cut method with that used when actual arithmetic means were used, we would find that $(\Sigma dx \times \Sigma dy)/N$ is the correction factor in the numerator and $(\Sigma dx)^2/N$ and $(\Sigma dy)^2/N$ are the correction factors introduced in the denominator on account of the use of assumed means in place of actual means.

Alternative Method There is yet another method of calculating the coefficient of correlation. Instead of taking deviations from the assumed mean, we can calculate r by using the original values in the two series. In such a case, the formula would be

$$r = \frac{\Sigma XY - \frac{\Sigma X \times \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}}$$

Example 17.3 We shall now use this formula in respect of our previous data.

Solution

Table 17.3 Calculation of Correlation Coefficient

X	X^2	Y	Y^2	XY
25	625	16	256	400
18	324	11	121	198
32	1024	20	400	640
21	441	15	225	315
35	1225	26	676	910
29	841	28	784	812
$\Sigma X = 160$	**$\Sigma X^2 = 4480$**	**$\Sigma Y = 116$**	**$\Sigma Y^2 = 2462$**	**$\Sigma XY = 3275$**

$$r = \frac{3275 - \frac{160 \times 116}{6}}{\sqrt{\left(4480 - \frac{(160)^2}{6}\right)\left(2462 - \frac{(116)^2}{6}\right)}}$$

$$= \frac{3275 - \frac{18560}{6}}{\sqrt{\left(4480 - \frac{25600}{6}\right)\left(2462 - \frac{13456}{6}\right)}}$$

$$= \frac{3275 - 3093.33}{\sqrt{(4480 - 4266.67)(2462 - 2242.67)}}$$

$$= \frac{181.67}{\sqrt{213.33} \times \sqrt{219.33}} = \frac{181.67}{14.61 \times 14.81} = 0.84$$

It will be seen that the alternative method gives the same answer ($r = 0.84$) as was given by the earlier method.

Correlation of Grouped Data

When we are given the grouped data to determine the correlation coefficient between the two variables, we have to proceed in a somewhat different manner. From the grouped data, we here mean a two-way frequency table wherein the values of one variable are in the rows while those of the other variables are in columns. These values can be either discrete or continuous. The frequencies in each class are shown in cells in the body of the table. The total of these cells will be the product of number of rows and number of columns. Normally, '*m*' is used to denote the number of rows and *'n'* to denote the number of columns. A two-way table is given below.

Table 17.4 Sales Revenue and Advertising Expenditure

Sales Revenue (Rs lakh)	*Advertising Expenditure (Rs lakh)*				*Total*
	5–15	15–25	25–35	35–45	
75–125	3	4	4	8	19
125–175	8	6	5	7	26
175–225	2	2	3	4	11
225–275	3	3	2	2	10
Total	**16**	**15**	**14**	**21**	**66**

It will be seen that the above table has four rows and four columns excluding the row and column for totals.

The formula used for calculating the coefficient of correlation is:

$$r = \frac{\Sigma fdxdy - \frac{\Sigma fdx \times \Sigma fdy}{N}}{\sqrt{\left(\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}\right)\left(\Sigma fdy^2 - \frac{(\Sigma fdy)^2}{N}\right)}}$$

This formula can be simplified as

$$r = \frac{N\Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{[N\Sigma fdx^2 - (\Sigma fdx)^2][N\Sigma fdy^2 - (\Sigma fdy)^2]}}$$

It may be noted that these formulae are similar to those used earlier when assumed mean was used. The only difference is that here the deviations are multiplied by the corresponding frequencies.

Steps for Calculating Correlation Coefficient for Grouped Data A number of steps are required before we can apply any one of the above formulae for calculating correlation coefficient. These *steps* are:

(i) Record the mid-points of the class intervals for both X and Y variables.

(ii) Choose an assumed mean in X series and calculate the deviations from it. The same procedure is to be used for Y series.

(iii) To simplify calculations, step deviations can be taken by dividing deviations by a common factor.

(iv) Obtain the product of dx and the corresponding frequencies in each cell. Write the figure thus obtained in the right-hand corner of each cell. The same procedure is to be followed for Y series. If this is inconvenient, an alternative of this is to write these values within brackets as we have done. This will give Σfdx and Σfdy.

(v) All the values obtained in (iv) above are to be added up to obtain $\Sigma fdxdy$.

(vi) Multiply dx with the respective frequencies, add them up to obtain Σfdx.

(vii) Multiply fdx in each cell by the corresponding dx to obtain Σfdx^2.

(viii) In the same manner, multiply dy with the respective frequencies, add them up to obtain Σfdy.

(ix) In the same manner as done in (vii) above, multiply dy and fdy to obtain Σfdy^2.

(x) Having obtained all the requisite values, viz. $\Sigma fdxdy$, Σfdx, Σfdy, Σfdx^2 and Σfdy^2, substitute them in one of the formulae given above.

Example 17.4 The data given in Table 17.4 are used to calculate the correlation coefficient. Table 17.5 shows the worksheet for this purpose. This is followed by the calculation of correlation coefficient.

Table 17.5 Computation of Correlation Coefficient

X / Y			5–15	15–25	25–35	35–45	Total			
		MV	10	20	30	40				
	MV	$dx \rightarrow$ / $dy \downarrow$	–1	0	1	2	*f*	*fdy*	fdy^2	*fdxdy*
75–125	100	–1	3(3)	4(0)	4(–4)	8(–16)	19	–19	19	–17
125–175	150	0	8(0)	6(0)	5(0)	7(0)	26	0	0	0
175–225	200	1	2(–2)	2(0)	3(3)	4(8)	11	11	11	9
225–275	250	2	3(–6)	3(0)	2(4)	2(8)	10	20	40	6
Total		***f***	**16**	**15**	**14**	**21**	**66**	**12**	**70**	**–2**
		fdx	–16	0	14	42	40			
		fdx^2	16	0	14	84	114			
		fdxdy	–5	0	3	0	–2			

Check

Solution A point worth noting is that when we subtract a common figure say 50 from X series such that $X' = X–50$ and another figure, say 70, from Y series such that $Y' = Y–70$, the data are simplified and calculations become easier. If calculations are made using the revised series X' and Y' (or X and Y', or X' and Y), the coefficient of correlation in each case will be the same. One can verify this by undertaking the necessary calculations.

$$r = \frac{\Sigma fdxdy - \frac{\Sigma fdx \times \Sigma fdy}{N}}{\sqrt{\left(\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}\right)\left(\Sigma fdy^2 - \frac{(\Sigma fdy)^2}{N}\right)}}$$

$$r = \frac{-2 - \frac{40 \times 12}{66}}{\sqrt{\left(114 - \frac{(40)^2}{66}\right)\left(70 - \frac{(12)^2}{66}\right)}}$$

$$= \frac{-2 - 7.27}{\sqrt{(114 - 24.24)\,(70 - 2.18)}}$$

$$= \frac{-9.27}{\sqrt{89.76} \times \sqrt{67.82}} = \frac{-9.27}{9.47 \times 8.24} = -0.119$$

In the above example, one worksheet shows all the data and as such it becomes unwieldy and confusing at times. In order to overcome this situation, we can show the calculations in different tables. Let us take another example to illustrate this procedure.

Example 17.5 Family income and its percentage spent on food in the case of 100 families gave the following frequency distribution:

Food (percentage)	*Family income (Rs '000)*				
	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	–

Calculate the coefficient of correlation and interpret the value.

Solution Let us take family income as X variable and expenditure on food (percentage) as Y variable. Taking mid–values of the class intervals, we have

Mid–values (X) : 7.5 12.5 17.5 22.5 27.5

Mid–values (Y) :12.5 17.5 22.5 27.5

Let $u_i = (X_i – 17.5)/5$ and $v_i = (Y_i – 22.5)/5$

Therefore, different u values are –2 –1 0 1 2

and different v values are –2 –1 0 1

Table 17.6 Calculation of Σfu and Σfu^2

u_i	f_i	f_iu_i	$f_iu_i^2$
–2	10	–20	40
–1	20	–20	20
0	40	0	0
1	20	20	20
2	10	20	40
Total	**100**	**0**	**120**

Table 17.7 Calculation of Σfv and Σfv^2

v_i	f_i	f_iv_i	$f_iv_i^2$
–2	10	–20	40
–1	20	–20	20
0	30	0	0
1	40	40	40
Total	**100**	**0**	**100**

Table 17.8 Calculation of *fuv*

v_i \ u_i	–2	–1	0	1	2	Total
–2	–	–	–	3 | –6	7 | –28	–34
–1	–	4 | 4	9 | 0	4 | –4	3 | –6	–6
0	7 | 0	6 | 0	12 | 0	5 | 0	–	0
1	3 | –6	10 | –10	19 | 0	8 | 8	–	–8
Total	**–6**	**–6**	**0**	**–2**	**–34**	**–48**

This gives $\Sigma fuv = -48$

$$r = \frac{\Sigma fuv - \dfrac{\Sigma fu \times \Sigma fv}{N}}{\sqrt{\left(\Sigma fu^2 - \dfrac{(\Sigma fu)^2}{N}\right)\left(\Sigma fv^2 - \dfrac{(\Sigma fv)^2}{N}\right)}}$$

$$r = \frac{-48 - \dfrac{0 \times 0}{100}}{\sqrt{\left(120 - \dfrac{(0)^2}{100}\right)\left(100 - \dfrac{(0)^2}{100}\right)}}$$

$$r = \frac{-48}{\sqrt{120} \times \sqrt{100}} = \frac{-48}{10.95 \times 10} = -0.438$$

Thus, we find that the coefficient of correlation is negative (–0.438). It is understandable as the income of families increases, the proportion of expenditure on food declines. This is the famous Engel's Law. The result of this exercise (the negative correlation) is in conformity with this law.

t Test for a Correlation Coefficient

In a number of examples given earlier in this chapter, the correlation coefficient r was computed to measure the extent of relationship between two variables. It may be noted that we often use the sample correlation coefficient for descriptive purposes as a point estimate of the population correlation coefficient ρ. This means r is used as if it is a parameter ρ, which it estimates, However, r can be used as an estimate of p provided the assumption of normal distribution of the two variables holds good.

The most frequently used test to examine whether the two variables X and Y are correlated is the t test. To apply this test, we first set up the two hypotheses as follows:

$H_0 : \rho = 0$ (Absence of correlation)

$H_1 : \rho \neq 0$ (Presence of correlation)

where p is the population correlation coefficient.

The formula used for the t test is as follows:

$$t = \frac{r - \rho}{\sqrt{(1 - r^2)/(n - 2)}}$$

where r is the sample correlation coefficient.

The test statistic t follows a t distribution with $n - 2$ degrees of freedom. Let us take an example.

Example 17.6 Refer to Example 17.3 where $r = 0.84$ and $n = 6$. Determine whether there is significant association between advertising expenditure and sales revenue.

Solution We apply the t test and use the formula

$$t = \frac{r - p}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$= \frac{0.84 - 0}{\sqrt{\frac{1 - (0.84)^2}{6 - 2}}}$$

$$= \frac{0.84}{\sqrt{\frac{1 - (0.7056)}{4}}}$$

$$= \frac{0.84}{\sqrt{\frac{0.2944}{4}}}$$

$$= \frac{0.84}{\sqrt{0.0736}} = 3.1$$

The critical value of t for 4 df at 0.05 level of significance is ±2.776. As the calculated value of t is more than the critical value, the null hypothesis is rejected. This means that there is statistically significant correlation between the two variables.

Example 17.7 Calculate the correlation coefficient between the two series, given below. Examine whether there is a significant relationship between the two variables.

X	10	20	30	40	50
Y	3	2	1	5	4

Solution

Worksheet

X	X^2	Y	Y^2	XY
10	100	3	9	30
20	400	2	4	40
30	900	1	1	30
40	1600	5	25	200
50	2500	4	16	200
$\Sigma X = 150$	$\Sigma X^2 = 5500$	$\Sigma Y = 15$	$\Sigma Y^2 = 55$	$\Sigma XY = 500$

$$r = \frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

$$= \frac{500 - \frac{(150)(15)}{5}}{\sqrt{\left[5500 - \frac{(150)^2}{5}\right]\left[55 - \frac{(15)^2}{5}\right]}}$$

$$= \frac{500 - 450}{\sqrt{(5500 - 4500)(55 - 45)}}$$

$$= \frac{50}{\sqrt{1000 \times 10}} = \frac{50}{\sqrt{10000}} = \frac{50}{100} = 0.5$$

In order to examine whether $r = 0.5$ indicates whether the relationship between X and Y is statistically significant, we apply the t test.

The two hypotheses are:

$H_0 : p = 0$ (No correlation)

$H_1 : p \neq 0$ (Correlation)

where p indicates correlation coefficient of the population.

The formula for the t test is

$$t = \frac{r - p}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$= \frac{0.5 - 0}{\sqrt{\frac{1 - (0.5)^2}{5 - 2}}}$$

$$= \frac{0.5}{\sqrt{\frac{0.75}{3}}}$$

$$= \frac{0.5}{\sqrt{0.25}} = \frac{0.5}{0.5} = 1$$

The critical value of t with 3 degrees of freedom and at 10% level of significance is 2.353. As the calculated value of t is less than the critical value of t, the null hypothesis is accepted. The conclusion is that the relationship between X and Y is not statistically significant.

Assumptions of the Karl Pearsonian Correlation

1. The two variables X and Y are linearly related. This implies that when the individual pairs are plotted on a graph resulting in a scatter diagram. If the points are joined together, a straight line will be formed.
2. The two variables are affected by several causes, which are independent, so as to form a normal distribution. For example, relationships between price and demand, price and supply, advertising expenditure and sales, length of experience and earnings and so on are affected by several factors such that the series result into a normal distribution.

17.6 COEFFICIENT OF DETERMINATION

When $r = 1$; or -1; or 0, the interpretation of r does not pose any problem. When $r = 1$; or -1, all the points lie on a straight line in a graph showing a perfect positive or a perfect negative correlation. When the points are extremely scattered on a graph, then it becomes evident that there is almost no relationship between the two variables. However, when it comes to other values of r, we have to be careful in its interpretation. Suppose we get a correlation of $r = 0.9$, we may say that $r = 0.9$ is 'twice as good' or 'twice as strong' as a correlation of $r = 0.45$. It may be noted that this comparison is wrong. The strength of r is judged by coefficient of determination, r^2 for $r = 0.9$, $r^2 = 0.81$. We multiply it by 100, thus getting 81 per cent. This suggests that when r is 0.9 then we can say that 81 per cent of the total variation in the Y series can be attributed to the relationship with X. When $r = 0.45$, $r^2 = 0.2025$ and in percentage terms it is 20.25. This means that $r = 0.45$ suggests that only 20.25 per cent of the total variations in the Y series can be attributed to the relationship with X. Now, we find that $r = 0.9$ is

four times as strong as $r = 0.45$ and not "twice as strong" as we earlier thought. The use of r^2 is a convenient way of interpreting r.

The relationship between r and r^2 can be seen from the following figures:

r	r^2	r	r^2
1.0	1.0	0.5	0.25
0.9	0.81	0.4	0.16
0.8	0.64	0.3	0.09
0.7	0.49	0.2	0.04
0.6	0.36	0.1	0.01

It will be seen that when r is maximum at 1, r^2 equates it. But as the value of r decreases from its maximum value of 1, there is sharper decline in r^2. It may also be noted that the value of r is always greater than that of r^2 except when $r = 0$ and 1. In both the cases, r^2 will also be 0 and 1. Another point to note is that r^2 will not be able to show the direction of r—whether it is positive or negative. If we are told that in a given problem r^2 is 0.81, we can only say that r is very high, being 0.9, but we cannot say whether it is +0.9 or –0.9. Another limitation of the concept of the coefficient of determination perhaps suggests that the independent variable X is in a determining position, that is, there is a causal relationship between the two variables. Instead, the statistical evidence suggests that there is covariation regardless of the fact whether it is causal or not. To understand whether the relationship between the two variables is causal, we have to collect further evidence and we cannot depend on the quantitative result alone.

17.7 RANK CORRELATION

Sometimes we come across statistical series that are ranked according to size. This is because the exact magnitude of individual items in the series cannot be ascertained. In such cases, Karl Pearsonian coefficient of correlation cannot be calculated. Instead, another method of correlation, popularly known as *Spearman's correlation*, is used. As the name implies, this method is based on the ranks (or order) of the observations rather than on a specific distribution of X and Y. The method is very handy, involving simple computations.

We may have two types of numerical problems in rank correlation:

(i) When actual ranks are given

(ii) When ranks are not given

1. In the first case, when actual ranks are given, the differences of the two ranks $(R_1 - R_2)$ are taken and these are denoted by 'd'.
2. The differences are squared and their total (Σd^2) obtained.
3. Then the following formula is applied:

$$r_s = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

where r_s denotes Spearman's correlation or rank correlation.

In the second case, when the ranks are not given, that is, when actual data are given, we have to assign ranks. We may do so by taking highest value as 1 or the lowest value as 1. One of these two approaches is to be followed. But we must follow the same approach for all the observations consistently.

When the two observations are the same, then the normal practice is to assign an average rank to the two observations. Suppose two individuals are rated equal at seventh place, they are each given the rank (7 + 8)/2, that is, 7.5. In other words, when two observations are the same, then they are assigned the average of the ranks that they would have got had both of them been slightly different from each other. Once the ranks have been assigned in the two series, the other steps will be the same as given earlier.

We shall now illustrate how the coefficient of rank correlation can be computed in both the cases. In the first example, we deal with the ranks in the two series already given to us.

Example 17.8 Suppose that 10 salesmen employed by a company were given a month's training. At the end of the specified training, they took a test and were ranked on the basis of their performance. They were then posted to their respective areas. At the end of six months, they were rated in respect of their sales performance. These ranks are shown below:

Salesmen	1	2	3	4	5	6	7	8	9	10
Ranks obtained in training	4	6	1	3	9	7	10	2	8	5
Ranks based on sales performance	5	8	3	1	7	6	9	2	10	4

Calculate the coefficient of rank correlation and comment on the result.

Solution These data are given in Table 17.9. In order to calculate rank correlation, we have to calculate Σd^2. Table 17.9 also shows these calculations.

Table 17.9 Ranks of Salesmen in Respect of Training and Sales Performance

Salesmen	*Ranks obtained in Training (X)*	*Ranks on the Basis of Sales Performance (Y)*	*Difference (X – Y = 'd')*	*Difference Squared (d^2)*
1	4	5	–1	1
2	6	8	–2	4
3	1	3	–2	4
4	3	1	2	4
5	9	7	2	4
6	7	6	1	1
7	10	9	1	1
8	2	2	0	0
9	8	10	–2	4
10	5	4	1	1
				$\Sigma d^2 = 24$

To compute the coefficient of rank correlation, the following formula is used:

$$r_s = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

$$r_s = 1 - \frac{(6)(24)}{(10)[(10^2 - 1)]} = 1 - \frac{144}{(10)(99)} = 0.855$$

A coefficient of 0.855 shows that there is a very high degree of correlation between the performance in training and the sales performance of the ten salesmen.

Now, the significance of the coefficient of rank correlation can be tested. It may be noted that for small values of n (i.e. when n is ≤ 30), the distribution of r_s is not normal. Unlike other small sample data, it is not appropriate to use the t test for testing hypotheses about the rank correlation. In such cases, we use Spearman's Rank Correlation Values as given in Appendix Table 7. We apply this test and set up the null hypothesis that there is no correlation between the ranked data. In other words, the null hypothesis is that $r_s = 0$. The alternative hypothesis is that $r_s \neq 0$. We test the hypothesis at 5 per cent level of significance. From Appendix Table 7, we find the critical value of r_s for $n = 10$ at 5 per cent level of significance as 0.6364.

As the calculated value of rank correlation r_s is 0.855, which is more than the critical value, the null hypothesis is rejected. In other words, the performance in training and the sales performance of a sample of ten salesmen are associated.

Example 17.9 Find out Spearman's coefficient of correlation between the two kinds of assessment of postgraduate students' performance in a college:

Name of students	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Internal Assessment (Int. Asmt.) (out of 100 marks)	51	63	73	46	50	60	47	36	60
External Assessment (Ex. Asmt.) (out of 100 marks)	49	72	74	44	58	66	50	30	35

Solution

Table 17.10 Calculation of Rank Correlation

Name	*Int. Asmt.*	*Ranks R_1*	*Ext. Asmt.*	*Ranks R_2*	*d $(R_1 - R_2)$*	*d^2*
A	51	5	49	6	–1	1
B	63	2	72	2	0	0
C	73	1	74	1	0	0
D	46	8	44	7	1	1
E	50	6	58	4	2	4
F	60	3.5	66	3	0.5	0.25
G	47	7	50	5	2	4
H	36	9	30	9	0	0
I	60	3.5	35	8	–4.5	20.25
						30.50

It will be seen that in case of internal assessment, two students obtained the same marks, viz. 60. As such, they should have the same rank. Had there been different marks, then ranks should have been 3 for student E and 4 for student I. Now, each of them gets a rank of (3 + 4)/2 = 3.5. In view of this, the next rank has to be 5.

$$r_s = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

$$r_s = 1 - \frac{(6)(30.5)}{(9)[(9^2-1)]} = 1 - \frac{183}{720} = 0.746 = 0.75 \text{ approx.}$$

If we refer again to Appendix Table 7, we find that that critical value of r_s for $n = 9$ and at 5 per cent level of significance is 0.6833. As our calculated value of r_s is 0.75, which is greater than the critical value of 0.6833, the null hypothesis that there is no relationship between the ranked variables is rejected. In other words, we conclude that there is a high degree of positive relationship between the internal assessment and the external assessment.

Sometimes we come across problems where one or two items are missing or wrongly entered. Later on when we come to know the actual position, we have to revise our earlier calculation. Let us take a simple problem of this type.

Example 17.10 The coefficient of rank correlation of the marks obtained by 10 students in Statistics and accountancy was found to be 0.2. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct value of the coefficient of rank correlation.

Solution The formula for Spearman's correlation is

$$r_s = 1 - \frac{6\Sigma d^2}{N(N^2-1)}$$

$$r_s = 1 - \frac{6\Sigma d^2}{10(10^2-1)} \quad \text{or} \quad 0.2 = 1 - \frac{6\Sigma d^2}{990}$$

or $\qquad 198 = 990 - 6\Sigma d^2 \qquad \therefore \quad \Sigma d^2 = 132$

As the difference in ranks between the two subjects for one student was wrongly taken as 9 instead of 7, the necessary adjustment in Σd^2 should be

$$\Sigma d^2 = 132 - 9^2 + 7^2 \quad = 132 - 81 + 49 \quad = 100$$

The correct r_s will be

$$r_s = 1 - \frac{6 \times 100}{990} = 1 - \frac{600}{990} = 0.39$$

Limitations of Spearman's Method of Correlation It may be noted that Spearman's r is a *distribution-free* or nonparametric measure of correlation. This is due to the fact that no strict assumptions are made regarding the underlying distribution from which the sample observations have been taken. As such, the result may not be as dependable as in the case of ordinary correlation where the distribution is known. This is a limitation of Spearman's correlation though it is much easier to compute than r. Another limitation of rank correlation is that it cannot be applied to a grouped frequency distribution. Further, when the number of observations is quite large, say 30 or more, where ranks are not given and one has to assign ranks to the observations in the two series, then such an exercise becomes rather tedious and time-consuming. This becomes a major limitation of rank correlation.

Additional Examples

Example 17.11 You are given the following information:

$\Sigma x^2 = 90 \quad \sigma_y = 2.5 \quad \Sigma xy = 60$ and $\quad r = 0.8$

x and y are the deviations from their respective means. Find the number of observations.

Solution The formula for correlation coefficient when deviations are used is:

$$r = \frac{\Sigma xy}{n \cdot \sigma_x \cdot \sigma_y}$$

Substituting the values given in this formula,

$$0.8 = \frac{60}{n\sqrt{\left(\frac{90}{n}\right)}(2.5)}$$

or $$(0.8)^2 = \frac{(60)^2}{n^2 \times \left(\frac{90}{n}\right) \times 6.25}$$

or $$0.64 = \frac{3600}{n \times 90 \times 6.25}$$

or $$0.64 = \frac{3600}{n \times 562.5}$$

or $(0.64)\ (562.5\ n) = 3600$

or $360\ n = 3600$

$\therefore \quad n = 3600/360 = 10$

Example 17.12 Calculate correlation coefficient between X and Y variables, using the following information:

$n = 12 \quad \Sigma X = 96 \quad \Sigma Y = 120 \quad \Sigma (X-8)^2 = 150 \quad \Sigma (Y-10)^2 = 200$ and $\quad \Sigma (X-8)\ (Y-10) = 50$

Solution

$$\bar{X} = \frac{\Sigma X}{n} = \frac{96}{12} = 8$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{120}{12} = 10$$

$\Sigma (X-8)^2$ = Summation of the squared deviations from the mean of x series = Σx^2

$\Sigma (Y-10)^2$ = Summation of squared deviations from the mean of y series = Σy^2

$\Sigma (X-8)(Y-10) = \Sigma xy$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{50}{\sqrt{150 \times 200}}$$

$$= \frac{50}{\sqrt{30000}}$$

$$= \frac{50}{173.2}$$

$$= 0.29$$

Example 17.13 A person while calculating correlation coefficient between two variables X and Y obtained the following results:

$$n = 30 \quad \Sigma X = 120 \quad \Sigma X^2 = 600 \quad \Sigma Y = 90 \quad \Sigma Y^2 = 250 \text{ and } \quad \Sigma XY = 356$$

It was, however, later discovered at the time of checking the calculations that two pairs of observations (8, 10) and (12, 7) were wrongly entered instead of (8, 12) and (10, 8). Determine the correct value of correlation coefficient.

Solution We have to make the necessary adjustments in the given values as follows:

Corrected $\Sigma X = 120 - 8 - 12 + 8 + 10 = 118$

Corrected $\Sigma X^2 = 600 - 8^2 - 12^2 + 8^2 + 10^2 = 556$

Corrected $\Sigma Y = 90 - 10 - 7 + 12 + 8 = 93$

Corrected $\Sigma Y^2 = 250 - 10^2 - 7^2 + 12^2 + 8^2 = 309$

Corrected $\Sigma XY = 356 - (8 \times 10) - (12 \times 7) + (8 \times 12) + (10 \times 8) = 368$

$$r = \frac{\Sigma XY - \dfrac{\Sigma X \cdot \Sigma Y}{n}}{\sqrt{\left[\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{n}\right]}}$$

$$= \frac{368 - \dfrac{118 \times 93}{30}}{\sqrt{\left[556 - \dfrac{(118)^2}{30}\right]\left[309 - \dfrac{(93)^2}{30}\right]}}$$

$$= \frac{368 - 365.8}{\sqrt{(556 - 464.13)(309 - 288.3)}}$$

$$= \frac{2.2}{\sqrt{91.87 \times 20.7}} = \frac{2.2}{\sqrt{1901.709}} = \frac{2.2}{43.61} = 0.05$$

Example 17.14 The data on aptitude score and productivity index of six workers in a factory are given below:

Workers	*A*	*B*	*C*	*D*	*E*	*F*
Aptitude score (X)	9	18	18	20	20	23
Productivity index (Y)	23	23	33	42	29	32

Calculate coefficient of correlation.

Solution

Worksheet

Workers	X	Y	X^2	Y^2	XY
A	9	23	81	529	207
B	18	23	324	529	414
C	18	33	324	1089	594
D	20	42	400	1764	840
E	20	29	400	841	580
F	23	32	529	1024	736
	108	182	2058	5776	3371

The formula for correlation coefficient is

$$r = \frac{\Sigma XY - \frac{\Sigma X \times \Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

$$= \frac{3371 - \frac{108 \times 182}{6}}{\sqrt{\left[2058 - \frac{(108)^2}{6}\right]\left[5776 - \frac{(182)^2}{6}\right]}}$$

$$= \frac{3371 - 3276}{\sqrt{(2058 - 1944)(5776 - 5520.67)}}$$

$$= \frac{95}{\sqrt{114 \times 255.33}}$$

$$= \frac{95}{\sqrt{29108}} = \frac{95}{170.61} = 0.56$$

Example 17.15 Calculate the coefficient of correlation between price and sales from the following data:

Price (Rs)	100	90	85	92	90	84	88	90
Sales (Units)	500	610	700	630	670	800	800	750

Interpret the value of r.

Solution

Price (Rs)			*Sales (Unit)*			
(x)	*d from 90* dx	d_x^2	*y*	*d from 700* dy	d_y^2	$dx \cdot dy$
100	10	100	500	–200	40000	–2000
90	0	0	610	–90	8100	0
85	–5	25	700	0	0	0
92	2	4	630	–70	4900	–140
90	0	0	670	–30	900	0
84	–6	36	800	100	10000	–600
88	–2	4	800	100	10000	–200
90	0	0	750	50	2500	0
719	–1	169	5460	–140	76400	–2940

We have to use the short-cut method as our calculations are based on arbitrary mean. For this, the following formula is used.

$$r = \frac{\Sigma dx dy - \dfrac{\Sigma dx \times \Sigma dy}{N}}{\sqrt{\left(\Sigma dx^2 - \dfrac{(\Sigma dx)^2}{N}\right)\left(\Sigma dy^2 - \dfrac{(\Sigma dy)^2}{N}\right)}}$$

Applying the above values in this formula, we get

$$r = \frac{-2940 - \left(\dfrac{-1 \times -140}{8}\right)}{\sqrt{\left(169 - \dfrac{(-1)^2}{8}\right)\left(76400 - \dfrac{(-140)^2}{8}\right)}}$$

$$= \frac{-2957.5}{\sqrt{(169 - 0.125)(76400 - 2450)}}$$

$$= -\frac{2957.5}{\sqrt{168.875 \times 73950}}$$

$$= -0.88$$

This shows a high degree of negative correlation.

Example 17.16 The following are the monthly figures of advertising expenditure and sales of a firm. It is generally found that advertising expenditure has an impact on sales after two months. Allowing for this time lag, calculate coefficient of correlation between expenditure on advertisment and sales.

Month	*Ad. Expenditure ('000 Rs)*	*Sales ('000 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

Solution As the problem indicates that it is generally found that advertising expenditure has an impact on sales after two months, it is necessary to adjust sales data. This means that for January advertising expenditure, sales figure for March is to be shown. This process will go on until we reach October advertising expenditure against which December sales figure is to be shown.

It may be noted that the figures are shown in thousand rupees. To simplify our calculations further, we divide each figure in Ad expenditure series by 10 and each figure in sales series by 100.

We now set up the Worksheet for calculation of correlation coefficient.

Worksheet

Month	*Ad Exp.*			*Sales*			
	x	dx	dx^2	y	dy	dy^2	
January	5	–7	49	16	–10	100	70
February	6	–6	36	20	–6	36	36
March	7	–5	25	22	–4	16	20
April	9	–3	9	25	–1	1	3
May	12	0	0	24	–2	4	0
June	15	3	9	26	0	0	0
July	14	2	4	28	2	4	4
August	16	4	16	29	3	9	12
September	17	5	25	31	5	25	25
October	19	7	49	39	13	169	91
	120		222	260		364	261

$$\bar{x} = \frac{\Sigma x}{n} = \frac{120}{10} = 12 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{260}{10} = 26$$

$$\Sigma dx^2 = 222 \qquad \Sigma dy^2 = 364$$

$$\Sigma dxdy = 261$$

$$r = \frac{\sum dxdy}{\sqrt{\sum dx^2 \times \sum dy^2}}$$

$$= \frac{261}{\sqrt{222 \times 364}}$$

$$= \frac{261}{\sqrt{80808}} = \frac{261}{284.27} = 0.92$$

Example 17.17 The director of a management training programme is interested to know whether there is a positive association between a trainee's score prior to his/her joining the programme and the same trainee's score after the completion of the training. The director has obtained the scores of 10 trainees as follows:

Trainee	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Rank Score 1	1	4	10	8	5	7	3	2	6	9
Rank Score 2	2	3	9	10	5	6	1	4	7	8

Determine the degree of association between pre-training and post-training scores.

Solution

Worksheet

Trainee No.	*Score 1* X	*Score 2* Y	*Rank diff.* $d = X - Y$	$d^2 = (X - Y)^2$
1	1	2	–1	1
2	4	3	1	1
3	10	9	1	1
4	8	10	–2	4
5	5	5	0	0
6	7	6	1	1
7	3	1	2	4
8	2	4	–2	4
9	6	7	–1	1
10	9	8	1	1
	55	55	0	18

Spearman's rank correlation

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6(18)}{10(10^2 - 1)}$$

$$= 1 - \frac{108}{990}$$

$= 1 - 0.11$

$= 0.89$

In order to determine whether this spearman's correlation coefficient is statistically significant or not, we first set up the two hypotheses.

$H_0 : r_s = 0$

$H_1 : r_s \neq 0$

From Appendix Table 7, we find that the critical value of r_s for $n = 10$ at 5 per cent level of significance for a two-tail test is 0.6364. As the calculated value of r_s is greater than the critical value, H_0 is rejected. In other words, the association between the pre-training and post-training scores of 10 trainees is statistically significant.

Example 17.18 Find the coefficient of correlation between X and Y, given that, $10Y = 8X + 66$ and $40X = 18Y + 214$ are the regression of Y on X and X on Y, respectively.

Solution

$$10Y = 8X + 66 \quad (1)$$

or $$Y = \frac{8}{10}X + \frac{66}{10}$$

or $$byx = \frac{8}{10}$$

$$40X = 18Y + 214 \quad (2)$$

or $$X = \frac{18}{40}Y + \frac{214}{40}$$

or $$bxy = \frac{18}{40}$$

$$r = \sqrt{b\,x\,y \times b\,y\,x}$$

$$= \sqrt{\frac{18}{40} \times \frac{8}{10}}$$

$$= \sqrt{\frac{144}{400}}$$

$$= \sqrt{0.36}$$

$$= 0.6$$

Example 17.19 Calculate the correlation coefficient, given that, $\Sigma x = 15$, $\Sigma y = 15$, $\Sigma x^2 = 49$, $\Sigma y^2 = 49$, $\Sigma xy = 44$, $n = 5$.

Solution

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{44}{\sqrt{49} \times \sqrt{49}}$$

$$= \frac{44}{7 \times 7}$$

$$= \frac{44}{49}$$

$$= 0.897 \quad \text{or} \quad 0.9 \text{ approx.}$$

Now, $r = \sqrt{bxy \cdot byx}$

$$bxy = \frac{\Sigma xy}{\Sigma y^2} \quad \text{and} \quad byx = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{44}{49} \quad \text{and} \quad \frac{44}{49}$$

$$r = \sqrt{\frac{44}{49} \times \frac{44}{49}}$$

$$= \frac{44}{49} = 0.897 \quad \text{or} \quad 0.9 \text{ approx.}$$

Example 17.20 A computer, while calculating correlation coefficient between the two variables, X and Y, from 25 pair of observations, gave the following values:

$$N = 25 \quad \Sigma X = 125 \quad \Sigma X^2 = 650 \quad \Sigma Y = 100 \quad \Sigma Y^2 = 460 \quad \text{and} \quad \Sigma XY = 508.$$

It was later discovered that at the time of checking, the values

X	Y
6	14
8	6

were fed, while the correct values were

X	Y
8	12
6	8

Obtain the correct value of the correlation coefficient.

Solution

To begin with, it must be noted that the values given in the question are the original values for which the following formula for calculating the correlation coefficient is used.

$$r = \frac{\Sigma XY - \dfrac{\Sigma X \cdot \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}\right)}}$$

Substituting the given values in the above formula:

$$r = \frac{508 - \frac{(125)(100)}{25}}{\sqrt{\left(650 - \frac{(125)^2}{25}\right)\left(460 - \frac{(100)^2}{25}\right)}}$$

Now, let us look at these values carefully.

It will be seen that ΣX remains unchanged (earlier it was 6 + 8 = 14, and now it is 8 + 6 = 14).

Similarly, ΣX^2 would remain the same.

Turning to the values of Y too, we find there is no change.

Earlier Y	*Now Y*	*Correction*
14	12	–2
6	8	+2
Y^2	**Y^2**	
196	144	–52
36	64	+28
232	**208**	**–24**

However, ΣY^2 needs correction. Instead of 460, it should be 460 – 24 = 436.

Now, we come to ΣXY.

Earlier $(6 \times 14) + (8 \times 6) = 84 + 48 = 132$, which was included in $\Sigma XY = 508$.

Now, $(8 \times 12) + (6 \times 8) = 96 + 48 = 144$.

Thus, 144 – 132 = 12. $\Sigma XY = 508$ should now be 508 + 12 = 520.

Substituting the correct values in the given formula, we get

$$r = \frac{520 - \frac{(125)(100)}{25}}{\sqrt{\left(650 - \frac{(125)^2}{25}\right)\left(436 - \frac{(100)^2}{25}\right)}}$$

$$= \frac{520 - 500}{\sqrt{(650 - 625)(436 - 400)}}$$

$$= \frac{20}{\sqrt{25 \times 36}}$$

$$= \frac{20}{\sqrt{900}}$$

$$= \frac{20}{30}$$

$$= 0.67$$

17.8 SOME LIMITATIONS OF CORRELATION ANALYSIS

As was mentioned earlier, correlation analysis is a statistical tool, which should be properly used so that correct results can be obtained. Sometimes, it is indiscriminately used by management, resulting in misleading conclusions. We give below some *errors* frequently made in the use of correlation analysis.

1. Correlation analysis cannot determine cause-and-effect relationship. One should not assume that a change in Y variable is caused by a change in X variable unless one is reasonably sure that one variable is the cause while the other is the effect. Let us take an example.

   Suppose that we study the performance of students in their graduate examination and their earnings after, say, three years of their graduation. We may find that these two variables are highly and positively related. At the same time, we must not forget that both the variables might have been influenced by some other factors such as quality of teachers, economic and social status of parents, effectiveness of the interviewing process and so forth. If the data on these factors are available, then it is worthwhile to use multiple correlation analysis instead of bivariate one.

2. Another mistake that occurs frequently is on account of misinterpretation of the coefficient of correlation and the coefficient of determination. Although this was explained earlier, we may briefly mention it here as well. Suppose in one case $r = 0.7$, it will be wrong to interpret that correlation explains 70 per cent of the total variation in Y. The error can be seen easily when we calculate the coefficient of determination. Here, the coefficient of determination r^2 will be 0.49. This means that only 49 per cent of the total variation in Y is explained.

   Similarly, the coefficient of determination is misinterpreted if it is also used to indicate causal relationship, that is, the percentage of the change in one variable is due to the change in another variable.

3. Another mistake in the interpretation of the coefficient of correlation occurs when one concludes a positive or negative relationship even though the two variables are actually unrelated. For example, the age of students and their score in the examination have no relation with each other. The two variables may show similar movements but there does not seem to be a common link between them.

   To sum up, one has to be extremely careful while interpreting coefficient of correlation. Before one concludes a causal relationship, one has to consider other relevant factors that might have any influence on the dependent variable or on both the variables. Such an approach will avoid many of the pitfalls in the interpretation of the coefficient of correlation. It has been rightly said that the coefficient of correlation is not only one of the most widely used, but also one of the widely abused statistical measures.

GLOSSARY

Coefficient of determination (r^2)	The square of the linear correlation coefficient. It measures the proportion of variation in Y values, which is explained by the linear relationship with X. It ranges from zero (none of the variation is explained) to 1 (all of the variations are explained).

Correlation analysis	Analysis of statistical data that is concerned with the question of whether there is a relationship between two variables.
Correlation coefficient (r)	A measure that indicates the extent of association between two variables. It varies from –1 to +1.
Cov (X, Y)	The covariance of X and Y, which is the average of the products of the deviations from the means for n pairs of X and Y series.
Curvilinear relationship	A relationship between two variables that is described by a curved line.
Direct relationship	A relationship between two variables showing that their movements are in the same direction.
Inverse relationship	A relationship between two variables showing that the two variables are moving in opposite directions.
Linear relationship	A particular type of association between two variables when the points in a scatter diagram cluster around a straight line with non-zero slope.
Pearson's correlation coefficient	A measure of correlation propounded by Karl Pearson. It is also known as the product moment correlation coefficient. It is the most widely used correlation coefficient.
Rank correlation	A method to determine correlation when the data are not available in numerical form and, as an alternative, the method of ranking is used.
Rank-correlation coefficient	A measure of the degree of association between two variables that is based on the ranks of observations instead of their numerical values.
Scatter diagram	A plot of the paired observations of X and Y that shows a broad pattern of relationship between the two variables.

LIST OF FORMULAE

1. Coefficient of correlation with original data

$$r = \frac{\Sigma XY - \frac{\Sigma X \times \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right)}}$$

$$= \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

2. Coefficient of correlation with original data when deviations are taken from means

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

3. Coefficient of correlation by direct method

$$r = \frac{\Sigma xy}{N\sigma_x \cdot \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}}$$

where r = coefficient of correlation, Σxy = sum of the product of deviations in X and Y series from their arithmetic means, σ_x= standard deviation of the series X, σ_y = standard deviation of the series Y and N = total number of observations.

4. Coefficient of correlation by the short–cut method

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx \times \Sigma dy}{N}}{\sqrt{\left(\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}\right)\left(\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}\right)}}$$

where dx and dy are the deviations of individual observations in X and Y series from their respective assumed means, Σdx and Σdy are the sums of the deviations in X and Y series, respectively.

5. Coefficient of correlation in grouped data (two-way frequency distribution)

$$r = \frac{\Sigma fdxdy - \frac{\Sigma fdx \times \Sigma fdy}{N}}{\sqrt{\left(\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}\right)\left(\Sigma fdy^2 - \frac{(\Sigma fdy)^2}{N}\right)}}$$

where fdx and fdy are the deviations in X and Y series multiplied by the corresponding frequencies.

6. A simplified form of the above formula

$$r = \frac{N\Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{[N\Sigma fdx^2 - (\Sigma fdx)^2][N\Sigma fdy^2 - (\Sigma fdy)^2]}}$$

7. t test for a correlation coefficient

$$t = \frac{r - p}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

where p is population correlation coefficient and r is sample correlation.

8. Coefficient of determination (r^2)

$$r^2 = \frac{a\Sigma Y + b\Sigma XY - n\bar{Y}^2}{\Sigma Y^2 - n\bar{Y}^2}$$

This is a short-cut formula for calculating r^2, where sample data are used.

9. Coefficient of rank correlation between two ranked variables

$$r_s = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

where r_s stands for Spearman's rank correlation, d for the difference between the ranks of the two variables and N for number of paired observations.

QUESTIONS

17.1 Given below are fifteen statements. Indicate in each case whether the statement is true or false:

(a) An r^2 value close to zero is an indication of a strong relationship between X and Y.
(b) An r^2 value measures how strong the relationship between X and Y is provided it is linear.
(c) If one variable is increasing while the other is declining, then there is an inverse correlation between the two variables.
(d) The coefficient of correlation must always be between zero and +1.
(e) Correlation indicates a causal relationship between the two variables.
(f) There is no difference between coefficient of correlation and coefficient of determination.
(g) A scatter diagram can give us a broad idea whether the two variables are related or not.
(h) If $r = 0.7$, it represents 70 per cent of the total variation in Y.
(i) A spurious correlation indicates that the two variables are related, but in reality there is no common link between the two.
(j) If two series X and Y plotted on a graph, move in opposite directions, then there is an absence of correlation.
(k) If one variable is constant in the two series X and Y, then the coefficient of correlation is zero.
(l) Correlation analysis is a method of obtaining the equation that represents relationship between two variables.
(m) As the value of r decreases from its maximum value of 1, there is a sharper decline in the coefficient of determination.
(n) The Spearman's r is a distribution-free measure of correlation.
(o) Rank correlation can be applied both to individual observations and to a grouped frequency distribution.

Multiple Choice Questions (17.2 to 17.12)

17.2 A Scatter diagram is

(a) a statistical test (b) linear
(c) curvilinear (d) a graph showing x and y values

17.3 Which of the following correlation coefficients shows the highest degree of association?

(a) 0.9 (b) 0.95 (c) –0.89
(d) –1 (e) 1 (f) Both (d) and (e)

17.4 Which of the following correlation coefficients shows the lowest degree of association?

(a) –0.95 (b) 0.75 (c) 0.38 (d) 0.1

17.5 Which of the following measurement scales is required for the valid calculation of Karl Pearson's correlation coefficient?
(a) ordinal (b) interval (c) ratio (d) nominal

17.6 The value of r^2 for a particular situation is 0.81. What is the coefficient of correlation?
(a) 0.81 (b) 0.9 (c) 0.09 (d) none of these

17.7 Which of the following is the highest range of r?
(a) 0 and 1 (b) –1 and 0 (c) –1 and 1 (d) none of these

17.8 Which of the following is most likely to be an inverse relationship?
(a) Between income and expenditure on education.
(b) Between price increase and demand for a certain product.
(c) Between average number of hours studied per day and the performance of the students in the examination
(d) Between advertising expenditure and sales of a product.

17.9 When the correlation coefficient between x and y is positive, then as variable x decreases, variable y
(a) remains the same (b) increases
(c) decreases (d) changes linearly

17.10 Which of the following measurement scales is required for the valid calculation of Spearman correlation coefficient?
(a) ordinal (b) interval (c) nominal (d) ratio

17.11 Many studies show that advertising expenditure and sales have a high degree of positive association. Which of the following correlation coefficient is consistent with the above statement?
(a) –0.8 (b) 0.4 (c) –0.3 (d) 0.75

17.12 Suppose you are told that there is a direct relationship between incentives and productivity. Which of the following statements would be correct?
(a) Productivity tends to be low when more incentives are given.
(b) Productivity tends to be more when incentives are more.
(c) Productivity and incentives move in different directions.
(d) High or low incentives hardly make any impact on productivity.

17.13 What are the advantages and limitations of Karl Pearson's coefficient of correlation?

17.14 "A high degree of positive correlation does not necessarily mean that the two variables in question are related." Comment.

17.15 Differentiate between the coefficient of determination and the coefficient of correlation.

17.16 What mistakes frequently occur while interpreting the coefficient of correlation?

17.17 Show that if one of the variables in a bivariate series is constant, then their coefficient of correlation will be zero.

17.18 What is a scatter diagram? How does it help in studying the correlation between the variables in respect of both of its direction and degree?

17.19 How would you interpret the value of a coefficient of correlation?

17.20 You are called upon to calculate the coefficient of correlation of bivariate grouped data. Explain, in proper sequence, the steps involved in such an exercise.

17.21 What is covariance? How does it relate to the coefficient of correlation? What is the range within which it must always lie and why?

17.22 What is meant by correlation? Do you think that correlation always signifies a cause-and-effect relationship between the two variables?

17.23 What do you understand by a linear correlation coefficient? Within what range can a correlation coefficient assume a value?

17.24 What are the different methods of finding correlation between the two variables?

17.25 Given below are some concepts. Explain these, using graph to illustrate each concept:

(a) Perfect positive linear correlation
(b) Non-linear correlation
(c) Perfect negative correlation
(d) Spurious correlation

17.26 Indicate in each of the following cases whether the coefficient of correlation is likely to be positive, negative or spurious:

(a) Height and weight of students in a class
(b) Advertising expenditure and sales of a product
(c) The IQs of husbands and wives
(d) Weights of students and shoe sizes
(e) Price of tea per kg and the demand for it
(f) Dividends and profits of a company over a period of 10 years
(g) Amount of rainfall and agricultural output

17.27 Write the applicable assumptions of Pearsonian coefficient of correlation. Also, describe important properties of the coefficient of correlation.

17.28 Explain how you would use the t test to test the hypothesis that r is significant.

17.29 What is the rank correlation? Which formula is normally used for calculating it?

17.30 Why is rank correlation important in Business Statistics? How does it differ from Karl Pearson's coefficient of correlation?

17.31 What is coefficient of rank correlation? Bring out its usefulness as also its limitations.

17.32 Calculate the coefficient of correlation between X and Y series from the following data:

	X series	*Y series*
No. of observations	15	15
Mean	25	18
Standard deviation	3.01	3.03
Square of deviations from their means	136	138
Summation of product deviations of X and Y series from their means		122

17.33 Calculate coefficient of correlation between X and Y from the following data:

Marks in English *(X)*	2	5	4	6	9
Marks in Mathematics (*Y*)	3	4	4	8	9

17.34 In Problem 17.33, suppose 20 is added to each item in X series and 30 to each item in Y series, then calculate coefficient of correlation. Do you find any difference in the answer as compared to the previous one?

17.35 A functional relationship exists between the two variables given below:

X	6	4	2	0	2	4	6
Y	–3	–2	–1	0	1	2	3

Explain why the coefficient of correlation is zero.

17.36 Calculate the coefficient of correlation between *X* and *Y* from the data given below:

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

17.37 Find the coefficient of correlation between the age and the sum assured from the following table:

	Sum Assured in Rs '000				
Age Group (years)	*10*	*20*	*30*	*40*	*50*
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	—	—

17.38 Calculate Karl Pearson's coefficient of correlation from the following data:

	X Series			
Y series	*10–13*	*6–9*	*2–5*	*Total*
20–24	2	1	—	3
15–19	1	3	—	4
10–14	—	2	2	4
5–9	—	1	1	2
0–4	—	1	1	2
Total	**3**	**8**	**4**	**15**

17.39 Calculate Karl Pearson's coefficient of correlation between *X* and *Y* from the bivariate sample of 140 pairs of *X* and *Y* as distributed below:

Y \ *X*	*10–20*	*20–30*	*30–40*	*40–50*
10–20	20	26	—	—
20–30	8	14	37	—
30–40	—	4	18	3
40–50	—	—	4	6

17.40 Calculate Karl Pearson's coefficient of correlation from the data given below:

Marks	*Age in Years*				
	18	*19*	*20*	*21*	*22*
20–25	3	2	—	—	—
15–20	—	5	4	—	—
10–15	—	—	7	10	—
5–10	—	—	—	3	2
0– 5	—	—	—	3	1

17.41 The following frequency distribution relates to the age and salary of 100 employees working in an organisation. Find the coefficient of correlation.

Age (Yrs)	*Salary (Rs)*			
	1300–1400	*1400–1500*	*1500–1600*	*1600–1700*
20–30	4	6	5	2
30–40	2	5	8	5
40–50	8	12	20	2
50–60	0	8	12	1

17.42 Calculate the coefficient of correlation between age and sum assured from the data given below and comment on the value.

	Sum assured in Rs lakh				
Age group	*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**

17.43 Compute the coefficient of correlation between the security prices and the amounts of annual dividend on the basis of the following data:

Annual dividend (Rs) / *Security prices (Rs)*	*4–8*	*8–12*	*12–16*	*16–20*	*Total*
120–140	–	–	2	4	6
100–120	–	1	2	3	6
80–100	–	2	3	–	5
60– 80	2	2	2	–	6
40– 60	4	2	1	–	7
Total	**6**	**7**	**10**	**7**	**30**

17.44 Using the following data on weight (Y lbs) and height (X inches) of children, calculate the correlation coefficient.

Y lbs	*X Inches*				*Total*
	47	*48*	*49*	*50*	
41	2	1			3
42	1	2	2		5
43		2	3	1	6
44		1	4	2	7
Total	**3**	**6**	**9**	**3**	**21**

17.45 Write the expression for coefficient of correlation in terms of covariance and variance of the variables involved. Prove that if X and Y are independent variables then they are linearly uncorrelated. Is the converse true?

17.46 Two persons were asked to watch ten specified TV programmes and offer their evaluation by rating them 1 to 10. These ratings are given below:

TV Programme	*Ranks Given by*	
	X	*Y*
A	4	2
B	6	3
C	3	4
D	9	9
E	1	5
F	5	7
G	2	1
H	7	10
I	10	8
J	8	6

Calculate Spearman's coefficient of correlation of the two ratings.

17.47 Table below shows data on the cost of advertisement and circulation (X) and return–on–inquiry cost (Y) for a sample of 14 magazines. Can it be inferred that the costs on advertisement and return–on–inquiry cost are correlated? Test at 5 per cent level of significance, using the Spearman's rank correlation coefficient.

	X	*Y*
1	4.1	17.4
2	5.6	36.0
3	3.8	29.7
4	2.5	22.2
5	1.3	90.8
6	3.3	92.0
7	1.3	65.8
8	1.5	78.6
9	2.7	98.9
10	1.6	21.9
11	1.6	28.7
12	1.8	48.0
13	3.0	59.0
14	1.7	32.8

17.48 A study was conducted to find out whether there is correlation between consumers' perceptions of a TV commercial and their interest in purchasing a specific product (both variables were measured on a scale). The results are $n = 65$ and $r = 0.37$. Do you think that there is a linear relationship between the two variables?

17.49 Compute the correlation coefficient between the corresponding values X and Y from the given data:

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

17.50 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

17.51 From the following data, obtain the two regression equations and calculate the correlation coefficient.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

17.52 Find Karl Pearson's coefficient of correlation between X and Y for the following data:

X	3	4	8	9	6	2	1
Y	5	3	7	7	6	9	2

17.53 A researcher wants to explain attitude towards respondents' city of residence, in terms of duration of residence in the city. The attitude is measured on an 11-point scale. (1 = do not like the city, 11-very much like the city), and the duration of residence is measured in terms of the number of years the respondent has lived in the city. In a pre-test of 12 respondents, the data, as shown in the table, are obtained. Find the correlation between attitude and duration.

Respondent	*Attitude towards the City*	*Duration of Residence*
1	6	10
2	9	12
3	8	12
4	3	4
5	10	12
6	4	6
7	5	8
8	2	2
9	11	18
10	9	9
11	10	17
12	2	2

17.54 The following table gives the frequency of marks obtained by 67 students in an intelligence test. Measure the degree of relationship between age and marks.

	Age in years				
Test marks	*18*	*19*	*20*	*21*	*Total*
200–250	4	4	2	1	11
250–300	3	5	4	2	14
300–350	2	6	8	5	21
350–400	1	4	6	10	21
Total	10	19	20	18	67

17.55 Eleven cities are ranked according to their pollution levels and occurrence of pulmonary diseases.

City	:	A	B	C	D	E	F	G	H	I	J	K
Pollution	:	4	7	9	1	2	10	3	5	6	8	11
Pulmonary diseases	:	5	4	7	3	1	11	2	10	8	6	9

Is there any relation between pollution and occurrence of pulmonary diseases? Comment on the findings.

17.56 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 to determine which pair of judges has the nearest approach to common taste in beauty.

17.57 The following are the ratings of aggressiveness (X) and amount of sales (Y) during last year for eight sales people. Is there a significant rank correlation between the two measures ? Use the 0.10 significant level.

X	30	17	35	28	42	25	19	28
Y	35	31	43	46	50	32	33	42

17.58 An article in Concrete Research presented data on compressive strength, x, and intrinsic permeability, y, of various concrete mixes and cures. Summary quantities are $n = 14$, $\Sigma y_i = 572$, $\Sigma y_i^2 = 23530$, $\Sigma x_i = 43$, $\Sigma x_i^2 = 157.42$ and $\Sigma x_i y_i = 1697.80$. Assume that the two variables are related according to the simple linear regression model. Calculate the least square estimates of the slope and intercept.

17.59 M/s Standard Engineering Company manufactures various equipment required for the chemical industry.

Since it undertakes turnkey projects, the manufactured items are not standardized. The Directors of the company are working on a tender for an export order, and are looking at various cost factors. The components require 4500 hours of machining. Mr Joshi, Manager of the production unit, has been asked to submit a report on the production costs. He has estimated relevant direct costs but is facing problems while estimating indirect costs. He seeks your help on the basis of the following data, which relate to X, machine hours ('00), and Y, indirect costs (Rs '000). Assess whether there exists a linear relation in X and Y, and, if yes, to what extent? Also estimate the required indirect costs when X is 5000, machine hours.

X (Hours '00)	40	24	8	40	32	24	16	48	32	16
Y (Rs '000)	96	88	48	110	80	64	56	120	88	54

The McGraw-Hill Companies

CHAPTER 18

MULTIPLE REGRESSION AND CORRELATION ANALYSIS

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the meaning and significance of multiple linear regression
- calculate the multiple linear regression equation and interpret the result
- differentiate between partial and multiple correlation and calculate them and interpret the result
- know the major advantages and limitations of multiple correlation analysis.

Chapter Prerequisites

Before starting work on this chapter, make sure you have fully understood the topics covered in Chapters 16 and 17.

18.1 INTRODUCTION

In the preceding two chapters, our discussion on regression and correlation analysis was confined to only two variables. However, in real life, we come across several situations where the relationship is not that simple. One variable may be affected by two or more independent variables. For example, sale of a product, Y, may be related to a number of independent variables such as price, income, advertising expenditure, seasons, number, size and location of retail outlets, quality of the product and so forth. If in such cases, we take cognizance of only one independent variable, then the magnitude of the error in the result is likely to be high. In view of this, it is desirable to use two or more independent variables in the estimating equation. The statistical technique of extending linear regression so as to consider two or more independent variables is known as *multiple linear regression*. In this chapter, we shall study multiple linear regression and correlation analysis.

Before discussing multiple regression, let us summarise the results of simple linear regression discussed in Chapter 16. It may be recalled that the method of least squares was used to obtain an estimating equation—$Y = a + bX$, where values of a and b are obtained by the method of least squares. The normal equations used are:

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

These two normal equations are used when the problem relates to two variables only.

18.2 MULTIPLE REGRESSION

In case of multiple regression, many formulae can be used to ascertain the relationships among variables. The most frequently used method among social scientists is that of linear equations.

The multiple linear regression takes the following form:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where Y is the dependent variable, which is to be predicted; X_1, X_2, X_3, … X_k are the K known variables on which the predictions are to be based and a, b_1, b_2, b_3, … b_k are parameters, the values of which are to be determined by the method of least squares.

An example will make the method in respect of multiple regression clear.

Example 18.1 The following data relate to radio advertising expenditures, newspaper advertising expenditures and sales. Fit a regression $Y = a + b_1X_1 + b_2X_2$.

Radio ad. exp. ('000 Rs) (X_1)	4	7	9	12
Newspaper ad. exp. ('000 Rs) (X_2)	1	2	5	8
Sales (Rs lakh) (Y)	7	12	17	20

Solution It may be noted here that as there are three variables, viz. Y, X_1 and X_2, there will be three normal equations as shown below:

$$\Sigma Y = na + b_1\Sigma X_1 + b_2\Sigma X_2 \quad (1)$$
$$\Sigma X_1Y = a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \quad (2)$$
$$\Sigma X_2Y = a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 \quad (3)$$

In order to use these normal equations, it is necessary to find the numerical values of the terms involved in the equations. For this purpose, we have to set up the worksheet as was done in Chapter 16. Table 18.1 gives the worksheet.

Table 18.1 Worksheet for Calculation of Regression Coefficients

X_1	X_2	Y	X_1^2	X_1X_2	X_2^2	X_1Y	X_2Y
4	1	7	16	4	1	28	7
7	2	12	49	14	4	84	24
9	5	17	81	45	25	153	85
12	8	20	144	96	64	240	160
32	**16**	**56**	**290**	**159**	**94**	**505**	**276**

Applying the above values in the normal equations,

$$56 = 4a + 32b_1 + 16b_2 \quad (1)$$
$$505 = 32a + 290b_1 + 159b_2 \quad (2)$$
$$276 = 16a + 159b_1 + 94b_2 \quad (3)$$

Multiplying (1) by 8 and subtracting (2) from (4),

$$448 = 32a + 256b_1 + 128b_2 \quad (4)$$
$$505 = 32a + 290b_1 + 159b_2 \quad (2)$$
$$-57 = -34b_1 - 31b_2 \quad (5)$$

Multiplying (3) by 2 and subtracting (2) from (6),

$$552 = 32a + 318b_1 + 188b_2 \quad (6)$$
$$505 = 32a + 290b_1 + 159b_2 \quad (2)$$
$$47 = 28b_1 + 29b_2 \quad (7)$$

Multiplying (5) by 14 and (7) by 17 and subtracting (9) from (8)

$$798 = 476b_1 + 434b_2 \quad (8)$$
$$799 = 476b_1 + 493b_2 \quad (9)$$
$$-1 = -59b_2$$

$\therefore \quad b_2 = 1/59 = 0.0169$

Substituting the value of $b_2 = 1/59$ in (5) above,

$$57 = 34b_1 + (31 \times 1/59)$$

or $\quad 34b_1 = 57 - 0.525$

or $\quad b_1 = 1.661$

Substituting the value of $b_1 = 1.661$ and $b_2 = 0.0169$ in (1) above,

$$56 = 4a + (1.661 \times 32) + (0.0169 \times 16)$$

or $\quad 4a = 56 - 53.152 - 0.2704$

or $\quad 4a = 2.5776$

$\therefore \quad a = 2.5776/4 = 0.6444$

Therefore, the multiple regression of Y on X_1 and X_2 is

$$Y = 0.6444 + 1.661X_1 + 0.0169X_2$$

The values of coefficients b_1 and b_2 give us a clear indication as to which variable X_1 or X_2 is more important. In our example, we find that the value of b_1 is 1.661 and that of b_2 is 0.0169. This means that one unit of change in X_1 leads to 1.661 units of change in the dependent variable Y, while one unit of change in X_2 leads to only 0.0169 unit of change in the dependent variable Y. Thus, it becomes clear that radio advertising expenditure (X_1) is more important than newspaper advertising expenditure (X_2).

The above example related to two independent variables. At times, we may have problems where more than two independent variables are involved. The main problem of obtaining a linear equation in more than two variables which best describes a given set of data is that of finding numerical values for $b_1, b_2, b_3, \ldots$ and b_k. This is usually done by the method of least squares. In other words, we minimise the sum of squares $\Sigma(Y - \hat{Y})$, where as before, the Y's are the observed values and the $\hat{Y}$*'s* are the values calculated by means of linear estimating equation. It may be noted that the problem of finding the values of $b_1, b_2, b_3, \ldots$ and b_k is the same as was done earlier in Chapter 16. However, the calculation of these values becomes more tedious because the method of least squares yields as many normal equations as there are unknown constants $b_1, b_2, b_3, \ldots$ and b_k.

Advantage of Computers in Multiple Regression To overcome the problem of tedious calculations involved in multiple regression analysis, computers are generally used. The use of computers facilitates us enormously as several independent variables can be handled. As a result, we can have a better estimating equation. We can ascertain whether adding another independent variable will

improve our results or not. We can see the magnitude of r^2, which indicates what proportion of the variation in the dependent variable is explained by the independent variables.

Example 18.2 We take another example of multiple regression analysis. This also relates to sales and advertising. However, in this example we have taken personal selling (number of selling agents) as one independent variable and overall advertising as another independent variable.

Table 18.2 Original Data

Sales Territory	*Sales (Lakh Rs) (Y)*	*Advertising ('000 Rs) (X_1)*	*Number of Selling Agents (X_2)*
1	100	40	10
2	80	30	10
3	60	20	7
4	120	50	15
5	150	60	20
6	90	40	12
7	70	20	8
8	130	60	14

Solution We set up the worksheet in Table 18.3.

Table 18.3 Worksheet for Calculation of Multiple Regression Coefficients

X_1	X_1^2	X_2	X_2^2	X_1X_2	Y	X_1Y	X_2Y
40	1,600	10	100	400	100	4,000	1,000
30	900	10	100	300	80	2,400	800
20	400	7	49	140	60	1,200	420
50	2,500	15	225	750	120	6,000	1,800
60	3,600	20	400	1,200	150	9,000	3,000
40	1,600	12	144	480	90	3,600	1,080
20	400	8	64	160	70	1,400	560
60	3,600	14	196	840	130	7,800	1,820
320	**14,600**	**96**	**1,278**	**4,270**	**800**	**35,400**	**10,480**

The totals of columns contained in Table 18.3 can be written with notations as follows:

$$\Sigma X_1 = 320 \qquad \Sigma X_1^2 = 14{,}600$$
$$\Sigma X_2 = 96 \qquad \Sigma X_2^2 = 1{,}278$$
$$\Sigma X_1X_2 = 4{,}270 \qquad \Sigma Y = 800$$
$$\Sigma X_1Y = 35{,}400 \qquad \Sigma X_2Y = 10{,}480$$

Substituting these values in the normal equations as given in the preceding example,

$$800 = 8a + 320b_1 + 96b_2 \quad (1)$$
$$35{,}400 = 320a + 14{,}600b_1 + 4{,}270b_2 \quad (2)$$
$$10{,}480 = 96a + 4{,}270b_1 + 1{,}278b_2 \quad (3)$$

Multiplying (1) by 40 and then subtracting (2) from (4),

$$32{,}000 = 320a + 12{,}800b_1 + 3{,}840b_2 \quad (4)$$
$$35{,}400 = 320a + 14{,}600b_1 + 4{,}270b_2 \quad (2)$$
$$-3{,}400 = -1{,}800b_1 - 430b_2$$
or $$340 = 180b_1 + 43b_2 \quad (5)$$

Multiplying (1) by 12 and then subtracting (3) from (6),

$$9{,}600 = 96a + 3{,}840b_1 + 1{,}152b_2 \quad (6)$$
$$10{,}480 = 96a + 4{,}270b_1 + 1{,}278b_2 \quad (3)$$
$$-880 = -430b_1 - 126b_2$$
or $$440 = 215b_1 + 63b_2 \quad (7)$$

Multiplying (5) by 43 and (7) by 36,

$$14{,}620 = 7{,}740b_1 + 1{,}849b_2 \quad (8)$$
$$15{,}840 = 7{,}740b_1 + 2{,}268b_2 \quad (9)$$
$$-1{,}220 = -419b_2 \text{ [by subtracting (9) from (8)]}$$
$$\therefore \quad b_2 = 1{,}220/419 = 2.9116945$$

Substituting the value of

$b_2 = 2.9116945$ in (7) above,

$$440 = 215b_1 + (2.9116945 \times 63)$$
or $$215b_1 = 440 - 183.43675 \text{ or } b_1 = 256.56325/215 = 1.1933174$$

Substituting the value of $b_1 = 1.1933174$ and $b_2 = 2.9116945$ in (1) above,

$$800 = 8a + (1.1933174 \times 320) + (2.9116945 \times 96)$$
or $$800 = 8a + 381.86156 + 279.52267$$
or $$8a = 800 - 661.38423$$
or $$a = 138.61577/8 = 17.326971$$

$\therefore$ The regression equation is $Y = 17.327 + 1.193X_1 + 2.912X_2$ (rounding the values to 3 decimal places).

Interpretation of the Regression Equation

The regression equation in the above example was

$$Y = 17.327 + 1.193X_1 + 2.912X_2$$

The b's are called partial regression coefficients and indicate the average change in Y for a unit change in X, holding the other X's constant. In the above equation, for example, $b_1 = 1.193$ shows that sales increase by 1.193 units for every one thousand rupees of expenditure on advertising; $b_2 = 2.912$ shows that sales increase by 2.912 units for every one person employed as a selling agent. Of the two variables, personal selling is far more important than advertising for increasing sales.

It may be noted that all the above calculations were carried out with absolute values. An alternative method based on deviations from the mean can be used in deriving the regression equation.

The coefficient of multiple determination (R^2) for this multiple linear regression can be calculated by the following formula:

$$R^2 = \frac{\Sigma(Y_i - \bar{Y})^2 - \Sigma(Y_i - \hat{Y})^2}{(Y_i - \bar{Y})^2}$$

where R^2 is coefficient of determination
Y_i = value of i^{th} item in Y series
$\bar{Y}$ = mean of the Y series
$\hat{Y}$ = computed value of i^{th} item in Y series on the basis of the regression

It has the same meaning and interpretation here as it has in the case of simple regression. The coefficient of determination is the ratio of the explained variation to the total variation. In the above formula, the term $\Sigma (Y_i - \bar{Y})^2$ shows total variation and the second term $\Sigma (Y_i - \hat{Y})^2$ shows the explained variation. The above formula is now applied to the foregoing example. The worksheet is given in Table 18.4.

Table 18.4 Worksheet for Computing Coefficient of Determination

X_1 ('000 Rs)	X_2 No of S. Agents	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$
40	10	100	94	6	36	0	0
30	10	80	82	–2	4	–20	400
20	7	60	62	–2	4	–40	1,600
50	15	120	121	–1	1	20	400
60	20	150	147	3	9	50	2,500
40	12	90	100	–10	100	–10	100
20	8	70	64	6	36	–30	900
60	14	130	130	0	0	30	900
		800			**190**		**6,800**

$$\bar{Y} = 800/8 = 100$$

$$R^2 = \frac{\Sigma(Y_i - \bar{Y})^2 - \Sigma(Y_i - \hat{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

$$= \frac{6,800 - 190}{6,800} = 0.972$$

The value of $R^2 = 0.972$ shows that 97.2 per cent of the total variation observed in the sales is explained by the regression equation. In other words, merely 2.8 per cent of the total variation in the dependent variable, Y, remains unexplained by the regression equation.

This shows that the multiple regression is able to explain the variation in the dependent variable in a convincing manner by using appropriate variables.

18.3 THE STANDARD ERROR OF ESTIMATE

It may be recalled that in Chapter 16, while dealing with the bivariate distribution, the standard error of estimate was calculated after coefficients of regression were worked out. To measure the dispersion around the multi-regression plane, the following formula is used:

$$S_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - k - 1}}$$

where S_e = standard error of estimate

$\hat{Y}$ = corresponding estimated value of Y from the regression equation

n = number of observations

k = number of independent variables

The denominator in the above formula $n - k - 1$ shows that the standard error has $n - k - 1$ degrees of freedom. This shows that the degrees of freedom are reduced by the number of variables (k) + 1 that have been estimated from the same sample.

In our preceding exercise, we set up a worksheet (Table 18.4), where we have calculated $Y - \hat{Y}$ for each individual observation in the Y series. The next column in the worksheet calculates $(Y - \hat{Y})^2$ for each individual item. The total of this item is 190. We can now apply the formula

$$S_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - k - 1}} = \sqrt{\frac{190}{8 - 2 - 1}} \quad k = 2 \text{ as } X_1 \text{ and } X_2 \text{ are the two independent variables}$$

$$= \sqrt{\frac{190}{5}} = \sqrt{38} = 6.164$$

As was done in the case of simple linear regression, we can use the standard error of estimate and the t distribution to calculate an *approximate confidence interval* around our estimated value $\hat{Y}$. The regression equation for this problem was

$$\hat{Y} = 17.327 + 1.193X_1 + 2.912X_2$$

Suppose we are interested to know what will our sales be if X_1 = Rs 70,000 and X_2 = 15. We apply these values in the above equation.

$$\hat{Y} = 17.327 + (1.193 \times 70) + (2.912 \times 15)$$
$$= 17.327 + 83.510 + 43.680$$
$$= 144.517$$

Suppose we want to be 95 per cent confident that the actual sales of the product should be within ±1 standard error of estimate from $\hat{Y}$. Further, as our n = 8, we have $n - k - 1$ degrees of freedom, that is, 8 – 2 – 1 = 5. The critical value of t for 5 degrees of freedom at 95 per cent level of significance is 2.571.

$$\hat{Y} + t\,(S_e) = \text{Rs } 144.517 + (2.571 \times 6.164) = 160.036 \text{ or } 160$$
$$\hat{Y} - t\,(S_e) = \text{Rs } 144.517 - (2.571 \times 6.164) = 128.669 \text{ or } 129$$

Since sales figures are given in lakh of rupees, these values are Rs 129 lakh and Rs 160 lakh. We can be 95 per cent confident that the sales of the product will be between Rs 129 lakh and Rs 160 lakh.

18.4 TESTING THE SIGNIFICANCE OF MULTIPLE REGRESSION

In order to test the overall significance of multiple regression, we perform a test of hypothesis using the F test. We want to know whether the value of R^2 really shows that the independent variables explain

the dependent variable Y, or it might have happened by chance. In other words, we are trying to find whether the regression as a whole is significant. To ascertain this, our two hypotheses would be

$H_0 : B_1 = B_2 = \ldots B_k = 0$ (This means that Y does not depend on X's.)

H_1 : At least one $B_i \neq 0$ (This means that Y depends on at least one of X's.)

In order to examine the null hypothesis, we have to look at the following three terms:

SST = Total sum of squares (i.e. the explained part) = $\Sigma (Y - \overline{Y})^2$

SSR = Regression sum of squares (i.e. the explained part) = $\Sigma (\hat{Y} - \overline{Y})^2$

SSE = Error sum of squares (i.e. the unexplained part) = $\Sigma (Y - \hat{Y})^2$

These three terms are related by the equation

$$\text{SST} = \text{SSR} + \text{SSE}$$

This equation shows that the total variation in Y can be split into two parts, viz. the explained part and the unexplained part.

Another point to note relates to the degrees of freedom. Each of these three terms has some degrees of freedom. SST has $n - 1$ degrees of freedom. SSR has k degrees of freedom, as there are k independent variables that are used to explain Y. Finally, SSE has $n - k - 1$ degrees of freedom on account of our having used n observations to estimate $k + 1$ constants, $a, b_1, b_2, \ldots b_k$. On the basis of these three terms and the respective degrees of freedom, we can obtain F ratio as follows:

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)}$$

If this F ratio is larger than the critical value of F at a given level of significance (obtained from the Appendix Table 6), the null hypothesis is to be rejected. The rejection of null hypothesis implies that the regression as a whole is significant.

Example 18.3 Having explained our approach to verify the null hypothesis, let us take an example to illustrate its application. We use the data given in Table 18.4, which also contains the original data earlier given in Table 18.2.

$\text{SST} = \Sigma (Y - \overline{Y})^2 = 6{,}800$

$\text{SSR} = \Sigma (\hat{Y} - \overline{Y})^2 = 6{,}610$

$\text{SSE} = \Sigma (Y - \hat{Y})^2 = 190$

We have to calculate F-statistic to test the significance of regression.

Solution

Table 18.5 Calculation of Test Statistic

Source	*Sum of Squares (SS)*	*Degrees of Freedom (DF)*	*Mean Sum of Squares (MS)*
SSR	6,610	2	3,305
SSE	190	5	38
Total	**6,800**	**7**	**971.4**

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} \quad \text{or} \quad \frac{\text{MSR}}{\text{MSE}}$$

$$F = \frac{6,610/2}{190/5} = \frac{3,305}{38} = 86.97$$

For 2 numerator degrees of freedom and 5 denominator degrees of freedom, the Appendix Table 6 shows the values of $F = 13.27$. This is the upper limit of the acceptance region for a significance level of $\alpha = 0.01$. Since our calculated F value of 86.97 is much higher than 13.27, the null hypothesis is rejected. This means that the multiple regression $Y = 17.327 + 1.193X_1 + 2.912X_2$ (obtained earlier) is highly significant.

18.5 PARTIAL AND MULTIPLE CORRELATION

The discussion so far was confined to multiple regression. We now discuss partial and multiple correlation.

The partial correlation shows the relationship between two variables, excluding the effect of other variables. In a way, the partial correlation is a special case of multiple correlation. It may be noted that there is a difference between a simple correlation and a partial correlation. The simple correlation does not include the effect of other variables as they are completely ignored. There is almost an implicit assumption that the variables not considered do not have any impact on the dependent variable. But such is not the case in the partial correlation, where the impact of other independent variables is held constant.

Let us take our previous example where three variables are involved, viz. sales, advertising expenditure and personal selling, that is, number of selling agents. Obviously, here sales is the dependent variable and the other two are independent variables. Let us give notations to these variables:

Say X_1 = Sales
X_2 = Advertising expenditure
X_3 = Number of selling agents
r = Coefficient of correlation

Here, $r_{12.3}$ shows the relation between sales (X_1) and advertising expenditure (X_2), excluding the effect of number of selling agents (X_3). In the same manner, $r_{13.2}$ shows the coefficient of correlation between sales (X_1) and number of selling agents (X_3), excluding the effect of advertising expenditure (X_2). The third partial correlation will be $r_{23.1}$, showing the coefficient of correlation between advertising expenditure (X_2) and number of selling agents (X_3), excluding the effect of sales (X_1). It should be obvious that this relationship would not be proper as sales (X_1) cannot be taken as an independent variable that would have an impact on the other two variables.

As regards notations, it should be noted that the subscripts on the left of dot (.) indicate the variables related, while the subscript on the right of dot indicates the variable excluded.

Partial Correlation Coefficient

If we denote $r_{12.3}$ as the coefficient of partial correlation between X_1 and X_2, holding X_3 constant, then

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Likewise,

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

shows $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3, holding X_2 constant.

Similarly,

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

where $r_{23.1}$ is the coefficient of partial correlation between X_2 and X_3, holding X_1 constant. Thus, we find that when three variables X_1 X_2 and X_3 are given, there will be three coefficients of partial correlation. Each of these coefficients of partial correlation will give the relationship between two variables while the third is held constant. It may also be noted that the squares of coefficients of partial correlation are called coefficients of partial determination. For example, if $r = 12.3$ is 0.7 then $r_{12.3}^2$ is 0.49. This means that 49 per cent of the variation is explained by this relationship when the third variable X_3 is held constant.

Partial correlation coefficient is helpful in deciding whether an additional variable is to be included or not. On the basis of the number of independent variables held constant, we distinguish between varying orders of coefficients of correlation. When only two variables X and Y are involved, there is no independent variable held constant; as such this is called as *zero-order correlation coefficient*. When one independent variable is held constant, then it is called *first-order correlation coefficient*. Likewise, when two independent variables are held constant, then it is known as the *second-order correlation coefficient*, and so forth.

It may also be noted that a coefficient of a given order can be expressed in terms of the next lower order. For example, a problem of first-order partial coefficient of correlation (involving three variables) can be expressed in terms of zero-order correlation, that is, simple correlation. This enables us to simplify the computations involved in case of three or more independent variables.

We shall now take a few examples to show how the coefficient of partial correlation can be calculated.

Example 18.4 In a problem involving three variables X_1, X_2 and X_3, we find that

$r_{12} = 0.8$, $r_{13} = 0.6$, $r_{23} = 0.5$

Find the values of coefficients of partial correlation $r_{23.1}$ and $r_{13.2}$.

Solution

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

$$= \frac{0.5 - (0.8 \times 0.6)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.6)^2}}$$

$$= \frac{0.5 - 0.48}{\sqrt{1-0.64}\sqrt{1-0.36}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

$$= \frac{0.6 - (0.8 \times 0.5)}{\sqrt{1-(0.8)^2}\sqrt{1-(0.5)^2}}$$

$$= \frac{0.6 - 0.4}{\sqrt{1-0.64}\sqrt{1-0.25}}$$

$$= \frac{0.02}{\sqrt{0.36}\sqrt{0.64}} \qquad = \frac{0.2}{\sqrt{0.36}\sqrt{0.75}}$$

$$= \frac{0.02}{0.6 \times 0.8} = 0.04 \qquad = \frac{0.2}{0.6 \times 0.87} = 0.38$$

Example 18.5 In a trivariate distribution, the simple coefficients of correlation are as follows:

If $r_{12} = 0.86$ $r_{13} = 0.65$ and $r_{23} = 0.72$, calculate the coefficient of partial correlation $r_{12.3}$.

Solution

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$$

$$= \frac{0.86 - (0.65)(0.72)}{\sqrt{1 - (0.65)^2}\ \sqrt{1 - (0.72)^2}}$$

$$= \frac{0.86 - 0.468}{\sqrt{1 - 0.4225}\ \sqrt{1 - 0.5184}}$$

$$= \frac{0.392}{\sqrt{0.5775}\ \sqrt{0.4816}} = \frac{0.392}{0.7599 \times 0.6939}$$

$$= \frac{0.392}{0.527} = 0.744$$

18.6 MULTIPLE CORRELATION

Unlike the partial correlation, multiple correlation is based on three or more variables without excluding the effect of anyone. It is denoted by R as against r, which is used to denote simple bivariate correlation coefficient. The subscripts are used in the same manner as in the case of partial correlation.

In case of three variables X_1, X_2 and X_3, the multiple correlation coefficients will be:

$R_{1.23}$ = Multiple correlation coefficient with X_1 as a dependent variable while X_2 and X_3 as independent variables.

$R_{2.13}$ = Multiple correlation coefficient with X_2 as a dependent variable while X_1 and X_3 as independent variables.

$R_{3.12}$ = Multiple correlation coefficient with X_3 as a dependent variable while X_1 and X_2 as independent variables.

It may be recalled that the concepts of dependent and independent variables were non-existent in case of simple bivariate correlation. In contrast, the concepts of dependent and independent variables are introduced here in multiple correlation.

Symbolically, the multiple correlation coefficient can be shown as follows:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}}$$

As is the case with simple bivariate correlation, the coefficient of multiple correlation lies between 0 and 1. As R becomes closer to 0, it shows the relationship is becoming more and more negligible. In contrast, as it moves closer to 1, it shows that the relationship is becoming more and more strong. If R is 1, the correlation is called perfect. It may be added that when R is 0 showing the absence of a linear relationship, it is just possible that there may be a non-linear relationship among the variables. Another point to note is that multiple coefficient of correlation is always positive. This is in contrast to simple bivariate coefficient of correlation, which may vary from – 1 to +1.

We can obtain the coefficient of multiple determination by squaring the multiple coefficient of correlation $R_{1.23}$.

Let us take an example.

Example 18.6 Given the following zero-order coefficients of correlation, calculate multiple coefficient of correlation taking first variable as dependent and the other two variables as independent:

$r_{12} = 0.56 \quad r_{13} = 0.38$ and $r_{23} = 0.69$

Solution

As we are required to find $R_{1.23}$, we have to use the following formula:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.56)^2 + (0.38)^2 - 2(0.56)\,(0.38)\,(0.69)}{1 - (0.69)^2}}$$

$$= \sqrt{\frac{0.3136 + 0.1444 - 0.293664}{1 - 0.4761}}$$

$$= \sqrt{\frac{0.458 - 0.293664}{0.5239}}$$

$$= \sqrt{\frac{0.164336}{0.5239}}$$

$$= \sqrt{0.313678} = 0.56$$

18.7 MULTICOLLINEARITY IN MULTIPLE REGRESSION

At times, we come across a problem in multiple-regression analysis that is known as multicollinearity. This problem arises on account of a high degree of correlation between the independent variables, which reduces the reliability of the regression coefficients. Suppose that our data set consists of three

variables—Y is production of steel, X_1 is the index of industrial production, and X_2 is the GNP (Gross National Product). It should be obvious that industrial production (X_1) and GNP (X_2) are not independent of each other as industrial production is one of the components in GNP. In case both these variables have a perfect correlation between them, then it is unnecessary to use both of them. Only one variable can be used and that will give a better result.

Multicollinearity is a problem of degree. When the degree of correlation is minor among independent regression variables, the effect of multicollinearity may not be serious. In contrast, when there is a strong correlation, then the effect of multicollinearity is serious on the regression as it affects it adversely. In such cases, it is advisable to find out the extent of multicollinearity existing in regression.

One of the methods used for this purpose is a correlation matrix of the independent regression variables. It is an array of pairwise correlation coefficients between the independent variables X_i. Table 18.6 is an example of a correlation matrix based on hypothetical data.

Table 18.6 A Correlation Matrix

Variables	X_1	X_2	X_3	X_4	...
X_1	1				
X_2	0.8	1			
X_3	0.7	0.6	1		
X_4	0.6	0.5	0.9	1	
.					
.					
.					

It can be seen from the correlation matrix that we can find out as to which independent variables are highly correlated with one another and thus pose the problem of multicollinearity when they both are included in the regression equation. In our example, we find that variables X_3 and X_4 are highly correlated having 0.9 correlation coefficient. Again, variables X_1 and X_2 with 0.8 correlation coefficient are highly correlated. A high degree of correlation suggests that the two variables contain more or less the same information about Y that gives rise to multicollinearity when both the variables are included in the regression equation. It may be mentioned that correlation is shown as 1 in the correlation matrix whenever the same variable is involved, which is obvious as relationship between X_1 and X_1 will always be one. Same is the case with other variables. When correlation is 1 between say, X_1 and X_2, it is advisable to drop one variable from the regression equation.

In conclusion, it may be said that the problem of multicollinearity is an important one. Whenever we come across a case of multicollinearity or we suspect that it exists on the basis of information available, we should try to solve it The simplest method of solving such a problem is to remove the closely related variables from the regression equation. Alternatively, it is advisable to collect additional information to resolve this problem.

Additional Examples

Example 18.7 In a three variate (X_1, X_2 and X_3) multiple correlation analysis, the following results were found ($\bar{X}_i$ = mean, S_i = *SD* of X_i and r_{ij} = correlation coefficient between X_i and X_j = 1, 2, 3 in a sample of size 20):

$r_{12} = 0.6$ $r_{13} = 0.4$ $r_{23} = 0.2$ $\bar{x}_1 = 60$ $\bar{x}_2 = 70$

$\bar{x}_3 = 80$ $s_1 = 3$ $s_2 = 4$ $s_3 = 5$

Find the regression line of x_1 on x_2 and x_3. Also find X_1 when $x_2 = 74$ and $x_3 = 85$.

Solution The regression equation of X_1 on X_2 and X_3 is

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_2}\right)(X_3 - \bar{X}_3)$$

$$X_1 - 60 = \left[\frac{0.6 - (0.4)(0.2)}{1 - (0.2)^2}\right]\left(\frac{3}{4}\right)(X_2 - 70) + \left[\frac{0.4 - (0.6)(0.2)}{1 - (0.2)^2}\right]\left(\frac{3}{4}\right)(X_3 - 80)$$

$$X_1 = 60 + \left(\frac{0.52}{0.96} \times \frac{3}{4}\right)(X_2 - 70) + \left(\frac{0.28}{0.96} \times \frac{3}{4}\right)(X_3 - 80)$$

$$X_1 = 60 + 0.4\,(X_2 - 70) + (0.21875)\,(X_3 - 80)$$

or $X_1 = 60 + 0.4\,X_2 - 28 + 0.21875\,X_3 - 17.5$

or $X_1 = 14.5 + 0.4\,X_2 + 0.219\,X_3$

When $X_2 = 74$ and $X_3 = 85$, then X_1 will be

$$X_1 = 14.5 + (0.4 \times 74) + (0.219 \times 85)$$

or $X_1 = 14.5 + 29.6 + 18.615$

or $X_1 = 62.715$ or 63 approx.

Example 18.8 Find the multiple linear regression of X_1 on X_2 and X_3 from the data relating to three variables given below:

X_1:	10	38	25	29	16	14
X_2:	12	39	20	31	16	23
X_3:	10	23	16	13	11	6

Solution

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_1^2	X_2^2	X_3^2
10	12	10	120	100	120	100	144	100
38	39	23	1482	874	897	1444	1521	529
25	20	16	500	400	320	625	400	256
29	31	13	899	377	403	841	961	169
16	16	11	256	176	176	256	256	121
14	23	6	322	84	138	196	529	36
132	141	79	3579	2011	2054	3462	3811	1211

$\Sigma X_1 = 132$ $\Sigma X_2 = 141$ $\Sigma X_3 = 79$

$\Sigma X_1X_2 = 3579$ $\Sigma X_1X_3 = 2011$ $\Sigma X_2X_3 = 2054$

$\Sigma X_1^2 = 3462$ $\Sigma X_2^2 = 3811$ $\Sigma X_3^2 = 1211$

The three normal equations are:

$$\Sigma X_1 = Na + b_{12.3}\,\Sigma X_2 + b_{13.2}\,\Sigma X_3$$
$$\Sigma X_1X_2 = a\Sigma X_2 + b_{12.3}\,\Sigma X_2^2 + b_{13.2}\,\Sigma X_2X_3$$
$$\Sigma X_1X_3 = a\Sigma X_3 + b_{12.3}\,\Sigma X_2X_3 + b_{13.2}\,\Sigma X_3^2$$

Substituting the above values in these equations,

$$132 = 6a + 141b_{12.3} + 79b_{13.2} \quad (1)$$
$$3579 = 141a + 3811b_{12.3} + 2054b_{13.2} \quad (2)$$
$$2011 = 79a + 2054b_{12.3} + 1211b_{13.2} \quad (3)$$

Multiplying equation (1) by 141 and equation (2) by 6, we get

$$18612 = 846a + 19881b_{12.3} + 11139b_{13.2} \quad (4)$$
$$\underline{-}\,21474 = \underline{-}\,846a \underline{+}\,22866b_{12.3} \underline{+}\,12324b_{13.2} \quad (5)$$
$$-2862 = \quad -2985b_{12.3} - 1185b_{13.2} \quad (6)\text{By subtraction}$$

or $$2862 = 2985b_{12.3} + 1185b_{13.2} \quad (6)$$

Multiplying equation (2) by 79 and equation (3) by 141, we get

$$282741 = 11139a + 301069b_{12.3} + 162266b_{13.2} \quad (7)$$
$$283551 = \underline{-}\,11139a \underline{+}\,289614b_{12.3} \underline{+}\,170751b_{13.2} \quad (8)$$
$$-810 = \quad +11455b_{12.3} - 8485b_{13.2} \quad (9)$$

Multiplying equation (6) by 11455 and equation (9) by 2985,

$$32784210 = 34193175b_{12.3} + 13574175b_{13.2} \quad (10)$$
$$-2417850 = 34193175b_{12.3} - 25327725b_{13.2} \quad (11)$$
$$+ \qquad - \qquad +$$

$$35202060 = 38901900b_{13.2}$$

$\therefore$ $$b_{13.2} = 0.9$$

Substituting the value of $b_{13.2} = 0.9$ in equation (6),

$$2862 = 2985\,b_{12.3} + (1185 \times 0.9)$$

or $$2862 = 2985\,b_{12.3} + 1066.5$$

or $$2985\,b_{12.3} = 2862 - 1066.5$$

or $$b_{12.3} = \frac{1795.5}{2985}$$
$$= 0.6$$

Substituting the value of $b_{12.3} = 0.6$ and the value of $b_{13.2} = 0.9$ in equation (1), we get

$$132 = 6a + (141 \times 0.6) + (79 \times 0.9)$$

or $$132 = 6a + 84.6 + 71.1$$

or $$6a = 132 - 84.6 - 71.1$$

or $$6a = -23.7$$

$\therefore$ $$a = \frac{-23.7}{6}$$
$$= -3.95$$

Hence, the regression equation is

$$X_1 = -3.95 + 0.6\,X_2 + 0.9\,X_3$$

Example 18.9 From heights (X_1) in inches, weights (X_2) in kg and ages (X_3) in years of a group of students, the following means, variances and correlation were obtained

$$\bar{x}_1 = 40 \qquad \bar{x}_2 = 50 \qquad \bar{x}_3 = 20$$
$$s_1 = 3 \qquad s_2 = 2 \qquad s_3 = 2$$
$$r_{12} = 0.4 \qquad r_{23} = 0.5 \qquad r_{13} = 0.7$$

where $\bar{x}_i$ = mean of x, s_i^2 = var (x_i) and y_{ij} = correlation x_i for $i, j = 1, 2, 3$. Find the multiple regression equation of x_3 on x_1 and x_2 and estimate the value of x_3 when $x_1 = 43$ inches and $x_2 = 54$ kg.

Solution Multiple regression equation of X_3 on X_1 and X_2 is:

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

where

$$b_{31.2} = \frac{\sigma_3}{\sigma_1} \cdot \frac{r_{31} - r_{32}\, r_{12}}{1 - r_{12}^2} \text{ and}$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2} \cdot \frac{r_{32} - r_{31}\, r_{12}}{1 - r_{12}^2}$$

Substituting the values in $b_{31.2}$:

$$b_{31.2} = \frac{2}{3} \times \frac{0.7 - (0.5 \times 0.4)}{1 - (0.4)^2}$$

$$= \frac{2}{3} \times \frac{0.7 - 0.20}{1 - 0.16}$$

$$= \frac{2}{3} \times \frac{0.5}{0.84}$$

$$= \frac{1.0}{2.52} = 0.397 \quad \text{or} \quad 0.4$$

Substituting the values in $b_{32.1}$:

$$b_{32.1} = \frac{2}{2} \times \frac{0.5 - (0.7 \times 0.4)}{1 - (0.4)^2}$$

$$= 1 \times \frac{0.5 - 0.3}{0.84}$$

$$= 1 \times \frac{0.2}{0.84}$$

$$= 0.238 \text{ or } 0.24$$

Hence, the required equation is

$$X_3 - 20 = 0.4\,(X_1 - 40) + 0.24(X_2 - 50)$$

or

$$X_3 = 20 + 0.4(X_1 - 40) + 0.24(X_2 - 50)$$

When $X_1 = 43$ and $X_2 = 54$, then X_3 will be

$$X_3 = 20 + 0.4\,(43 - 40) + 0.24\,(54 - 50)$$
$$= 20 + 1.2 + 0.96$$
$$= 22.16 \text{ years.}$$

Example 18.10 From a trivariate distribution, the following correlation coefficients are obtained:

$r_{12} = 0.7 \qquad r_{13} = 0.6, \qquad r_{23} = 0.4.$

Show that $1 - R^2_{1.23} = (1 - r^2_{12.3})\,(1 - r^2_{13.2})$, where $R_{1.23}$ and $r_{13.2}$ are multiple correlation coefficient and partial correlation coefficient, respectively.

Solution Substituting the values given in the question in the following formula,

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1 - r^2_{12})\,(1 - r^2_{23})}}$$

$$= \frac{0.6 - (0.7 \times 0.4)}{\sqrt{[1 - (0.7)^2][1 - (0.4)^2]}}$$

$$= \frac{0.32}{\sqrt{0.51 \times 0.84}}$$

$$= \frac{0.32}{\sqrt{0.4284}}$$

$$= \frac{0.32}{0.6545} = 0.4889$$

Formula for multiple correlation $R_{1.23}$ is:

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

Substituting the values given in the above formula,

$$R^2_{1.23} = \frac{(0.7)^2 + (0.6)^2 - (2 \times 0.7 \times 0.6 \times 0.4)}{1 - (0.4)^2}$$

$$= \frac{0.49 + 0.36 - 0.336}{1 - 0.16}$$

$$= \frac{0.514}{0.84} = 0.6119$$

$$1 - R^2_{1.23} = 1 - 0.6119 = 0.3881$$

$1 - R^2_{1.23} = (1 - r^2_{12.3})\,(1 - r^2_{13.2})$

$r_{13.2} = 0.4889$ as calculated earlier

Now,

$$1 - R^2_{1.23} = [1 - (0.7)^2]\,[1 - (0.4889)^2]$$
$$= (1 - 0.49)\,(1 - 0.239)$$
$$= 0.51 \times 0.761$$
$$= 0.3881$$

This is the same figure as obtained earlier

Hence, $1 - R^2_{1.23} = (1 - r^2_{12.3})\,(1 - r^2_{13.2})$

Example 18.11 In a problem involving three variables, it is given that

$$r_{12} = 0.41, \quad r_{13} = 0.71, \quad r_{23} = 0.65$$

Find $R_{1.23}$ and $r_{12.3}$.

Solution

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.41)^2 + (0.71)^2 - (2 \times 0.41)(0.71)(0.65)}{1 - (0.65)^2}}$$

$$= \sqrt{\frac{0.1681 + 0.5041 - (0.82)(0.71)(0.65)}{1 - 0.4225}}$$

$$= \sqrt{\frac{0.1681 + 0.5041 - 0.37843}{0.5775}}$$

$$= \sqrt{\frac{0.29377}{0.5775}}$$

$$= \sqrt{0.5086924}$$

$$= 0.71$$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$$

$$= \frac{0.41 - (0.71 \times 0.65)}{\sqrt{1 - 0.71^2}\ \sqrt{1 - 0.65^2}}$$

$$= \frac{0.41 - 0.4615}{\sqrt{1 - 0.5041}\ \sqrt{1 - 0.4225}}$$

$$= \frac{-0.0515}{\sqrt{0.4959}\ \sqrt{0.5775}}$$

$$= \frac{-0.0515}{0.7042 \times 0.7599}$$

$$= \frac{-0.0515}{0.5351}$$

$$= -0.0962$$

Example 18.12 Calculate the equation of regression of X_1 on X_2 and X_3, and estimate X_1 when $X_2 = 165$ and $X_3 = 175$.

Given

$$\bar{X}_1 = 170 \qquad \bar{X}_2 = 160 \qquad \bar{X}_3 = 168$$
$$\sigma_1 = 2.4 \qquad \sigma_2 = 2.7 \qquad \sigma_3 = 2.7$$
$$r_{12} = 0.28 \qquad r_{13} = 0.49 \qquad r_{23} = 0.51$$

Solution

The regression equation of X_1 on X_2 and X_3 is

$$X_1 - \bar{X}_1 = \frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\left(\frac{\sigma_1}{\sigma_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{\sigma_1}{\sigma_2}\right)(X_3 - \bar{X}_3)$$

$$X_1 - 170 = \frac{0.28 - (0.49 \times 0.51)}{1 - (0.51)^2}\left(\frac{2.4}{2.7}\right)(X_2 - 160) + \frac{0.49 - 0.28 \times 0.51}{1 - (0.51)^2}\left(\frac{2.4}{2.7}\right)(X_3 - 168)$$

$$X_1 - 170 = \frac{0.28 - 0.2499}{1 - 0.2601}\left(\frac{2.4}{2.7}\right)(X_2 - 160) + \frac{0.49 - 0.1428}{1 - 0.2601}\left(\frac{2.4}{2.7}\right)(X_3 - 168)$$

$$X_1 - 170 = \frac{0.0301}{0.7399}\left(\frac{2.4}{2.7}\right)(X_2 - 160) + \frac{0.3472}{0.7399}\left(\frac{2.4}{2.7}\right)(X_3 - 168)$$

$$X_1 - 170 = \frac{0.0301}{0.7399} \times 0.8888\,(X_2 - 160) + \frac{0.3472}{0.7399} \times 0.8888\,(X_3 - 168)$$

$$X_1 - 170 = \frac{0.02675}{0.7399}\,(X_2 - 160) + \frac{0.3086}{0.7399}\,(X_3 - 168)$$

$$X_1 - 170 = 0.0361(X_2 - 160) + 0.4171(X_3 - 168)$$
$$X_1 - 170 = 0.0361X_2 - 5.776 + 0.4171X_3 - 70.0728$$
$$X_1 = 170 + 0.0361X_2 - 5.776 + 0.4171X_3 - 70.0728$$

Hence, $X_1 = 94.1512 + 0.0361X_2 + 0.4171X_3$

When $X_2 = 165$ and $X_3 = 175$, then X_1 will be

$$X_1 = 94.1512 + (0.0361 \times 165) + (0.4171 \times 175)$$
$$X_1 = 94.1512 + 6.0885 + 72.9925$$
$$X_1 = 173.2322$$

or $X_1 = 173$ approx.

18.8 ADVANTAGES AND LIMITATIONS OF MULTIPLE CORRELATION ANALYSIS

Advantages

Multiple correlation analysis is useful in as much as it shows the degrees of association between one variable taken as a dependent variable while the remaining variables two or more are taken as the independent variables.

Another advantage of multiple correlation coefficient is that it serves as a measure of goodness-of-fit for a given series of data. As a result, it is regarded as a measure of the accuracy of estimates made by reference to the estimating equation.

Limitations

A major limitation of multiple correlation analysis is that it assumes that the relationship amongst the variables is linear. However, we find that in practice a large number of relationships are not linear and follow some other pattern. As such, the linear regression coefficients are unable to describe curvilinear data.

Another limitation is based on the assumption that the effects of independent variables on the dependent variable are quite separate from each other and hence additive. Accordingly, a given change in one independent variable has the same effect on the dependent variable regardless of the size of the other independent variable or variables.

Finally, the amount of work involved in the calculation of multiple linear correlation is enormous. This apart, it is not so easy to interpret the results accurately as very few persons are well-versed with the basic concept. The method, on the whole, is complex.

At the end, it may be noted that in this chapter the discussion was confined to only three variables; the calculations involved in case of three variables can be carried out smoothly with a pocket calculator. However, when the number of variables is four or more, calculations become tedious and even a pocket calculator may take quite some time. In such cases, we may have to use computers. The increasing use of computers has greatly facilitated statisticians to handle complex problems of multivariate analysis. In fact, many specific programmes based on the requirements of statisticians are available.

GLOSSARY

Coefficient of multiple correlation, R	The positive square root of R^2.
Coefficient of multiple determination, (R^2)	The proportion of total sum of squares (SST) that is explained by the multiple regression model. It measures how well the multiple regression fits the given data.
Computed F ratio	A statistic used to test the significance of the regression as a whole.
Computed t	A statistic used for testing the significance of an independent variable.
Model	A general mathematical relationship relating a dependent variable (Y) to independent variables $X_1, X_2, X_3, \ldots, X_k$.
Multicollinearity	A statistical problem in multiple–regression analysis arising from the existence of correlation between two or more independent variables. It reduces the reliability of regression coefficients.
Multiple regression	A technique of analysing data, which simultaneously investigates the effect of two or more independent variables on a dependent variable. It is, thus, an extension of the simple regression technique.
Partial correlation	It is the correlation between two variables while the remaining variables are held constant.

Partial regression coefficients	In a multiple regression, these are the coefficients of independent variables. The name 'partial' suggests that each one of these measures the effect of the corresponding independent variable on the dependent variable when the remaining independent variables are held constant.
Standard error of a regression coefficient	A measure of uncertainty regarding the actual value of a regression coefficient.

LIST OF FORMULAE

1. Multiple linear regression model:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

2. Estimated multiple regression model:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

3. Normal equations when two independent variables are involved:

$$\Sigma Y = na + b_1 \Sigma X_1 + b_2 \Sigma X_2$$

$$\Sigma X_1 Y = a\Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2$$

$$\Sigma X_2 Y = a\Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2$$

4. Standard error of estimate:

$$S_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - k - 1}}$$

where $n - k - 1$ stands for degrees of freedom.

5. Coefficient of multiple determination:

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

where $\text{SSR} = \Sigma (\hat{Y} - \bar{Y})^2$ and $\text{SST} = \Sigma (Y - \bar{Y})^2$

6. Value of the test statistic F for the test of overall significance of the multiple regression model:

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n - k - 1)}$$

where SSR stands for regression sum of squares (i.e. the explained part) and SSE for error sum of squares (i.e. the unexplained part), k for numerator degrees of freedom and $n - k - 1$ for denominator degrees of freedom.

QUESTIONS

18.1 Given below are twelve statements. Indicate in each case whether the statement is true or false:

(a) The multiple linear regression is not a superior analytical tool compared to the simple linear regression.

(b) Multiple regression enables us to use more of the information available as compared to simple regression.

(c) When there is a high degree of correlation between explanatory variables, we should always be able to identify the separate contribution of these variables in a regression.

(d) In order to test whether the value of Y in a multiple regression is really dependent on the value of X_1, the null hypothesis to be tested is $B_1 = 0$.

(e) Even when a regression model includes all the relevant explanatory variables, the residuals need not be random.

(f) In the partial correlation, the impact of other independent variables is held constant.

(g) The coefficient of multiple determination shows how well the multiple regression fits the given data.

(h) If we know SST and SSR for a multiple regression, then SSE can be easily calculated.

(i) In a multiple regression exercise, t statistic is used for testing the significance of an explanatory variable.

(j) The problem of multicollinearity arises when the independent variables are not mutually statistically independent.

(k) To test the significance of multiple linear regression as a whole, F test is used.

(l) Inclusion of additional variables in a multiple regression must always result in the reduction of the standard error of estimate.

Multiple Choice Questions (18.2 to 18.9)

18.2 In the multiple regression equation $y = a + b_1x_1 + b_2x_2$, y is independent of x_1 when

(a) $b_1 = 1$ (b) $b_1 = -1$ (c) $b_1 = 0$ (d) none of these

18.3 Given a regression equation $\hat{Y} = 23 + 4.7x_1 + 3.9x_2 - 1.7x_3$, the value of b_2 in this equation is

(a) 23 (b) –1.73 (c) 3.9 (d) 4.7

18.4 When a multiple correlation coefficient $R_{1.23} = 1$, then it shows

(a) reasonably good relationship
(b) lack of linear relationship
(c) perfect relationship
(d) none of these

18.5 When a multiple correlation coefficient $R_{1.23} = 1$, then $R_{2.13}$ is

(a) 1 (b) –1 (c) 0 (d) none of these

18.6 The range of a partial correlation coefficient $r_{12.3}$ is

(a) 0 to 1 (b) –1 to 0 (c) –1 to 1 (d) none of these

18.7 While calculating the standard error of estimate, the degrees of freedom used are $n - k - 1$. What does the k stand for?

(a) number of independent variables
(b) number of observations in the sample
(c) mean of the sample values of the dependent variable
(d) none of these

18.8 In a multiple regression problem, the presence of multicollinearity would show

(a) Low standard errors for the coefficients
(b) A sharp increase in a t value for one of the variables when another variable is dropped from the model

(c) Significant t values for the coefficients
(d) All of the above

18.9 Which of the following is the correct expression of r^2?
(a) 1 – SST/SSR (b) 1 – SSR/SST (c) 1 – SSE/SST (d) none of these

18.10 Distinguish between simple linear regression and multiple linear regression.

18.11 Why is it necessary to prefer multiple regression to simple regression in estimating a dependent variable?

18.12 What are the regression coefficients involved in a problem of multiple regression? How would you interpret them?

18.13 Describe the steps involved in the process of multiple regression analysis.

18.14 Do you think that the procedure in multiple regression differs significantly from that used in simple regression? Why or why not?

18.15 What is partial correlation? How does it differ from multiple correlation?

18.16 What is multicollinearity in multiple regression? Explain its concept by giving a suitable example.

18.17 When is multiple regression needed? Explain with the help of an example.

18.18 How are the coefficients of independent variables in a multiple regression model interpreted? Explain.

18.19 What are the assumptions of a multiple regression analysis?

18.20 Find the multiple linear regression equation of X_1 on X_2 and X_3 from the data relating to three variables given below:

X_1:	11	17	26	28	31	35	41	49	63	69
X_2:	2	4	6	5	8	7	10	11	13	14
X_3:	2	3	4	5	6	7	9	10	11	13

18.21 Ishwinder, owner of a business unit, is concerned about the sales behaviour of his product. He narrates that there are many factors that might help explain sales but believes that advertising and prices are major determinants. He has collected the following data:

Sales (Units sold)	33	61	70	82	17	24
Advertising (No. of ads.)	3	6	10	13	9	6
Price (Rs)	125	115	140	130	145	140

(i) Calculate the regression equation to predict sales from advertising and price.
(ii) If advertising is 7 and price is Rs 132, what sales would you predict?

18.22 Find the multiple 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	3
X_3:	30	24	20	14	10	4

Also predict the value of X_1 when $X_2 = 10$ and $X_3 = 22$.

18.23 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 (years) (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?

18.24 The following table shows the data for three variables X_1, X_2 and X_3. Find the least square regression equation of X_3 on X_1 and X_2.

X_1:	12	30	42	54	72	90
X_2:	14	12	8	6	5	3
X_3:	2	3	4	7	10	16

If $X_1 = 20$ and $X_2 = 10$, what will be the corresponding value of X_3?

18.25 Compute the coefficient of multiple linear correlation of X_3 on X_1 and X_2 from the data given below:

X_1:	3	5	6	8	12	14
X_2:	16	10	7	4	3	2
X_3:	90	72	54	42	30	12

18.26 Ashwin, owner of a business unit, is concerned about the sales pattern of his product. He realises that there are many factors that might help explain sales, but believes that advertising and prices are major determinants. He has collected the following data:

Sales (Units sold)	37	65	75	87	22	29
Advertising (No. of ads.)	7	10	14	17	13	10
Price (Rs)	129	115	140	130	145	140

(i) Calculate the regression equation to predict sales from advertising and price.

(ii) If advertising is 11 and price is Rs 132, what sales would you predict?

18.27 In a trivariate distribution:

$\bar{x}_1 = 28.2$ $\quad \bar{x}_2 = 4.91$ $\quad \bar{x}_3 = 594$

$S_1 = 4.4$ $\quad S_2 = 1.1$ $\quad S_3 = 80$

$r_{12} = 0.80$ $\quad r_{13} = -0.56$ $\quad r_{23} = -0.40$

Find the regression coefficient $r_{23.1}$ and $r_{1.23}$. Also estimate the value of X_1 when $X_2 = 6.0$ and $X_3 = 650$.

18.28 Sums, sums of squares and sums of products of three different characters x_1 x_2 and x_3 for 20 sample units are given below. Set up the linear regression equation for estimating x_1 from x_2 and x_3. Also calculate the multiple correlation coefficient of x_1 x_2 and x_3.

$\Sigma x_1 = 1{,}908$ $\quad \Sigma x_2 = 1{,}541$ $\quad \Sigma x_3 = 1{,}502$

$\Sigma x_1^2 = 1{,}84{,}766$ $\quad \Sigma x_2^2 = 1{,}19{,}951$ $\quad \Sigma x_3^2 = 1{,}13{,}104$

$\Sigma x_1 x_2 = 1{,}48{,}344$ $\quad \Sigma x_2 x_3 = 1{,}15{,}754$ $\quad \Sigma x_3 x_1 = 1{,}43{,}626$

18.29 The following information about a trivariate population is known:

$\bar{x}_1 = 60$	$\bar{x}_2 = 70$	$\bar{x}_3 = 100$
$\sigma_1 = 3$	$\sigma_2 = 4$	$\sigma_3 = 5$
$r_{23} = 0.4$	$r_{31} = 0.6$	$r_{12} = 0.7$

Determine the regression equation of X_1 on X_2 and X_3.

18.30 The multiple regression equation is given by $Y = 17.2 + 1.2X_1 + 2.9X_2$ where Y = sales in rupees, X_1 = advertisement expenses and X_2 = number of personal selling agents. Interpret the values of 1.2 and 2.9 and obtain the value of Y for $X_1 = 100$ and $X_2 = 20$.

18.31 For the past twelve months, the manager of RINNO detergent has been running a series of advertisements in the local newspaper. The advertisements (ads) are scheduled and paid for a month before they appear. Each ad contains a two-for-one-coupon, which entitles the bearer to get one RINNO soap of lesser value free with his/her purchase of RINNO soap. The manager has collected and tabulated the following data. Fit a regression equation.

Month	*Number of ads*	*Cost of ads ('000 Rs.)*	*Sale of RINNO Soaps ('000 Rs)*
1	8	10.2	36.5
2	6	8.4	22.6
3	8	11.4	37.6
4	10	11.1	35.2
5	12	13.9	43.6
6	11	12.0	38.0
7	9	9.3	30.1
8	7	9.7	35.3
9	12	12.3	46.4
10	8	11.4	34.2

18.32 Price, x, and advertising expenditure, z, are possible explanatory variables for sales, y. The following data are available from six test market areas.

Sales y (Units)	*Price x (Rs)*	*Advertising expenditure z (Rs)*
1000	10	100
1080	9	120
1200	9	200
1020	10	50
910	11	0
850	12	50

Determine the regression equation. Interpret the constants in terms of this problem.

The McGraw-Hill Companies

CHAPTER 19 TIME SERIES ANALYSIS AND FORECASTING

Learning Objectives

By the end of your work on this chapter, you should be able to

- recognise and define different components of a time series
- obtain trend, seasonal index, cyclical and irregular movements by using appropriate methods
- understand the importance as well as the different methods of forecasting
- measure the forecasting error
- understand and apply criteria for choosing an appropriate forecasting method.

Chapter Prerequisites

Before starting work on this chapter, make sure you are quite conversant with

1. the basic ideas of graphic presentation of data
2. the method of least squares

19.1 INTRODUCTION

Whenever one looks at an economic situation, whether it relates to an individual firm, industry or a nation as a whole, one will find a continuous movement of economic activity. In order to describe this flow of economic activity, the statistician uses a time series. The term 'time series' indicates a set of observations concerning any activity against different periods of time. The duration of time period may be hourly, daily, weekly, monthly or yearly. The following are few examples of time series data:

1. Profits earned by a company for each of the past five years.
2. Workers employed by a company for each of the past 15 years.
3. Number of students registered for the MBA programme of an institute for each of the past five years.
4. The weekly wholesale price index for each of the past 30 weeks.
5. Number of fatal road accidents in Delhi for each day for the past two months.

One can thus think of several types of time series relating to various aspects. Before we proceed further to the analysis of time series, it is necessary to understand the importance of a time series analysis.

Importance of Time Series Analysis

There are several reasons for undertaking a time series analysis. *Firstly*, the analysis of a time series enables us to **understand the past behaviour or performance**. We can know how the data have changed over time and find out the probable reasons responsible for such changes. If the past performance, say, of a company, has been poor, it can take corrective measures to arrest the poor performance.

Secondly, a time series analysis helps directly in business planning. A firm can know the long-term trend in the sale of its products. It can find out at what rate sales have been increasing over the years. This may help it in making projections of its sales for the next few years and plan the procurement of raw material, equipment and manpower accordingly.

Thirdly, a time series analysis enables one to study such movements as cycles that fluctuate around the trend. A knowledge of cyclical pattern in certain series of data will be helpful in making generalisations in the concerned business or industry.

Finally, a time series analysis enables one to make meaningful comparisons in two or more series regarding the rate or type of growth. For example, growth in consumption at the national level can be compared with that in the national income over specified period. Such comparisons are of considerable importance to business and industry.

The foregoing advantages of a time series analysis clearly point out its utility to business and industry. Having looked into the major advantages of a time series analysis, we now turn to the components of a time series.

19.2 COMPONENTS OF A TIME SERIES

A time series may contain one or more of the following four components:

1. Secular trend
2. Seasonal variation
3. Cyclical variation
4. Irregular variation

It is ordinarily assumed that there is a multiplicative relationship amongst these four components. In symbols, this can be written as

$$Y = T \times S \times C \times I$$

where Y is the result of the four components, T = trend, S = seasonal variation, C = cyclical variation and I = irregular variation. An alternative approach is to assume that the relationship amongst these compotents is additive, that is,

$$Y = T + S + C + I$$

Let us understand the meaning of these terms.

1. *Secular trend* shows that the variable has either increased or decreased over a long period of time. For example, to understand the rate of growth in per capita income over the last 20 years, we will first study original data and then, by using one of the several methods available for time series analysis, determine the secular trend. In the same manner, we can ascertain the secular trend in sale of cars. Thus, secular trend for any variable of interest can be obtained. Figure 19.1(a) shows secular trend superimposed on original data.
2. *Seasonal variation* The second component in time series is seasonal variation, which shows patterns of change within a year. Such change repeats itself from year to year. For example,

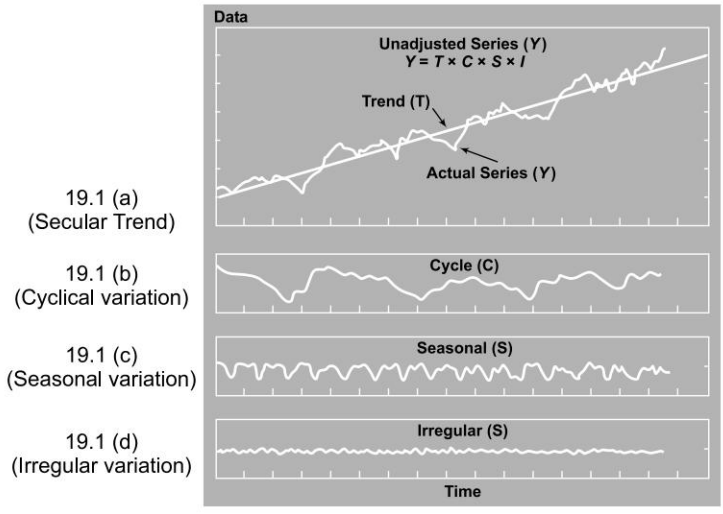


Fig. 19.1 **Components of a Time Series**

woollens are in demand during the winter season. Likewise, sale of cold drinks is more in summer as compared to other seasons. Another example of seasonal variation is found in the production of sugar. The production is high during the season when mills obtain fresh crop of sugar cane as raw material. Figure 19.1(c) shows the seasonal variation.

3. *Cyclical Fluctuations* The third component in time series is cyclical variation or fluctuation, In business, there are some periods when the business activity is at its peak, while in some other periods, it recedes and even goes below the trend line. In fact, business cycle shows wavelike variability, i.e., swings of excessive activity followed by slack times and *vice versa*. Such ups and downs go on repeatedly over time. The time between peak activity and slack season must be at least one year, and it may even go to 15 to 20 years. It may be noted that cyclical fluctuations do not follow any regular pattern. Figure 19.1(b) shows the cyclical movements.
4. *Irregular Variation* This is the fourth and the last component in time series. As the name implies, such variations are irregular, that is, unpredictable. One cannot say whether there would be rise or fall in a certain variable. For example, the occurrence of earthquake in a certain country would throw all activity out of gear. Invasion of a country by another is another example. It is generally believed that irregular variations are usually of smaller magnitudes. Figure 19.1(d) shows the movement of irregular variation.

There is no unanimity amongst statisticians in this regard. The effects of these components might be multiplicative, additive or might be a combination of several other forms. On account of this, different assumptions will give different results. Having said this, we may add that the assumption that the time series consists of multiplicative components is most frequently used.

As a time series involves the methods by which the different components can be separated, we shall now discuss these methods beginning with the trend.

19.3 THE TREND

The trend is the long-term movement of a time series. In other words, an increase or decrease in the values of a variable occurring over a period of several years gives a trend. If the values of a variable remain stationery over several years, then we can say that there is no trend in that time series. When we study the growth in industrial production from 1990 to 2000, we are trying to find the trend in industrial production for this time period. The trend may be increasing or decreasing, as will be clear from two graphs presented below:

It will be seen that Fig. 19.2 shows both actual sales and the trend. It is a graph of increasing trend as is shown by a straight line passing through the actual sales data. In contrast, Fig. 19.3 shows a decreasing trend as sales have declined over time. It may be noted that both the trends are linear. The trends can be non-linear as well. There are some other types of trend such as parabolic and logarithmic. We are concerned here with only the straight–line trends.

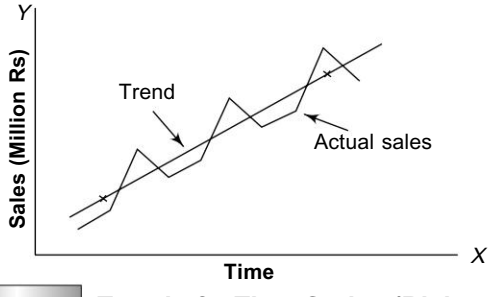


Fig. 19.2 **Trend of a Time Series (Rising Trend)**

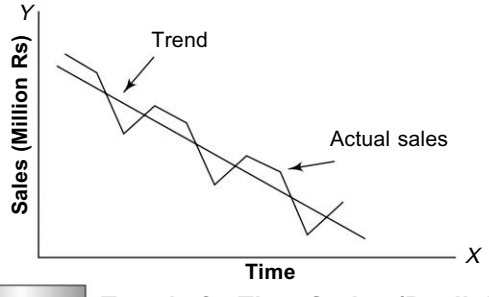


Fig. 19.3 **Trend of a Time Series (Declining Trend)**

There are various methods of fitting a straight line to a time series such as freehand method, the method of semi-averages, the method of moving averages, and the method of least squares. These methods are discussed below.

The Freehand Method

This is the simplest method of finding a trend line. The procedure involves first the plotting of the time series on a graph and then fitting a straight line through the plotted points in such a way that the straight line shows the trend of the time series. This is illustrated with the hypothetical data in Table 19.1.

Table 19.1 **Sale of Product *Y*, 1995–2001**

Year (X)	*Sale (Y) Million Rs*
1995	10
1996	20
1997	15
1998	25
1999	37
2000	35
2001	40

Let X show the year, starting with 1995, and Y show the sale of a product in million rupees. The sale figures are plotted in Fig. 19.4. We now join the points thereby getting the production line from 1995 to 2001. We draw a straight line by observation that gives a broad trend of the sales. It may be noted that the freehand method is not an accurate method of fitting a trend as different persons may fit different trend lines to the same set of data.

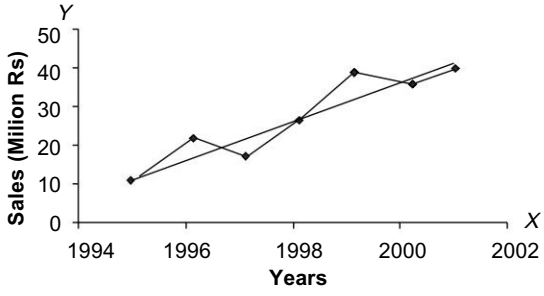


Fig. 19.4 **Trend by Freehand Method**

Let us assume that the trend line goes through the points for 1995 and 2000. Now, the trend line can be found by setting up the equation for the straight line that passes through the two points 1995 and 2000.

In order to ascertain the required equation, we first assign a sequence to the different years for which the data are given, starting from 1995, which is indicated by zero, that is, the *year of origin*. Then 1996 is 1, 1997 is 2 and so forth. Since the straight line connects two points, let us assume that these two points are at the point of origin and at the end of 2001. Thus, the coordinates of the two points selected become (0,10) and (6,40). On the basis of this information, we may formulate the following two equations for a straight line:

$$Y = a + bX$$

$$10 = a + 0b \quad (1)$$

$$40 = a + 6b \quad (2)$$

Solving the two equations gives $a = 10$ and $b = 5$. This gives the equation for the trend line as

$$\hat{Y} = 10 + 5X$$

Origin : 1st July 1995

X : 1-year units

It may be noted that we have used the symbol $\hat{Y}$ instead of Y. This is because the values of Y obtained from the equation are the *computed* or the *estimated* values and are not the actual ones. Another point to note is that it is necessary to specify the origin as well as the units of X. The general practice is to take the middle of the year (i.e., 1st July) as the point to represent the data for that year.

Advantages

1. The Freehand method is a very *simple method* of estimating trend.
2. There is *flexibility* in this method as it can be used regardless whether the trend is a straight line or non-linear.
3. If the statistician is well-conversant with the movement of the particular variable involved in the time series, the use of this method may even give a *better expression* to the long-term movement than the trend fitted by a rigid mathematical formula.

Limitations

1. This is a *highly subjective* method as the trend line fitted to the same set of data by this method will vary from one person to another.
2. In view of the trend line being highly subjective, it is *not an appropriate method for making predictions*.
3. It means very long experience on the part of the statistician to use this method, otherwise, the trend fitted would not be of much use.

The Method of Semi-averages

When the method of semi-averages is used, the given time series is divided into two parts preferably with the same number of years. The average of each part is calculated and then a trend line through these averages is fitted. This is illustrated with the data given in Table 19.2.

Table 19.2 Production from 1994–2001

Year	*X*	*Y*	*Average*
1994	0	10	
1995	1	12	
1996	2	18	60/4 = 15
1997	3	20	
1998	4	20	
1999	5	25	
2000	6	23	100/4 = 25
2001	7	32	

The average of first part of the data is 15 and that of the second part is 25. Since 15 is the average of 1994, 1995, 1996 and 1997, 15 is plotted in between 1995 and 1996, which is the middle of the 4-year period. Likewise, 25 is plotted in between 1999 and 2000. Then these points are joined by a straight line, which is a semi-average trend line. This is shown in Fig. 19.5.

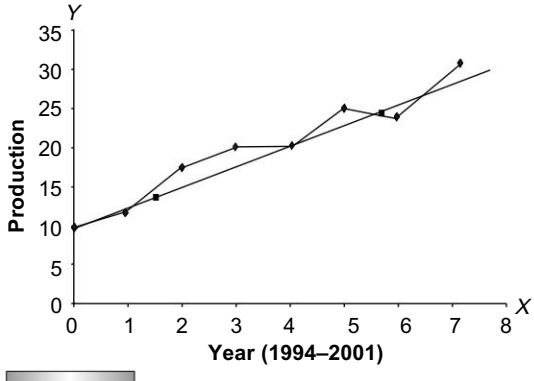


Fig. 19.5 Trend Line by the Method of Semi-averages

On the basis of the semi-average trend line, the following two simultaneous equations are arrived at:

$$15 = a + b \quad (1)$$

$$25 = a + 4b \quad (2)$$

Solving for a and b gives us $a = 11.67$ and $b = 3.33$. This gives the equation of the trend line

$$\hat{Y} = 11.67 + 3.33X$$

Origin: 1st July 1994

X : 1-year units

Applying this equation to estimate the production for 1996,

$$\hat{Y} = 11.67 + (3.33 \times 2)$$
$$= 18.33$$

As we know, the actual production in 1996 was 18. This shows that the discrepency between actual and estimated production is

$$\hat{Y}_{96} - Y_{96} = 18.33 - 18 = 0.33$$

It will be seen that there was a very close fit in this example because of a very negligible difference between the actual and the estimated figures.

It may be noted that in case of a time series pertaining to odd number of years, either the middle year is excluded from the computation or the series may be split in two inequal parts. A point worth emphasising is that in case a particular year has been an abnormal year on account of such factors as a prolonged strike disrupting production, it is advisable to omit that year so that the trend line may be more realistic.

Advantages

1. The method of semi-averages is simple to use as anyone can use it conveniently.
2. This is an objective method as anyone applying it to the given set of data would get the same trend line.

Limitations

1. This method will always give a straight line trend regardless of the nature of the given set of data. Thus, it assumes a straight line relationship, which may not be true.
2. This method may give wrong trend-line on account of the limitations of arithmetic average. In case there is an extreme value in either half or both halves of the time series, then the trend line will not give a realistic growth of the phenomenon being measured. In order to overcome this problem, it is necessary to ensure that the data do not have extreme values.

The Method of Moving Averages

The method of moving averages is used not only to fit trend lines but also to seasonal and cyclical variations. The following example illustrates how the method is used.

Example 19.1 Suppose we are given a time series data for 12 years—1989 to 2000 relating to sales of a certain business firm. These data are given below:

Year	*Sales (Million Rs)*	*Year*	*Sales (Million Rs)*
1989	10	1995	15
1990	15	1996	24
1991	20	1997	15
1992	25	1998	21
1993	15	1999	15
1994	12	2000	24

We are asked to find out the three-year moving averages, starting from 1989.

Solution Table 19.3 shows these figures. Taking the first three years—1989, 1990 and 1991, we add up their sales—10 + 15 + 20 = 45. This figure is written in column 3 of Table 19.3 against the mid-year, that is, 1990. Now, we drop the year 1989 and include the year 1992 in our calculation. Thus, the total of three years 1990, 1991 and 1992, which comes to 60, is entered in the third column against 1991. This process is continued until we reach the last three years—1998, 1999 and 2000. It may be noted that Table 19.3 does not have a total for the first year, 1989 and the last year, 2000.

The next step involves the calculation of moving averages by dividing the moving totals by 3. The moving averages are shown in the last column of Table 19.3.

Table 19.3 Calculation of Moving Averages

Year	*Sales (Million Rs)*	*Three-year Moving Total*	*Three-year Moving Average*
1989	10		
1990	15	45	15
1991	20	60	20
1992	25	60	20
1993	15	52	17
1994	12	42	14
1995	15	51	17
1996	24	54	18
1997	15	60	20
1998	21	54	18
1999	18	63	21
2000	24		

The moving averages have been plotted on a graph as shown in Fig. 19.6. In order to fit a trend line to these moving average points on the graph, the free-hand method, the semi-average method, the method of least squares (which is discussed later in this chapter) may be used. We have joined two points of moving averages to fit a trend line.

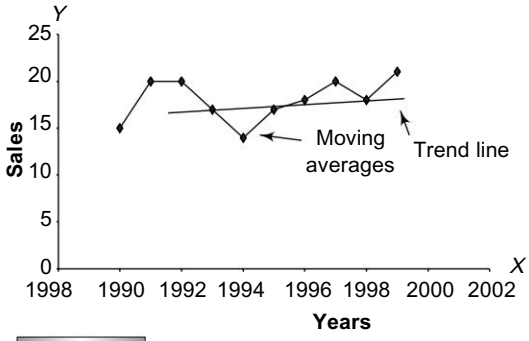


Fig. 19.6 **Graph showing Moving Averages and the Trend Line**

Effective Application of Moving Average Method In order to ensure that the moving average method is appropriate and it is applied effectively, it is necessary to ascertain first whether a regular, periodic cycle in the time series exists. In several cases, one would find that there is certain regularity in the series to allow the use of the moving average method. It may also be noted that if the basic nature of the time series is linear, it will give a linear trend. In case it is curvilinear, then its trend will be a curve. Another point to note is that the use of the moving average method is not confined to trend lines alone but is also extended to varying types of data that show regular periodic fluctuations. Later, we shall use it for removing seasonal fluctuations from a time series.

Advantages

1. The method of moving averages is simpler method than the method of least squares.
2. The general movement of the data is reflected in the moving averages. As such, the trend line is based on the data rather than on the statistician's choice of a mathematical function.
3. In case the period of moving averages coincides with that of cyclical fluctuations in the data, this method automatically removes such fluctuations.

Limitations

1. The method does not cover all the years so far as trend values are concerned.

2. It needs to be exercised very carefully as there are no definite rules regarding the period of moving averages.
3. This method cannot be used for forecasting as it is not represented by a mathematical function.
4. If the trend is non-linear, the moving averages may considerably deviate from it.

The Method of Least Squares

Among the methods of fitting a straight line to a series of data, this is the most frequently used method. As we have seen earlier that the equation of a straight line is $Y = a + bX$, where Y is the time period, say, year, X is the value of the item measured against time, a is the Y-intercept and b is the coefficient of X that shows the slope of the straight line. In order to find out the values of a and b, the following two equations are solved:

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

where n is the total number of observations in a series. These equations are called the normal equations.

This method, discussed earlier in Chapter 16 on Regression Analysis, is known as the method of least squares because deviations of the observed points (actual data) from the trend line (computed values) are the least, if it is used.

The following example will show how the method of least squares is used. It contains the series having odd numbers. This is followed by another example where the series has even number of years.

Example 19.2 ***Odd Number of Years***

Year X	*1996*	*1997*	*1998*	*1999*	*2000*
Y	270	285	295	315	330

We are required to fit a trend to these data, using the method of least squares.

Solution We know from the normal equations that we need the values of ΣY, ΣX, ΣXY and ΣX^2 so that the normal equations can be solved. In order to get these values, the following worksheet is set up.

Table 19.4 Worksheet

Year	*X*	*Y*	*XY*	X^2
1996	–2	270	–540	4
1997	–1	285	–285	1
1998	0	295	0	0
1999	1	315	315	1
2000	2	330	660	4
Total	**0**	**1,495**	**150**	**10**

The values are: $\Sigma X = 0$, $\Sigma Y = 1{,}495$, $\Sigma XY = 150$ and $\Sigma X^2 = 10$

Substituting these values in the two normal equations, we get

$$1{,}495 = 5a + b(0) \quad (1)$$
$$150 = a(0) + b(10) \quad (2)$$

Equation (1) gives the value of a as 1,495/5 = 299 and equation (2) gives the value of b as 150/10 = 15. Thus, the straight line trend that we get is

$$\hat{Y} = 299 + 15X$$

Origin : 1-7-1998

X : 1-year unit

Now, in order to fit the trend to the original data on a graph, first we plot the original values against the respective years. We then decide location of any two points as two points are needed to draw a straight line. Let us compute the value of X for 1997 and 2000.

$$\hat{Y}\ 1997 = a + bX$$
$$= 299 + 15\ (-1) = 284$$
$$\hat{Y}\ 2000 = a + bX$$
$$= 299 + 15\ (2) = 329$$

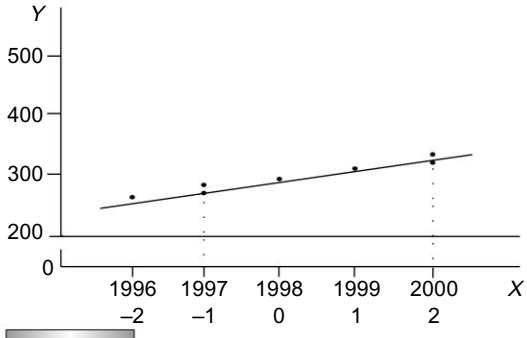


Fig. 19.7 **Trend Line Fitted to Data of Table 19.4**

We now draw a straight line with the help of these two calculated values. This is shown in Fig. 19.7.

We now take another example, which consists of even number of years.

Example 19.3 ***Even Number of Years***

Suppose we have the following data.

Year (X)	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
Y	15	14	18	20	17	24

We have to fit a trend to these data, again using the method of least squares.

Solution We set up the following worksheet.

Table 19.5 Worksheet

Year	*X*	*Y*	*XY*	X^2
1995	–5	15	–75	25
1996	–3	14	–42	9
1997	–1	18	–18	1
1998	1	20	20	1
1999	3	17	51	9
2000	5	24	120	25
Total	**0**	**108**	**56**	**70**

The two normal equations are:

$$\Sigma Y = na + b\Sigma X \quad (1)$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad (2)$$

Substituting the values in the two equations, we get

$$108 = 6a + b(0)$$
$$56 = a(0) + b(70)$$

The first equation gives $a = 108/6 = 18$ and the second equation gives $b = 56/70 = 0.8$. Thus, the trend line is

$$\hat{Y} = 18 + 0.8X$$

Origin : 1-1-1998

X : half-year units

A comparison of this equation with that in the earlier example will show that there are two changes here. *First*, unlike in the previous example, the origin is half-way between 1st July 1997 and 1st July 1998, that is, 1st January 1998. *Second*, the Xs are in half-year units. This is because the Xs are indicated as 1, 3, 5 and each year varies by 2 units. Thus, to move from 1st July 1998 to 1st July 1999, X has to move from 1 to 3. When X has to move from 1 to 2, it means it moves only half year, that is, from 1st July 1998 to 1st January 1999.

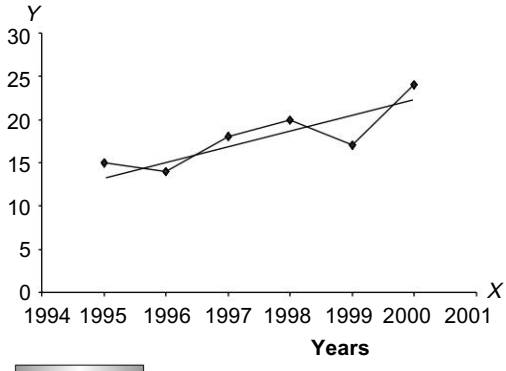


Fig. 19.8 **Trend Line Fitted to Data of Table 19.5**

Figure 19.8 shows the plot of original data along with the trend line.

If we want to avoid half-year units and would like to retain one-year X units, then we may multiply the X values by 1/2. In this example, the X has been given the values as

Year	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
	–5	–3	–1	1	3	5
	–2.5	–1.5	– 0.5	0.5	1.5	2.5

The last row in the above table shows the revised values of X, where X unit becomes 1 year instead of half-year.

Changing the Unit Value So far our discussion related to the trend equation pertaining to annual data. But in many cases the time series are given as annual monthly averages and monthly data. The trend line equation will be different in the three cases viz. the annual total, annual monthly average and monthly data. An illustration will bring out this distinction clearly.

Example 19.4 Suppose the annual salary of a person for 1998 was Rs 24,000, which was raised to Rs 27,600 in 1999 and to Rs 31,200 in 2000. Using this information, explain how the trend line equation will be differnt in case of annual total, annual monthly average and monthly data.

Solution This means that his monthly average salary was Rs 2,000 for 1998, Rs 2,300 for 1999 and Rs 2,600 for 2000. Thus, the annual increase of the monthly average salary was Rs 300. The monthly increase of the monthly average salary was Rs 300/12 = Rs 25. On the basis of these data, the following three equations can be constructed.

1. The annual total equation:

$$\hat{Y} = \text{Rs } 24{,}000 + 3{,}600\,X$$

$X = 0$ at July 1, 1999

where X is in 1-year units

2. The monthly average equation :

$$\hat{Y} = (\text{Rs } 24{,}000/12) + (3{,}600/12)\,X$$

or $\hat{Y} = \text{Rs } 2{,}000 + 300\,X$

$X = 0$ at July 1, 1999

where X is in 1-year units

3. The monthly equation :

$$\hat{Y} = (\text{Rs } 24{,}000/12) + [3{,}600/(12 \times 12)]\,X$$

or $\hat{Y} = \text{Rs } 2{,}000 + 25X$

$X = 0$ at July 1, 1999

where X is in 1-month units

It may be noted that the b coefficient in equation (2), which is Rs 3,600/12 = 300, shows the annual increase of the monthly average salary. In equation (3), the b coefficient is Rs $3{,}600/(12 \times 12) = 25$, which shows the monthly increase of the monthly average salary.

In keeping with our earlier approach of using the middle of the year to represent the annual data, it is necessary to use the same approach for monthly data. That is, the origin should be July 15 instead of July 1 as shown in equation (3).

In order to shift the origin from July 1, 1999 to July 15, the value of 'a' (in the equation $Y = a + bX$) has to be changed.

$$\hat{Y} = \text{Rs } 2{,}000 + 25 \times 1/2$$
$$= 2{,}000 + 12.5 = 2012.5$$

This will give the monthly equation for 15th July 1999 as

$$\hat{Y} = \text{Rs } 2{,}012.5 + 25X$$

Origin : 15-7-1999

X : 1-month units

Shifting the Origin The monthly average equation used in the preceding section was

$$\hat{Y} = \text{Rs } 2{,}000 + 300X$$

Origin : July 1, 1999

X : 1-year units

Suppose we are interested in shifting the origin to July 1, 2001. This can be done by finding the new Y-intercept a_{2001}. This is

$$a_{2001} = 2{,}000 + (300 \times 2) = 2{,}600$$

Thus, the equation becomes

$$\hat{Y} = 2{,}600 + 300X$$

Origin : July 1, 2001

X : 1-year units

In the same manner, we can shift the trend value to January 2001 instead of July 2001. What we have to do is to subtract the value of X for a six-month period from the value of Y-intercept. Taking our earlier example,

$$\hat{Y} = 2{,}600 + 300X$$

Origin : July 1, 2001

X : 1-year units

We can rewrite thus

$$\hat{Y} = 2{,}600 - (300 \times 1/2) + 300X$$

or $\hat{Y} = 2{,}450 + 300X$

Origin : January 1, 2001

X : 1-year units

Higher Degree Polynomial Trends

So far we have discussed a linear trend, which in fact is a polynomial trend of the first degree. The need to fit a curve to a time series is because at times a linear fit does not turn out to be a best fit. Second or third degree polynomial may be a better fit to a given time series.

In general, a polynomial trend has the form

$$\hat{Y} = a + bX + cX^2 + dX^3 + \ldots + rX^k$$

and the degree of polynomial is indicated by the highest component to which X is raised. It may be noted that as the degree of polynomial increases, a new term is added to the equation and one new direction of slope is possible. In practice, one finds that a polynomial higher than third degree is seldom used. We shall discuss here second degree polynomial only.

Second-degree Polynomial A second-degree polynomial is also known as a *quadratic trend* or *parabola*. It is given by the equation

$$\hat{Y} = a + bX + cX^2$$

Given a set of data, the trend can be obtained by the usual least square method. We explain this by taking an example.

Example 19.5 Suppose we have the following data:

Table 19.6 Production of TV Sets, 1995–1999

Year	*Production of TV Sets ('000)*
1995	2
1996	4
1997	8
1998	14
1999	22

In the second degree parabola, there will be three normal equations as shown below:

$$\Sigma Y = na + b\Sigma X + c\Sigma X^2$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$
$$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

If we put $\Sigma X = 0$, the above normal equations become

$$\Sigma Y = na + c\Sigma X^2 \quad (1)$$
$$\Sigma XY = b\Sigma X^2 \quad (2)$$
$$\Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4 \quad (3)$$

Let us set up a worksheet to obtain the numerical values of these terms.

Table 19.7 Worksheet for Calculating a Polynomial Trend

Years	X	X^2	Y	XY	X^2Y	X^3	X^4
1995	–2	4	2	–4	8	–8	16
1996	–1	1	4	–4	4	–1	1
1997	0	0	8	0	0	0	0
1998	1	1	14	14	14	1	1
1999	2	4	22	44	88	8	16
Total	**0**	**10**	**50**	**50**	**114**	**0**	**34**

Applying these values in the above equations, we get

$$50 = 5a + 10c \quad (1)$$
$$50 = 10b \quad (2)$$
$$114 = 10a + 34c \quad (3)$$

From (2) we find $b = 50/10 = 5$

Again,

$$50 = 5a + 10c \quad (1)$$
$$114 = 10a + 34c \quad (3)$$

Multiplying (1) by (2) and subtracting it from (3)

$$114 = 10a + 34c \quad (3)$$
$$100 = 10a + 20c \quad (4)$$
$$14 = 14c \qquad \therefore\ c = 1$$

Now, subtracting the value of $c = 1$ in (1),

$$50 = 5a + 10 \quad (1)$$
$$\therefore \quad 5a = 50 - 10$$
$$a = 40/5 = 8$$

Now, we have $a = 8$, $b = 5$ and $c = 1$

$$Y = a + bX + cX^2$$

or $$Y = 8 + 5X + 1X^2$$

Origin : July 1, 1997

X : 1-year units

If we have to find out the trend value for the year 2000, then

$$X = 3$$
$$Y = a + bX + cX^2$$

or $$Y = 8 + (5 \times 3) + (1 \times 9)$$

or $$Y = 32$$

Let us see how well the second degree parabola has fitted to the time series data. This is shown in Fig. 19.9. It will be seen that the parabolic trend has fitted extremely well to the original data. In fact, the $\hat{Y}$ values, if calculated, turn out to be the same as the original values.

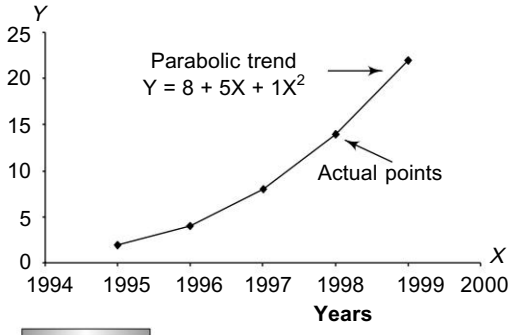


Fig. 19.9 **Parabolic Trend Fitted to Data of Table 19.9**

On the basis of the parabola fitted to the time series, we can say that the production of TV sets in that firm would be 32,000 in the year 2000. A word of caution is necessary here. One has to be more careful while using a parabola for forecasting when one is using a linear trend. This is because the slope of the second degree curve is continually increasing as can be seen from Fig. 19.9. Further, it is sometimes risky to use a parabola to forecast further into the future. In order to have a realistic forecast, one has to be fully aware of factors that may either increase or slacken the growth rate.

In case a given time series has an even number of elements, we can proceed in the same manner as was shown earlier for a linear trend. In such a case, the origin will shift to 1st January of the concerned year and X will be in half-year units.

19.4 SEASONAL VARIATION

Our discussion so far related to one component of the time series, the secular trend T. We now discuss another component—seasonal variation.

A seasonal variation is a recurring fluctuation that has a duration of one year. Some examples of seasonal variation are the consumption of soft drinks, sale of woollen cloth, sale of dry fruits, and so on. Soft drinks are taken during the summer season and during winter months their consumption is extremely low. In contrast, the woollens are in great demand during the winter season only. The same is the case of dry fruits, which are largely consumed in winter. Again, on account of major festivals there are fluctuations in the sale and consumption of certain items. For example, during Diwali festival, one finds that the sale of cloth is the maximum. In West Bengal, this pattern is noticeable during the Durga Pooja festival.

Reasons for Measuring Seasonal Variation If one knows in advance that the sales of a specific product are bound to be high in a particular month or week, one may hold a larger stock of it to match the increased demand in that month or week. In the same manner, if one knows that the price of some commodity moves in a particular fashion, one may buy it while the price is low and hold it for subsequent use or sale when the price is high. This will be true for any raw material that goes into the production of finished goods.

Yet another reason for measuring seasonal variation is to adjust the data statistically for such variation. It is easier to interpret seasonally adjusted data as it would reduce the confusion which might otherwise arise. For instance, if the data on sale of some commodity are not seasonally adjusted, a seasonal upswing in those data may be taken as an improved performance. This would obviously be wrong.

Our main concern here is with periodic series that have a period of one year. We will now focus on the methods used for measuring seasonal variation.

There are four methods of measuring seasonal variation:

1. The method of simple averages
2. The ratio to trend method
3. The moving average method
4. The link relative method

These are discussed below with suitable examples.

The Method of Simple Averages

This method can be better explained with the help of an example. Suppose we are given monthly data for two years as shown in the first two columns of Table 19.8.

Table 19.8 Consumption of Electricity (KWhrs)

Month	*1999 (1)*	*2000 (2)*	*Total (1999 + 2000) (3)*	*A. Mean (4)*	*Seasonal (5)*
January	342	392	734	367	116.48
February	309	349	658	329	104.42
March	299	343	642	321	101.88
April	268	312	580	290	92.04
May	250	290	540	270	85.69
June	236	274	510	255	80.93
July	242	282	524	262	83.15
August	263	305	568	284	90.13
September	288	328	616	308	97.75
October	321	365	686	343	108.86
November	343	379	722	361	114.57
December	365	417	782	391	124.10
			Total	**3,781**	**1,200.00**
			Average	**315.08**	**100.00**

It may be noted that (i) column (3) gives the total of two months for 1999 and 2000 (ii) column (4) shows the arithmetic mean for the two months. (iii) The total of column (4) comes to 3,781. This is divided by 12 in order to obtain the monthly average, which is 315.08. (iv) In order to obtain the seasonal variation *S*, figures given in column (4) are divided by 315.08 and multiplied by 100 as it is customary to express the seasonal factor as a base of 100. The values of *S* are shown in the last column of the table.

The above illustration consists of only two years because our purpose is to explain the method for computing seasonal variation. In real life, one may have to use a longer time series. Moreover, it is advisable to use a longer time series, which will give more realistic results. It may be noted that the above series is related to monthly data but the method, explained above, can be applied to any series with weekly or quarterly data.

The Ratio to Trend Method

Another method of measuring seasonal variation is the ratio to trend method. This method is an improvement over the method of simple averages. It involves a number of steps as indicated below:

(i) By using the method of least squares, the trend values are obtained.

(ii) The original time series is to be divided by the trend values obtained earlier. These figures are to be transformed into percentages. Now these percentage figures have three components of the time series, viz. seasonal, cyclical and irregular as the trend has been eliminated.

(iii) The percentages are to be averaged for each month or quarter or for any other time period in which the original data are available. This process will eliminate the effects of both cyclical and irregular movements. It may be noted that while averaging the percentages, median should be preferred to the arithmetic mean. This is because the latter gives undue weightage to extreme values, which are mainly on account of seasonal swings.

(iv) The figures obtained in (iii) above need to be adjusted to a total of 1,200 for the monthly series and to 400 for the quarterly series. This adjustment is carried out by multiplying each figure by a constant 'K'. The formula for this adjustment is:

For the monthly series: K = 1,200/Sum of the indices

For the quarterly series: K = 400/Sum of the indices

Let us take an example that uses this method for measuring seasonal variation.

Example 19.6 Find seasonal variation by the ratio-to-trend method for the following data:

Year	*Quarters*			
	1st	*2nd*	*3rd*	*4th*
1997	75	60	54	59
1998	88	65	63	80
1999	95	74	66	85
2000	100	78	73	93
2001	118	100	88	110

Solution For calculating the trend values, we have to use the method of least squares. For straight-line trend, $Y = a + bX$, the two normal equations are:

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

In order to find the numerical values for these terms, we have to set up a worksheet, which is given in Table 19.9.

Table 19.9 Computation of Straight-line Trend

Year	*Year X*	*Annual Total of Q. Data*	*Quarterly Average for the Year Y*	*XY*	X^2	*Trend Values*
1997	–2	248	62	–124	4	62
1998	–1	296	74	–74	1	71.6
1999	0	320	80	0	0	81.2

(*Contd.*)

(*Contd.*)

2000	1	344	86	86	1	90.8
2001	2	416	104	208	4	100.4
Total	**0**	**1,624**	**406**	**96**	**10**	

Applying these values in the two normal equations,

$$406 = 5a + 0b \quad \text{(i)}$$

$$96 = 0a + 10b \quad \text{(ii)}$$

$\therefore$ From equation (i), $a = 406/5 = 81.2$

From equation (ii), $b = 96/10 = 9.6$

From the above we get the trend line

$$\hat{Y} = 81.2 + 9.6X$$

Origin: July 1, 1999

X 1-year : units

As can be seen, yearly increment in the trend value is $b = 9.6$. Hence the increment per quarter is $9.6/4 = 2.4$. On this basis, we now determine the quarterly trend values, which are shown in Table 19.10.

For 1st July 1997, the trend value is $81.2 + (9.6 \times -2) = 62$. As the quarterly increment is 2.4, half of quarterly increment is 1.2. We obtain the trend values for 2nd and 3rd quarters of 1997 as $62 - 1.2$ and $62 + 1.2$, that is, 60.8 and 63.2 respectively. As a result, the trend value for the 1st quarter of 1997 would be $60.8 - 2.4 = 58.4$ and for the 4th quarter is $63.2 + 2.4 = 65.6$.

Table 19.10 Quarterly Trend Values

Year	*Quarters*			
	1st	*2nd*	*3rd*	*4th*
1997	58.4	60.8	63.2	65.6
1998	68.0	70.4	72.8	75.2
1999	77.6	80.0	82.4	84.8
2000	87.2	89.6	92.0	94.4
2001	96.8	99.2	101.6	104.0

Table 19.11 Trend Eliminated Values (Quarterly Values as Percentage of Trend Values)

Years	*Quarters*			
	1st	*2nd*	*3rd*	*4th*
1997	128.42	98.68	85.44	89.94
1998	129.41	92.33	86.54	106.38
1999	122.42	92.50	79.14	100.24
2000	114.68	87.64	79.35	98.52
2001	121.90	100.81	86.61	105.77
Total	**616.83**	**471.96**	**417.08**	**500.85**
A. Mean	**123.37**	**94.39**	**83.42**	**100.17**

As the total of arithmetic means of the four quarters (i.e. 123.37 + 94.39 + 83.42 + 100.17) is 401.35, it should be 400 for the quarterly series. It is, therefore, necessary to make some adjustment.

This adjustment can be made by using the following formula:

$$K = \frac{400}{\text{Sum of the quarterly indices}}$$

$$K = \frac{400}{401.35} = 0.996636$$

If we multiply the unadjusted indices by 0.996636, we will get the seasonal indices. Applying this adjustment factor, we get the following seasonal indices:

	Q_1	Q_2	Q_3	Q_4
Seasonal Index	122.96	94.07	83.14	99.83

It will be seen that these four seasonal indices add to 400.

The Moving Average Method

This method uses either of the two relationships

1. $Y = T \times S \times C \times I$
2. $Y = T \times S \times I$

In the second relationship, T is the trend cycle, that is, it consists of both T and C of (1) above.

To begin with, we assume that the seasonal component S has a 12-month period and that its pattern is the same each year. In case we use quarterly data, our assumption is the same—the seasonal component is the same each year. We make another assumption that the irregular variations I are independent for different years. In view of these assumptions, if a 12-month moving average is applied to a series, it will smooth out $S \times I$ and the remainder will be $T \times C$. In other words, the use of the moving average gives the trend and cycle, $T \times C$.

Symbolically,

$$\frac{T \times S \times C \times I}{T \times C} = S \times I$$

where the numerator is the original data and the denominator is the moving average, in the first term (T).

An example will help us in understanding this process clearly.

Example 19.7 Given the following quarterly data, you are asked to calculate the seasonal index.

Year	Q_1	Q_2	Q_3	Q_4
1	68	62	61	63
2	65	58	56	61
3	68	63	63	67
4	70	59	56	62
5	60	55	51	58

Solution The method is the same whether the monthly data or quarterly data are to be used for measuring the seasonal variation. We set up the following worksheet.

Table 19.12 Worksheet for Measuring Seasonal Variation

Year	*Quarters*	*Given Figures*	*4–Fig. Moving Total*	*2–Fig. Moving Total*	*Moving Average (5)/8*	*Given Figs. as Per cent of Moving Average*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
1	1	68				
	2	62	254			
	3	61	251	505	63.1	96.7
	4	63	247	498	62.2	101.3
2	1	65	242	489	61.1	106.4
	2	58	240	482	60.2	96.3
	3	56	243	483	60.4	92.7
	4	61	248	491	61.4	99.3
3	1	68	255	503	62.9	108.1
	2	63	261	516	64.5	97.7
	3	63	263	524	65.5	96.2
	4	67	259	522	65.2	102.8
4	1	70	252	511	63.9	109.5
	2	59	247	499	62.4	94.6
	3	56	237	484	60.5	92.6
	4	62	233	470	58.8	105.4
5	1	60	228	461	57.6	104.2
	2	55	224	452	56.5	97.3
	3	51				
	4	58				

Table 19.13 Calculation of Seasonal Index

Year	*Percentage to Moving Average*				
	I Q	*II Q*	*III Q*	*IV Q*	
1			96.7	101.3	
2	106.4	96.3	92.7	99.3	
3	108.1	97.7	96.2	102.8	
4	109.5	94.6	92.6	105.4	
5	104.2	97.3	–	–	Total ↓
Total	**428.2**	**385.9**	**378.2**	**408.8**	**1,601.1**
Average	**107.1**	**96.5**	**94.5**	**102.2**	**400.3**
Seasonal Index	**107.1**	**96.4**	**94.4**	**102.1**	**400**

It may be noted that in Table 19.13, the starting figure 96.7 is the original figure (61) taken as a percentage of the corresponding moving average (63.1). This figure of 96.7 is shown in the last column of Table 19.12 against third quarter of the 1st year.

Seasonal index is obtained by multiplying each quarterly average by an adjustment factor 400/400.3.

The Link Relative Method

This method, also known as the *Pearson's method*, is relatively complex as it involves the averaging of the link relatives. The use of this method to calculate seasonal indices involves the following steps:

(i) As a first step, link relatives of the seasonal figures are to be calculated. A link relative is the value of one season expressed as a percentage of the preceding season. Suppose our data are monthly, then

$$\text{Link relative} = \frac{\text{Figure for the current month}}{\text{Figure for the preceding month}} \times 100$$

In case the data are quarterly, then the link relative will be

$$\frac{\text{Figure for the current quarter}}{\text{Figure for the preceding quarter}} \times 100$$

(ii) The link relatives for each month or quarter are to be averaged. For this purpose, we can use either the mean or the median. As the arithmetic mean gives undue importance to extreme values, the median is preferable.

(iii) The average link relatives are now to be converted into chain relatives on the basis of the first season (month or quarter). The formula to calculate chain relatives is:

$$\frac{\text{Link relative of the current month} \times \text{Chain relative of the preceding month}}{100}$$

In case of quarterly data,

$$\frac{\text{Link relative of the current quater} \times \text{Chain relative of the preceding quarter}}{100}$$

As this will not add to 100 due to trend effect, some adjustment is needed.

(iv) The adjustment can be done by subtracting a 'correction factor' from each chain relative. For example, the correction factor (CF) for January will be CF(d) = 1/12 [second (new) CR for January – 100], CF for February will be 2d, for March 3d, and so forth. In quarterly series, the CF(d) will be ¼ (second new CR for 1st Q – 100). For 2nd Q 1d; for 3rd Q 2d and for 4th Q 3d will be the correction factor). This correction factor is used on the assumption that there is a linear trend.

(v) This is the final step wherein we have to ensure that the adjusted chain relatives add up to 1,200 if the data are monthly or to 400 if the data are quarterly. The figures thus arrived at are the required seasonal indices.

Let us take an example to illustrate the use of link relative method.

Example 19.8 Calculate the seasonal indices from the following data using the link relative method.

Year	Q_1	Q_2	Q_3	Q_4
1995	65	58	56	61
1996	68	63	63	67
1997	70	59	56	52
1998	60	55	51	58

Solution We have to first calculate the link relatives. This has been done by dividing the figure for 2nd quarter by the 1st quarter-figure and then multiplied by 100. For example, take the case of 1995 2nd quarter, for which the link relative is (58/65) × 100 = 89.23. In the same manner, link relatives for all quarters have been calculated. Table 19.14 shows these link relatives.

Table 19.14 Computation of Seasonal Indices by the Method of Link Relatives

Year	*Quarters*			
	1st	*2nd*	*3rd*	*4th*
1995	–	89.23	96.55	108.93
1996	111.48	92.65	100.00	106.35
1997	104.48	84.29	94.92	92.86
1998	115.38	91.67	92.73	113.73
Total	**331.34**	**357.84**	**384.20**	**421.87**
Average of LR	110.45	89.46	96.05	105.47
Chain relatives	100	$\frac{89.46 \times 100}{100}$ = 89.46	$\frac{96.05 \times 89.46}{100}$ = 85.93	$\frac{105.47 \times 85.926}{100}$ = 90.62
Adjusted CR	100.09	89.43	85.90	90.55
Seasonal indices	109.33	97.78	93.89	99.00

The new (second) CR for 1st quarter

$$= \frac{\text{LR of 1st quarter} \times \text{CR of last (4th) quarter}}{100}$$

$$= \frac{110.45 \times 90.62}{100} = 100.09$$

The difference between these chain relatives = 100.09 – 100 = 0.09

Difference per quarter = 0.09/4 = 0.0241

Adjustment factor, $d = 0.0241$

Adjusted CR for 2nd, 3rd and 4th quarters are obtained by subtracting d, $2d$ and $3d$ from the corresponding CRs. Finally, seasonal indices are obtained by adjusting the corrected CRs to a total of 400. This is done by multiplying each index by a constant factor K = 400/365.97. This is because the total of seasonal indices for the four quarters should be 400. The total of adjusted CR for the four quarters is 365.97. Hence, the constant factor K is 400/365.97. The adjusted seasonal indices are shown in the last row of Table 19.14.

As can be seen, this method is somewhat cumbersome on account of a number of calculations involved for link relatives followed by adjustments to be made.

The question now is: which of these four methods we should use in order to compute seasonal indices? Each method gives different indices. To a large extent, our choice will depend on the nature of data and the purpose of measuring seasonal variation. Without going further in details, we may say that the ratio-to-moving average method should be preferable as it has certain advantages over other methods.

Deseasonalisation

Having computed the seasonal indices by one of the four methods explained earlier, the question of their application arises. These seasonal indices are used in adjusting a time series. The process of making the necessary adjustment in a time series for seasonal fluctuations is known as deseasonalisation.

In order to deseasonalise a time series, we have to divide a figure in the time series, say, production in January by the corresponding seasonal index. It may be noted that the deseasonalised data will still have trend, cycle and irregular variations.

Symbolically,

$$\frac{Y_i}{S_i} = \frac{T_i \times C_i \times S_i \times I_i}{S_i} = T_i \times C_i \times I_i$$

An example will illustrate the process of deseasonalisation.

Example 19.9 Using the quarterly data for the first two years as given in Column (3) of Table 19.12, compute the deseasonalised series.

Solution Given below are the quarterly data for two years. The data are reproduced from Table 19.12.

Table 19.15 Computation of Deseasonalised Series

		Original Data	*Seasonal Index*	*Deseasonalised Series*
Year 1	Q1	68	107.1	63.49
	Q2	62	96.4	64.32
	Q3	61	94.4	64.62
	Q4	63	102.1	61.70
Year 2	Q1	65	107.1	60.69
	Q2	58	96.4	60.17
	Q3	56	94.4	59.32
	Q4	61	102.1	59.75

It may be noted that the quarterly seasonal indices are only four. As such, to complete the column in the above table, these indices have been repeated. When we have monthly time series, the seasonal index for each month will be different.

The original figure for year 1 quarter 1 is 68. To deseasonalise it, we have to divide it by the seasonal index for the 1st quarter and multiply by 100. Thus,

$(68/107.1) \times 100 = 63.49$

In the same manner, the computation of the deseasonalised figure for the 2nd quarter of year 1 is

$(62/96.4) \times 100 = 64.32$

In this manner, all other calculations have been made and the deseasonalised series is given in the last column of Table 19.15.

19.5 CYCLICAL VARIATION

Given monthly data, we may calculate the seasonal index numbers as explained in the preceding section. The original monthly data are then divided by the corresponding seasonal indices. The resulting data are thus deseasonalised or adjusted for seasonal variation. These data are inclusive of the trend, cyclical and irregular movements.

In order to measure cyclical variations, it is necessary that the deseasonalised monthly data are adjusted for trend. This can be done by dividing the data by corresponding trend values. The resulting data comprise both cyclical and irregular variations. In order to isolate irregular variations from these data, we have to use appropriate moving average of a few months' duration, say 3, 5 or 7 months. This would smooth out the irregular variations and leave only cyclical variations. If, on the other hand, a time series consists of only annual data, then the seasonal variation is non-existent as it completes its cycle within each year. In other words, we need to consider only the secular trend, cyclical and irregular components. As we can know secular trend using a trend line, we can isolate the two components —cyclical and irregular variation from the trend line. We can say that a major part of the variation left unexplained by the trend components is the cyclical variation.

When the time series comprises annual data, we can find out cyclical variation by dividing the actual value (Y) by the corresponding trend value $\hat{Y}$ for each item in the time series. The resultant of this calculation is then multiplied by 100. The figure thus arrived at gives us the measure of cyclical variation as a *per cent of trend*.

Symbolically,

$$C = \frac{Y}{\hat{Y}} \times 100$$

where C = cyclical variation
Y = actual time series value
$\hat{Y}$ = estimated trend value from the same item in the time series

It may be recalled that by applying the method of least squares as shown in Table 19.5, we obtained the trend line as $\hat{Y} = 18 + 0.8X$ when X was ½ year units. Converting it into one-year units gives us the trend line $\hat{Y} = 18 + 1.6X$ (origin : 1-1-1998). On the basis of this equation, we may now calculate $\hat{Y}$ for every single value of Y. This is shown in Table 19.16.

Table 19.16 Calculation of Per cent of Trend

X	Y	$\hat{Y}$	$Y/\hat{Y} \times 100$
1995	15	14	107.1
1996	14	16	87.5

(*Contd.*)

(*Contd.*)

1997	18	17	105.9
1998	20	19	105.3
1999	17	20	85.0
2000	24	22	109.1

The last column of the above table shows the percent of trend. It should be noted that when the two series Y and $\hat{Y}$ are plotted on a graph (Fig. 19.10), the actual values move above and below the trend line.

There is another method of measuring cyclical variation. Here we have to find the percentage deviation from the trend for each value. Symbolically,

$$\text{Relative cyclical residual} = \frac{Y - \hat{Y}}{\hat{Y}} \times 100$$

Where Y = actual time series value

$\hat{Y}$ = estimated trend value from the same point in the time series

The relevant calculations based on this formula are presented in Table 19.17.

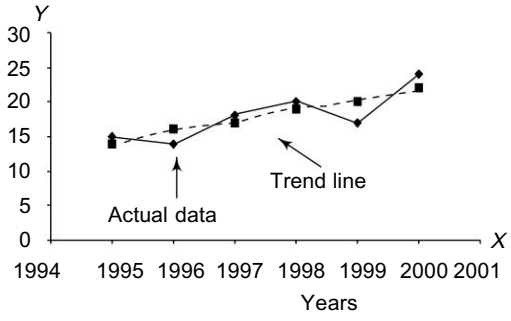


Fig. 19.10 **Cyclical Fluctuations Around the Trend Line for the Data in Table 19.16**

Table 19.17 **Calculation of Relative Cyclical Residuals**

X	Y	$\hat{Y}$	$Y/\hat{Y} \times 100$	$[(Y - \hat{Y})/\hat{Y}] \times 100$
1995	15	14	107.1	7.1
1996	14	16	87.5	–12.5
1997	18	17	105.9	5.9
1998	20	19	105.3	5.3
1999	17	20	85.0	–15.0
2000	24	22	109.1	9.1

Cyclical variations are often plotted as the percentages of trend (Fig. 19.11). It is worth emphasising that the procedure discussed here for cyclical variation can be used only for the past data and not for forecasting cyclical variations in the future. These two measures of cyclical variation, are percentages of the trend. For example, in 1999, the actual value was 85 per cent of the expected value for that year. For the same year, the relative cyclical residual indicated that the actual value was 15 per cent short of the expected value.

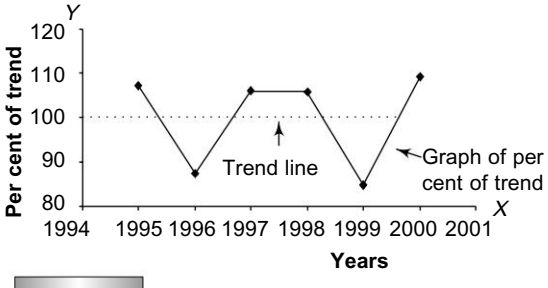


Fig. 19.11 **Graph of Percent of Trend Around the Trend Line for the Data in Table 19.17**

19.6 IRREGULAR VARIATION

In the preceding section, we have seen how cyclical variations can be measured. It may be reiterated that these variations are not merely cyclical but comprise both cyclical and irregular variations. We can obtain an estimate of the irregular variations by the following formula:

$$I = \frac{CI}{C}$$

where I = irregular variation
CI = cyclical and irregular variation
C = cyclical variation

However, in practice, we find that irregular variations are very erratic. In addition, both cyclical variation and irregular variation are so interwoven that it becomes extremely difficult to segregate or isolate irregular variation from cyclical variation. In the analysis of time series, generally, the focus is on trend and seasonal variation—the two main components of the time series.

Although it is extremely difficult to measure irregular variation, we can identify the cause for irregular variation. For example, take the case of strikes and lockouts, which are irregular factors that affect adversely the economic climate of the corporate sector.

Guidelines for Time Series Analysis

Before we pass on to the next section on Forecasting, the following guidelines may be adhered to, so that any mistakes or wrong analysis of time series may be avoided.

1. To begin with, one must plot the original data on a graph sheet. This would give a broad idea of the presence of secular trend, cyclical, as well as seasonal variations.
2. Choose one of the several methods available for constructing secular trend, though the method of least squares is frequently used.
3. In case the original data indicates that there are seasonal fluctuations, it is necessary to calculate a seasonal index. By using this index, the original series may be deseasonalised.
4. The deseasonalised series may be adjusted for trend. This adjusted deseasonalised series contains cyclical and irregular variations. In order to obtain cyclical variations, a moving average of 3,5 or 7 months may be attempted.
5. The cyclical variations obtained in the preceding step, may be plotted.
6. The foregoing steps, along with any additional information that is available may be used for making a forecast, which is discussed in detail in the next section.

19.7 FORECASTING

The discussion so far related to various aspects of time series analysis. We now look into another related aspect, viz. forecasting. Let us start with the question: Why is forecasting necessary?

The Importance of Forecasting

As we all know that economic and business conditions do not remain the same over time. In view of changes in economic and business conditions, it becomes necessary for management to keep itself abreast of the effects that such changes are likely to have on their organisations. In the absence of

realistic forecasts, management may find itself placed in an adverse situation resulting into losses. In fact, the need for forecasting is felt in other spheres as well. For example, take the case of the government. It has to make forecasts in respect of population growth, employment, revenues, etc. so that it can formulate appropriate policies for good governance. In the field of education too, forecasting is important. In the absence of forecasting, educationists are unable to plan for the future.This will have an adverse impact on the quality of education as they would be ill-equipped to provide adequate number of teaching and administrative staff as well as physical facilites. In short, we find that forecasting is quite necessary for planning for the uncertain future in different areas of the economy.

It is worthwhile at this stage to know the actual process of forecasting.

Forecasting Process

There are five steps involved in the forecasting process.

First, one has to decide the objective of the forecast. The statistician should know as to what will be the use of the forecast he is going to make.

Second, the time period for which the forecast is to be made should be selected. Is the forecast short-term, medium-term or long-term? Why should a particular period of forecast be selected?

Third, the method or technique of forecasting should be selected. One should be clear as to why a particular technique from amongst several techniques should be used.

Fourth, the necessary data should be collected. The need for specific data will depend on the forecasting technique to be used.

Finally, the forecast is to be made. This will involve the use of computational procedures.

In order to ensure that the forecast is really useful to the company, there should be good understanding between management and the analyst who is to make the forecast. The management should clearly spell out the purpose of the forecast and how it is going to help the company. It should also ensure that the analyst has a proper understanding of the operations of the company, its environment, past performance in terms of key indicators and their relevance to the future trend. If the analyst is well-informed with respect to these aspects, then he is likely to make more realistic and more useful forecast for the management.

Methods of Forecasting

Figure 19.12 shows different methods of forecasting.

The methods of forecasting can be divided into two broad categories, viz. subjective or qualitative methods and objective or quantitative methods. These can be further divided into several methods. Each of these methods is discussed below.

Subjective Methods In the subjective methods, judgement is an important ingredient. Before attempting a forecast, the basic assumptions regarding environmental conditions as also competitive behaviour must be provided to people involved in forecasting. An important advantage of subjective methods is that they are easily understood. Another advantage is that the cost involved in forecasting is quite low.

As against these advantages, subjective methods have certain limitations also. One major limitation is the varying perceptions of people involved in forecasting. As a result, wide variance is found in forecasts. Subjective methods are suitable when forecasts are to be made for highly technical products which have a limited number of customers. Generally, such methods are used for industrial products.

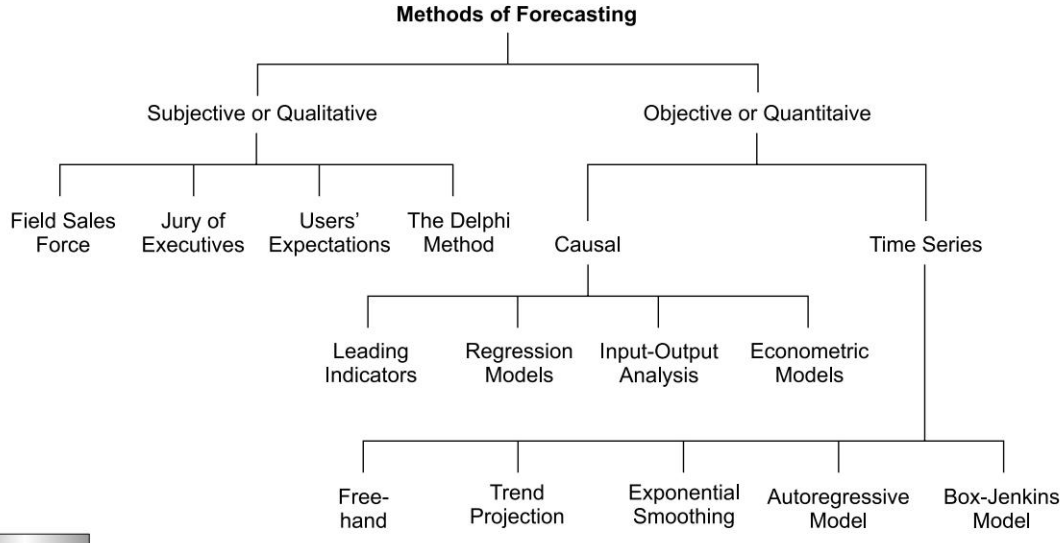


Fig. 19.12 Methods of Forecasting

Also, when cost of forecasting is to be kept minimum, subjective methods may be more suitable.

There are four subjective methods—field sales force, jury of executives, users' expectations and the Delphi method. These are discussed here briefly, the focus being on company sales forecasts.

Field Sales Force Some companies ask their salesmen to indicate the most likely sales for a specified period in the future. Usually the salesman is asked to indicate anticipated sales for each account in his territory. These forecasts are checked by district managers who forward them to the company's head office. Different territory forecasts are then combined into a composite forecast at the head office. This method is more suitable when a short-term forecast is to be made as there would be no major changes in this short period affecting the forecast. Another advantage of this method is that it involves the entire sales force which realises its responsibility to achieve the target it has set for itself. A major limitation of this method is that sales force would not take an overall or broad perspective and hence may overlook some vital factors influencing the sales. Another limitation is that salesmen may give somewhat low figures in their forecasts thinking that it may be easier for them to achieve those targets. However, this can be offset to a certain extent by district managers who are supposed to check the forecasts.

Jury of Executives Some companies prefer to assign the task of sales forecasting to executives instead of a sales force. Given this task each executive makes his forecast for the next period. Since each has his own assessment of the environment and other relevant factors, one forecast is likely to be different from the other. In view of this, it becomes necessary to have an average of these varying forecasts. Alternatively, steps should be taken to narrow down the differences in the forecasts. Sometimes this is done by organising a discussion between the executives so that they can arrive at a common forecast. In case this is not possible, the chief executive may have to decide which of these forecasts is acceptable as a representative one.

This method is simple. At the same time, it is based on a number of different viewpoints as opinions of different executives are sought. One major limitation of this method is that the executives' opinions are likely to be influenced in one direction on the basis of general business conditions.

Users' Expectations Forecasts can be based on users' expecations or intentions to purchase goods and services. It is diffcult to use this method when the number of users is large. Another limitation of this method is that though it indicates users' 'intentions' to buy, the actual purchases may be far less at a subsequent period. It is most suitable when the number of buyers is small such as in case of industrial products.

The Delphi Method This method too is based on the experts' opinions. Here, each expert has access to the same information that is available. A feedback system generally keeps them informed of each others' forecasts but no majority opinion is disclosed to them. However, the experts are not brought together. This is to ensure that one or more vocal experts do not dominate other experts.

The experts are given an opportunity to compare their own previous forecasts with those of the others and revise them. After three or four rounds, the group of experts arrives at a final forecast.

The method may involve a large number of experts and this may delay the forecast considerably. Generally, it involves a small number of participants.

It will be seen that both the jury of executive opinion and the Delphi method are based on a group of experts. They differ in that in the former, the group of experts meet, discuss the forecasts, and try to arrive at a commonly agreed forecast while in the latter the group of experts never meet. As mentioned earlier, this is to ensure that no one person dominates the discussion thus influencing the forecast. In other words, the Delphi method retains the wisdom of a group and at the same time reduces the effect of group pressure. An approach of this type is more appropriate when long-term forecasts are involved.

Quantitative or Objective Methods

These methods can be divided into two broad categories, namely:

1. Causal or Explanatory Methods
2. Time Series forecasting

We first discuss causal or explanatory methods.

Causal or Explanatory Methods Causal or explanatory methods are regarded as the most sophisticated methods of forecasting. These methods yield realistic forecasts provided relevant data are available on the major variables influencing changes in sales. There are three distinct advantages of causal methods. First, turning points in sales can be predicted more accurately by these methods than by time-series methods. Second, the use of these methods reduces the magnitude of the random component far more than it may be possible with the time-series methods. Third, the use of such methods provides greater insight into causal relationships. This facilitates the management in decision-making.

Causal methods are briefly discussed here.

Leading Indicators Sometimes one finds that changes in sales of a particular product or service are preceded by changes in one or more leading indicators. In such cases, it is necessary to identify leading indicators and to closely observe changes in them. One example of a leading indicator is the demand for various household appliances which follows the construction of new houses. Likewise, the demand for many durables is preceded by an increase in disposable income. Yet another example is of number of births. The demand for baby food and other goods for infants can be ascertained by the number of births in a territory. It may be possible to include leading indicators in regression models.

Regression Models Linear regression analysis is perhaps the most frequently used and the most powerful method among causal methods. As we have discussed regression analysis in detail in Chapters 16 and 18, here we shall only dwell on a few relevant points.

1. Regression models indicate linear relationships within the range of observations and at the times when they were made. For example, if a regression analysis of sales is attempted on the basis of independent variables of population sizes of 15 million to 30 million and per capita income of Rs 1000 to Rs 2500, the regression model shows the relationships that existed between these extremes in the two independent variables. If the sales forecast is to be made on the basis of values of independent variables falling outside the above ranges, then the relationships expressed by the regression model may not hold good.
2. Somtimes there may be a lagged relationship between the dependent and independent variables. In such cases, the value of dependent variables are to be related to those of independent variables for the preceding month or year as the case may be. The search for factors with a lead-lag relationship to the sales of a particular product is rather difficult. One should try out several indicators before selecting the one which is most satisfactory.
3. It may happen that the data required to establish the ideal relationship, do not exist or are inaccessible or, if available, are not useful. Therefore, the analyst has to be careful in using the data. He should be quite familiar with the varied sources and types of data that can be used in forecasting. He should also know about their strengths and limitations.
4. Finally, regression model reflects the association among variables. The causal interpretation is done by the analyst on the basis of his understanding of such an association. As such, he should be extremely careful in choosing the variables so that a real causative relationship can be established among the variables chosen.

Input-output Analysis Another method that is used for forecasting is the input-output analysis. Here, the analyst takes into consideration a large number of factors, which affect the outputs he is trying to forecast. For this purpose, input-ouput table is prepared where the inputs are shown horizontally as the column headings and the outputs vertically as the stubs. It may be mentioned that by themselves input-output flows are of little direct use to the analyst. It is the application of an assumption as to how the output of an industry is related to its use of various inputs that makes an input-ouput analysis a good method of forecasting. The assumption states that as the level of an industry's output changes, the use of inputs will change proportionately, implying that there is no substitution in production among the various inputs. This may or may not hold good.

The use of input-output analysis in sales forecasting is appropriate for products sold to governmental, institutional and industrial markets as they have distinct patterns of usage. It is seldom used for consumer products and services. It would be most appropriate when the levels and kinds of inputs required to achieve certain levels of outputs need to be known.

A major constraint in the use of this method is that it needs extensive data for a large number of items which may not be easily available. Large business organisations may be in a position to collect such data on a continuing basis so that they can use input-output analysis for forecasting. However, this is not possible in case of small industrial organisations on account of excessive costs involved in the collection of comprehensive data.

Econometric Models Econometrics is concerned with the use of statistical and mathematical techniques to verify hypotheses emerging in economic theory. An econometric model incorporates functional relationships estimated by these techniques into an internally consistent and logically self-contained framework. The use of econometric models is generally found at the macro level such as forecasting national income and its components. Such models show how the economy or any of its

specific segment operates. As compared to an ordinary regression equation, they bring out the causalities involved more distinctly. This merit of econometric models enables them to predict turning points more accurately. However, their use at the micro-level for forecasting has so far been extremely limited.

Time Series Forecasting Having briefly looked into causal or explanatory methods of forecasting, we now turn to the time series forecasting.

Freehand Method One of the methods of getting a secular trend is the freehand method. We have explained it eariler. It may be reiterated that it is the simplest method of finding the trend line, which is simply extended for forecast. It is highly subjective method as the trend line fitted to the same set of data will vary from one person to another as such it is the most inappropriate method to be used for forecasting.

Trend Projection Another method is the method of least squares in fitting the trend. Earlier in this chapter, some examples have been given using the least squares method. The trend is forecast simply by substituting the appropriate value of t (i.e. the year for which the forecast is desired) in the least squares line, as was done in Section 19.3. In case the data are monthly or quarterly, this value is to be multiplied by the seasonal index. The construction of seasonal index has been earlier explained using different methods. Finally, we measure the cyclical component and try to ascertain what it is likely to be at the point for which the forecast is being made. We multiply this component to obtain TSC. Since the irregular movements cannot be forecast as they are random fluctuations, we forecast the three "regular" components T, S and C.

This method of least squares is far superior though we must remember that all forecasts into the future are based on the assumption that the characteristics displayed by the existing data will continue to influence future values. If this assumption does not hold, even with statistics proving a good fit to known data, forecasts could be most inaccurate.

We now discuss the other three time series methods of forecasting which are being increasingly used in recent times. These are Exponential Smoothing, Autoregressive method and Box-Jenkins model. This discussion will be followed by the discussion on Measuring the Forecasting Error. Finally, the chapter will end with a discussion on the choice of a Forecasting Model, which undoubtedly is very important for obtaining realistic forecasts.

Exponential Smoothing A method which is often useful in forecasting time series is exponential smoothing. When a large number of forecasts are to be made for a number of items, exponential smoothing is particularly suitable as it combines the advantages of simplicity of computation and flexibility. It may be used for short-term forecasts (one period into the future) particularly when there is no long-term trend in a time series data or when the trend is not clear.

This method uses differential weights to time-series data. The heaviest weight is assigned to the most recent data and the least weight to the most remote data in the time series. It is a type of moving average that 'smooths' the time-series of its sharp variations.

The formula used for exponential smoothing is based on three terms:

(i) The present observed value of the time series Y.
(ii) The previous computed exponentially smoothed value E_{i-1}
(iii) A subjectively assigned weighting factor or smoothing coefficient W.

Thus, the formula is

$$E_i = WY_i + (1 - W)\, E_{i-1}$$

where E_i = value of the exponentially smoothed series being computed in time period i

E_{i-1} = value of the exponentially smoothed series computed in the preceding time period $i - 1$

Y_i = observed value of the time series in period i

W = subjectively assigned weight whose value is between 0 and 1.

It should be evident from the above formula that the weighting factor or smoothing coefficient affects the results substantially. As such we have to select it very carefully. In case our purpose is served just by eliminating unwanted cyclical and irregular fluctuations, a small value of smoothing coefficient should be preferred. On this basis the more recent values will have low weightage and exponentially smoothed value will have higher weightage. In case our purpose is forecasting, a higher value of smoothing coefficient should be preferred as it will give a higher weightage to the more recent values and a lower weightage to the exponentially smoothed value.

In order to explain the actual process used in exponential smoothing, let us take an example.

Example 19.10 Sales data of a firm for the years 1995 to 2000 are given below:

Years	*Sales (million Rs)*
1995	15
1996	24
1997	15
1998	20
1999	22
2000	28

Let us select $W = 0.5$ and another $W = 0.3$. We will use these weights so that we will get two series of exponentially smoothed values.

The exponentially smoothed value for 1995, i.e. first year E_1 is simply the observed value for that year, being 15. In other words, E_1 is 15. Now, for subsequent years calculations are given below:

1996 $E_2 = WY_2 + (1 - W)\, E_1$

$= (0.5)\,(24) + (1 - 0.5)\,15$

$= 12 + 7.5$

$=$ Rs 19.5 million

1997 $E_3 = WY_3 + (1 - W)\, E_2$

$= (0.5)\,(15) + (1 - 0.5)\,(19.5)$

$= 7.5 + 9.75$

$=$ Rs 17.25 million

This process will continue until we calculate the exponentially smoothed values for the latest year. In the same manner, calculations have been done with smoothing coefficient $W = 0.3$. Table 19.18 shows the exponentially smoothed values along with the original time series.

Table 19.18 Exponentially Smoothed Values of Sales of a Business Firm

Year	*Sales (million Rs)*	*W = 0.5*	*W = 0.3*
1995	15	15.00	15.00
1996	24	19.50	17.70
1997	15	17.25	16.89
1998	20	18.63	17.82
1999	22	20.32	19.07
2000	28	24.16	21.75

In order to forecast sales for the year 2001, we will take the smoothed value for the latest year in the time series (2000) as its estimate. Symbolicaly, $\hat{Y}_i + 1 = E_i$. For a smoothing coefficient of 0.5 that projection is 24.16 milion rupees and for a smoothing component of 0.3 it is 21.75 million rupees.

Autoregressive Model Another approach to forecasting with annual time series data involves the fitting of an autoregressive model. Sometimes, the values of a time series data are highly correlated with the values that proceed and succeed them. In such cases an autoregression model is used for forecasting.

A first-order autocorrelation refers to the magnitude of the association between consecutive values while a second-order autocorrelation refers to the magnitude of the relationship between values which are two periods apart.

The first-order autoregressive model may be expressed as:

$$\hat{Y}_i = b_0 + b_1 Y_{i-1}$$

The second-order autoregressive model may be expressed as:

$$\hat{Y}_i = b_0 + b_1 Y_{i-1} + b_2 Y_{i-2}$$

It will be seen from the above two models that the first-order autoregressive model is similar to the simple linear regression, while the second-order autoregressive model is similar to the multiple regression having two explanatory variables.

Similarly, we can have the p^{th}-order autoregressive model which may be expressed as:

$$\hat{Y}_i = b_0 + b_1 Y_{i-1} + b_2 Y_{i-2} + \ldots + b_p Y_{i-p}$$

The p^{th}-order autoregressive model deals with relationships between values up to p periods apart.

A question that is very relevant here is: How can we decide as to which model will be the most appropriate? It is rather difficult to know this. We must weigh the advantage of simplicity against the possibility of overlooking an important autocorrelation behaviour that may exist in the data. On the other hand, we must be equally concerned with choosing a high-order model estimation of several parameters, which may be unnecessary particularly when the total number of observations in the time series is not large. This is understandable because when we choose, for example, a second-order model, then there is a loss of first two observations. Each higher-order will involve a loss of one additional observation. In view of this, one should be careful in deciding which order of model should be used. It may be noted that a statistical software package is invariably used to calculate the values of parameters in autoregressive models.

Box-Jenkins Method We may now briefly describe the Box-Jenkins method of forecasting, which uses a very different approach than what we have discussed so far. First of all, the analyst identifies a tentative model considering the nature of the past data. This tentative model and the data are entered in the computer. The Box-Jenkins programme then gives the values of the parameters included in the model. A diagnostic check is then conducted to find out whether the model gives an adequate description of the data. If the model satisfies the analyst in this respect, then it is used to make the forecast. In case the model is not satisfactory, then the computer points out diagnostic information, which is then used by the analyst in revising the model. This process is continued until the analyst obtains an appropriate model, which is used for making forecasts.

It may be pointed out that some studies used the Box-Jenkins model as well as some other methods and found that the Box-Jenkins model gave more accurate forecasts as compared to other methods. There is, however, limitation of this method that it requires at least 45 observations in the time series.

Measuring the Forecast Error It should always be remembered that all forecasts are based on the assumption that the characteristics displayed by the existing data will continue to influence future values. In case this assumption does not hold, even a very good forecasting technique may provide a most inaccurate forecast. The management must aim at maximum accuracy of the forecast, and in the majority of forecasting situations accuracy is indeed regarded the most important criterian for selecting a forecasting technique.

While there are several measures that have been suggested to measure the forecast error, a measure that is most commonly used is called the Mean Absolute Deviation (MAD). This is based on the size of the absolute value of the residuals.

Symbolically, $$MAD = \frac{\sum_{i=1}^{n} | Y_i - \hat{Y}_i |}{n}$$

where | | represents the absolute value.

MAD is thus an average of the absolute deviations between the actual (Y_i) and fitted ($\hat{Y}_i$) values in a particular time series. In case a forecast model fits the past time series perfectly, then the MAD is zero. When the forecasting model fits the past time series poorly, then the MAD is large. The concept of MAD is very useful when there are two or more models. Obviously in such cases the model which gives the lowest value will be the most appropriate and, as such, it should be chosen.

When two models give the same value of MAD, then the choice should go in favour of the simpler model.

The Choice of Forecasting Model

As there is a wide variety of forecasting methods, one is not sure as to which method should be adopted. The question of choosing the right method in a given situation assumes considerable importance. This should be clear by a couple of examples.

Suppose that a very optimistic forecast for sales has been made by a manufacturing company for its product. On this basis, it has expanded its manufacturing operations and produced a larger quantity of output than in the earlier years. However, as the time goes by, it realises that sales have been more or less at the previous level, resulting in a heavy loss on account of large stock of unsold product with the company. Apart from this, loss has resulted on account of higher inventory the company has maintained in order to produce a larger quantity of output.

We can take another example in contrast to the earlier one. A sale forecast for a particular product has been made. But it turns out that the forecast is not realistic—it is on a much lower side than the demand for that product. In view of a low forecast, the company has missed a very good opportunity of augmenting its sales and profits.

One can visualise several situations in business where either high or low forecasts would go against the interest of the company. This closely establishes a major point that efforts must be made to make the forecasts as realistic as possible.

In order to select a specific method of forecasting, it is necessary first to compare the requirements of the forecast with the capabilities of the proposed method. Generally, there are three requirements of any forecast. These are: (i) extent of accuracy desired, (ii) data required, and (iii) extent of time available. If accuracy, say, within ± 5 per cent is required, then methods that are judged to yield less than ± 5 per cent accuracy in forecast need not be considered. However, if no other method is expected to

give better accuracy than that, then one has to use any of these methods. If the *data* needed to use the proposed method are not available (as, for example, time series data in forecasting sales of new products), then obviously one has to consider some other method. In the same manner, if *time available* is too short, an elaborate method requiring considerable time to complete the forecast cannot be used. In such a situation, one has to compromise and use a less sophisticated and a less rigorous method.

When we apply these three screening criteria before selecting a particular method, we may find that a sizeable number of potential forecasting methods get eliminated. The management has to choose from amongst the remaining methods. The choice will normally depend on cost-benefit analysis where greater accuracy is weighed against added costs.

A point worth emphasizing is that whatever forecasting model is selected, it should be closely monitored. When a new data value (Y_i) becomes available, it must be compared with the forecasted value ($\hat{Y}_i$). In case the difference between Y_i and $\hat{Y}_i$ is considerable, one must re-evaluate the forecasting model. This re-evaluation may indicate that the forecasting model is not appropriate and as such it should be revised.

Additional Examples

Example 19.11 Calculate the seasonal index number from the following data:

Year	Q_1	Q_2	Q_3	Q_4
1987	105	100	90	115
1988	110	105	95	115
1989	100	95	95	105
1990	115	110	100	125

If the annual sales for 1991 are expected to be Rs 2,000 lakh, what are the likely sales for the individual quarters?

Solution

Worksheet

Year	*Quarter* (1)	*Original data* (2)	*4-quarter moving total* (3)	*2-figure moving total* (4)	*Moving average (4)/8* (5)	*Original data as % of M.A* (6)
1987	1	105				
	2	100				
	3	90	410	825	103	87
	4	115	415	835	104	111
1988	1	110	420	845	106	104
	2	105	425	850	106	99
	3	95	425	840	105	90
	4	115	415	820	103	112
1989	1	100	405	810	101	99
	2	95	405	800	100	95
	3	95	395	805	101	94

(*Contd.*)

(*Contd.*)

	4	105	410	835	104	101
1990	1	115	425	855	107	107
	2	110	430	880	110	100
	3	100	450			
	4	125				

Calculation of Seasonal Index

Year	Q_1	Q_2	Q_3	Q_4	
1987	–	–	87	111	
1988	104	99	90	112	
1989	99	95	94	101	
1990	107	100	–	–	
Total	**310**	**294**	**271**	**324**	**Total**
Average	103	98	90	108	**399**
S. Index	103	98	90	109	**400**

Likely Sales for 1991 (Quarterwise)

Rs Lakhs	Q_1	Q_2	Q_3	Q_4	
	515	490	450	545	2000

Example 19.12 Fit a straight line trend by the method of least square to the following data :

Year	1991	1992	1993	1994	1995	1996
Production (in tonnes)	24	25	29	26	22	24

Estimate the likely production for the year 1998.

Solution

Year	*Production (tonnes)*			
	X	Y	XY	X^2
1991	–5	24	–120	25
1992	–3	25	–75	9
1993	–1	29	–29	1
1994	1	26	26	1
1995	3	22	66	9
1996	5	24	120	25
	0	150	–22	70

The two normal equations are:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the above values in the normal equation, we get

$$150 = 6a + b\,(0) \quad (1)$$

$$-22 = a\,(0) + 70b \quad (2)$$

From equation (1), we get $a = \frac{150}{6} = 25$ and

From equation, (2) we get $b = \frac{-22}{70} = -0.314$

Hence, the straight line trend is

$$\hat{Y} = 25 - 0.314X$$

X : Half-year units

Origin 1.1.1994

Now, we have to estimate the likely production for the year 1998.

The year 1998 as on 1.1.1998 will be 9, we have to make it 10 as we are interested in production at the end of the year, which means 31.12.1998.

Thus, $Y = 25 - (0.314)\ (10)$

or $Y = 25 - 3.14 = 21.86$

Example 19.13 The trend equation for yearly sales of a commodity with 1st July 1991 as origin is $Y = 96 + 28.8\ X$.

(i) Determine the monthly trend equation with January 1992 as origin.

(ii) Compute the trend values for August 1991 and March 1992.

Solution

$$Y = 96 + 28.8\ X$$

Origin 1st July 1991

X : 1-year unit

Now, we have to determine the monthly trend equation with January 1992 as origin. This can be done by finding the new Y-intercept a Jan. 1992. This is: a Jan 2002 $= 96 + (28.8 \times \frac{1}{2}) = 96 + 14.4 = 110.4$

Thus, $\hat{Y} = 110.4 + 28.8\ X$

Origin 1st January 1992

X : One-year unit.

The monthly equation:

$$\hat{Y} = (110.4/12) + [28.8/(12 \times 12)]\ X$$
$$= 9.2 + 0.2\ X$$

Origin January 1, 1992.

where X is in 1 month units

(b) August 1991

We go back by $4\frac{1}{2}$ months as we have to determine the trend value in mid-August.

Hence, $\hat{Y} = 9.2 + (0.2)\ (-4.5)$

or $\hat{Y} = 9.2 - 0.9 = 8.3$

Similarly, for March 1992, we have to go forward by $2\frac{1}{2}$ months.

Hence, $\hat{Y} = 9.2 + (0.2 \times 2.5)\ X$

$= 9.2 + 0.5 = 9.7$

Example 19.14 The linear trend of sales of a company is Rs 6,50,000 in 1995 and it rises by Rs 16,500 per year.

(i) Write down the trend equation.

(ii) If the company knows that its sales in 2002 will be 10 per cent below the forecasted trend sales, find its expected sales in 2002.

Solution Since sales of the company amounted to

(i) Rs 6,50,000 in 1995 with an annual increase of Rs 16,500, we have to subtract this amount from Rs 6,50,000 to arrive its trend equation.

Hence, Y = Rs 6,50,000 – Rs 16,500 = Rs 6,33,500

$\therefore$ Y = Rs 6,33,500 + Rs 16,500 X

(ii) Rs 6,33,500 + (Rs 16,500 × 8)

Forecasted trend in 2002 is Rs 7,65,500

Less 10% of its 76,550

Rs 7,65,500 – Rs 76,550 = Rs 6,88,950

Example 19.15 The trend equation for annual sales of a product is—

$$Y = 102 + 36X$$

with 1st January 1990 as origin.

(i) Determine the monthly trend equation with 1st July 1992 as origin.

(ii) Compute the trend values of sales in October 1994.

Solution

$$Y = 102 + 36X$$

Origin 1.1.1990

X : 1-year units

$$\hat{Y} = 102 + (36 \times 2)$$
$$= 102 + 72$$
$$= 174\text{—origin 1.1.1992}$$

Hence, with new intercept in trend line is

$$\hat{Y} = 174 + 36X$$

Origin 1.1.1992

X : 1-year units

Now the origin is to be shifted to 1.7.1992

i.e. half-year from 1.1.1992

This means X is $\frac{1}{2}$.

The new intercept will be

$$\hat{Y} = 174 + 36 \times \frac{1}{2}$$
$$= 174 + 18$$
$$= 192$$
$$\hat{Y} = 192 + 36X$$

Origin 1.7.1992

X = 1-year unit

Now we have to convert this annual trend into monthly trend line as follows.

$$\hat{Y} = \frac{192}{12} + \frac{36}{12 \times 12} X$$

$\hat{Y} = 16 + 0.25\,X$ X is 0 at July 1, 1992

For October 1994 $\hat{Y} = 16 + (0.25)\,(27)$

$= 16 + 6.75$

$= 22.75$

Example 19.16 The sales of a company (in lakh of Rs) for the seven years are given below:

Year	1990	1991	1992	1993	1994	1995	1996
Sales	32	47	65	88	132	190	275

Find out the trend values by using the equation, $Y = ab^X$.

Solution

Year X	*Sales Y*	*Log y*	*Time centred x*	x^2	*xY*
1990	32	1.5051	–3	9	–4.5153
1991	47	1.6721	–2	4	–3.3442
1992	65	1.8129	–1	1	–1.8129
1993	88	1.9445	0	0	0
1994	132	2.1206	1	1	2.1206
1995	190	2.2788	2	4	4.5576
1996	275	2.4393	3	9	7.3179
	829	13.7733	0	28	4.3237

$$\text{Log } a = \frac{\Sigma \log Y}{n} = \frac{13.7733}{7} = 1.9676$$

$$\text{Log } b = \frac{\Sigma X \cdot \log Y}{\Sigma x^2} = \frac{4.3237}{28} = 0.1544$$

Trend of logs : $\hat{Y} = a + bX$

$= 1.9676 + 0.1544X$

or $\text{Log } \hat{Y} = \log a + X \log b$

Trend of data : $\hat{Y} = ab^X$

$\hat{Y} = \text{Antilog } 1.9676\ (\text{Antilog } 0.1544)^X$

or $\hat{Y} = (92.81)\ (1.427)^X$

Example 19.17 Eliminate trend by 'moving average method'.

Year	*I Q*	*II Q*	*III Q*	*IV Q*
1995	40	35	38	40
1996	42	37	39	38
1997	41	36	38	42

Solution

Year (1)	Cd (2)	(3)	4-Q Moving Total (4)	2-Q Moving Total (5)	M. Average (5)/8 (6)	Trend free series (3)/(6) (7)
1995	1	40				
	2	35				
	3	38	153	308	38.5	99
	4	40	155	312	39	102
1996	1	42	157	315	39.4	107
	2	37	158	314	39.3	94
	3	39	156	311	38.9	100
	4	38	155	309	38.6	98
1997	1	41	154	307	38.4	107
	2	36	153	310	38.8	93
	3	38	157			
	4	42				

Example 19.18 Fit a simple exponential trend to the following data, and estimate the sales for 1989.

Year	1985	1986	1987	1988	1989
Sales (Rs crore)	100	105	112	120	130

Solution

Year X	Sales Y	Log Y	Time centred x	x^2	x log y
1985	100	2.0000	–2	4	–4.0000
1986	105	2.0212	–1	1	–2.0212
1987	112	2.0492	0	0	0
1988	120	2.0792	1	1	2.0792
1989	130	2.1139	2	4	4.2278
	567	10.2635	0	10	0.2858

$$a = \frac{\Sigma Y}{n} = \frac{10.2635}{5} = 2.0527$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{0.2858}{10} = 0.02858 \text{ or } 0.0286 \text{ (upto 4 places)}$$

Trend of logs : $\hat{Y} = a + bX$

$= 2.0527 + 0.0286X$

or $\log \hat{Y} = \log a + \log b\,(X)$

Trend of data = $\hat{Y} = ab^X$

$=$ Antilog 2.0527 (Antilog 0.0286)X

or $112.9016\ (1.068)^x$

or $112.9\ (1.07)^x$, Origin 1.7.1987 X-unit 1 year

Estimate of sale for 1989:
Here X is 2

$= 112.9\ (1.07)^x$
or $112.9\ (1.07)^2$
or 112.9×1.1449
or 129.25921
or 129 approx.

Example 19.19 Fit an appropriate trend line to the following data:

Year (X)	*Population in (millions)(Y)*
1998	184
1999	187
2000	189
2001	192
2002	195
2003	197
2004	199
2005	201
2006	203
2007	205
2008	208

Solution

Worksheet

X	Y	*Log Y*	*Time centred X*	X^2	XY	*X log Y*
1998	184	2.2648	–5	25	–920	–11.3240
1999	187	2.2718	–4	16	–748	–9.0872
2000	189	2.2765	–3	9	–567	–6.8295
2001	192	2.2833	–2	4	–384	–4.5666
2002	195	2.2000	–1	1	–195	–2.2000
2003	197	2.2945	0	0	0	0
2004	199	2.2989	1	1	199	2.2989
2005	201	2.3032	2	4	402	4.6064
2006	203	2.3075	3	9	609	6.9225
2007	205	2.3118	4	16	820	9.2472
2008	208	2.3181	5	25	1040	11.5905
	2160	**25.1304**	**0**	**55**		**0.6582**

$$a = \frac{\Sigma Y}{n} = \frac{2100}{11} = 190.909$$

$$b = \frac{\Sigma X \log Y}{\Sigma X^2} = \frac{0.6582}{55} = 0.01196$$

Trend of logs :

$$\hat{Y} = a + bX$$
$$= 190.909 + 0.01196X$$

or $$\log \hat{Y} = \log a + \log b(X)$$

Trend of data $= \hat{Y} = ab^x$

$$= \text{Antilog } 190.909 \text{ (Antilog } 0.01196)$$

or $$= 811.0\ (1.046)^x$$

$$\hat{Y} = 811.0\ (1.046)^x$$

Example 19.20 Fit a straight line trend by the method of least squares to the following data. Assuming the same rate of change continues, forecast the sales for 2008.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Sales (Rs lakhs)	76	80	330	144	138	120	174	190

Calculate the trend values from 1997 to 2004.

Solution

Worksheet

Year	X	Y	XY	X^2
1997	–7	76	–532	49
1998	–5	80	–400	25
1999	–3	330	–990	9
2000	–1	144	–144	1
2001	1	138	138	1
2002	3	120	360	9
2003	5	174	870	25
2004	7	190	1330	49
	0	**1252**	**632**	**168**

The two normal equations are

$$\Sigma Y = na + b\,\Sigma X$$
$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2$$

Substituting the values in the two equations, we get

$$1252 = 8a + b(0) \quad (1)$$
$$632 = a\,(0) + b\,(168) \quad (2)$$

The first equation gives $a = 1252/8 = 156.5$ and

the second equation gives $b = 3.76$.

Hence, the trend line is

$$\hat{Y} = 156.5 + 3.76X$$

Origin 1.1.2001

X : half-year units

We are required to forecast for 2008. Since X is in half-year units, for 2004 X is 7, for 2005 X is 9, … and for 2008 X is 15. Hence, the forecast is

$$\hat{Y} = 156.5 + (3.76 \times 15)$$
$$= 156.5 + 56.4$$
$$= 212.9 \text{ (Rs. 212.9 lakhs)}$$

Example 19.21 The prices of a certain commodity during the period were as follows:

Year	*Jan-March*	*April-June*	*July-September*	*Oct-Dec.*
1980	321	348	348	348
1981	327	354	354	348
1982	342	350	381	345
1983	364	390	401	385

Compute the seasonal indices by the method of simple averages, and find the deseasonalised values.

Solution

We have to first find the average of each quarter. In other words, each quarter's total is to be divided by 4.

Quarter	*Quarterly Totals*	*Quarterly Averages*	*Seasonal Index*	*Adjusted Seasonal Index*
1	1354	338.5	23.7	94.8
2	1442	360.5	25.3	101.2
3	1484	371.0	26.0	104.0
4	1426	356.5	25.0	100.0
		1426.5	**100.0**	**400**

Since the seasonal index for the year should add to 400, we have to calculate the adjusted index for each quarter, by first taking 1426.5 = 100. Thus, we calculate the proportion for each quarter, which is to be multiplied by 4 to obtain, finally, the seasonal index.

Deseasonalised series has been arrived at by dividing the original data for each quarter by the corresponding seasonal index and then multiplying it by 100.

Table

Year		*Original data*	*Adjusted Seasonal Index*	*Deseasonalised series*
1980	Q_1	321	94.8	338.61
	Q_2	348	101.2	343.87
	Q_3	348	104.0	334.62
	Q_4	348	100.0	348.00
1981	Q_1	327	94.8	344.94
	Q_2	354	101.2	349.80

(*Contd.*)

(*Contd.*)

	Q_3	354	104.0	340.38
	Q_4	348	100.0	348.00
1982	Q_1	342	94.8	360.76
	Q_2	350	101.2	345.85
	Q_3	381	104.0	366.35
	Q_4	345	100.0	345.00
1983	Q_1	364	94.8	383.97
	Q_2	390	101.2	385.38
	Q_3	401	104.0	385.58
	Q_4	385	100.0	385.00

Example 19.22 The following data relate to an annual trend

$$\hat{Y} = 460 + 27.8X$$

Origin : July 1, 2007

X in terms of years

Y in terms of Rupees in million

You are asked to convert this annual trend equation into monthly terms.

Solution

$$\hat{Y} = 462 + 27.8X$$

Since, this is an annual trend equation, dividing these values by 12, we get

$$a = 462/12 = 38.5$$

$$b = 27.8/12 = 2.3167$$

Again, $b = 2.3167/12 = 0.193$

As it is monthly trend equation, the origin should be shifted from July 1, 2007 to July 15, 2007, i.e., to the middle of the month. Hence,

$$\hat{Y} = 38.5 + (0.193)(0.5)$$

or $38.5 + 0.0965$

Thus, we can now write,

$$\hat{Y} = 38.5 + 0.0965X$$

Origin : July 15, 2007

X in terms of months

Y in terms of Rupees in million

GLOSSARY

Cyclical component or fluctuations	In a time series, fluctuations around the trend line that last for more than one year.
Deseasonalisation	A statistical process by which the seasonal variation from a time series is eliminated.
Forecasting	Predicting the expected value of an item or variable of interest.

Irregular component	Variation in a time series that occurs due to chance.
Ratio-to-moving average method	A statistical technique used to measure seasonal fluctuations.
Relative cyclical residual	A measure of cyclical variation wherein the percentage deviation from the trend for each observation in the series is used.
Residual method	A method of describing the cyclical component of a time series. It is based on the assumption that most of the variation in a time series not explained by the secular trend is cyclical variation.
Seasonal index	A measure used to adjust monthly or quarterly data for seasonal fluctuations.
Seasonal variation	One of the components in a time series that occurs during different seasons or periods of less than one year duration.
Second-degree equation	An equation in time series analysis to describe a non-linear trend.
Secular trend or trend component	One of the components in a time series indicating the long-term movement of an item or variable.
Time series	Data arranged in relation to time. Such data have four components—trend, cycle, seasonal and irregular movements.

LIST OF FORMULAE

1. Multiplicative time series model:

$$Y = T \times C \times S \times I$$

where T, C, S and I are the trend, cyclical, seasonal and irregular components, respectively and Y is the value of the time series at a given time.

2. Additive time series model:

$$Y = T + C + S + I$$

The symbols connote the same components as mentioned in formula 1 above.

3. Regression model for linear trend:

$$Y = a + bX$$

Y is the value of the time series, X is the time period, a is the constant term and b is the slope of the trend line.

4. When the individual years (X) are changed into coded time values such that $\Sigma X = 0$, then

$$a = \frac{\Sigma Y}{n} \quad \text{and} \quad b = \frac{\Sigma XY}{\Sigma X^2}$$

5. Computed or estimated linear trend line:

$$\hat{Y} = a + bX$$

where $\hat{Y}$ is the computed or estimated value of a time series.

6. Multiplicative time series for annual data (i.e. no seasonal variation is involved):

$$Y = T \times C \times I$$

7. Seasonal index for a period:

$$= \frac{\text{Mean for the period}}{\text{Sum of means for all periods}} \times \text{Number of periods}$$

8. Deseasonalised value for a period:

$$= \frac{\text{Actual value for that period}}{\text{Seasonal index for that period}}$$

9. Parabolic trend:

$$Y = a + bX + cX^2$$

It is the second-degree curve as a new term cX^2 has been added to the linear trend, $Y = a + bX$.

10. Relative cyclical residual:

$$= \frac{Y - \hat{Y}}{\hat{Y}} \times 100$$

where Y = actual time series value,

$\hat{Y}$ = estimated trend value from the same point in the time series

11. Irregular variation

$$I = \frac{CI}{C}$$

where I = irregular variation

CI = cyclical and irregular variation

C = cyclical variation

12. Exponentional Smoothing:

$E_i = WY_i + 1\ (1 - W)\ E_{i-1}$

where E_i and E_{i-1} are values of the smoothed series in period i and $i - 1$, respectively

W = weighting factor

Y_i = observed value in period i.

13. Autoregressive Models:

$\hat{Y}_i = b_0 + b_1\ Y_{i-1}$ (1st order) – similar to simple regression

$\hat{Y}_i = b_0 + b_1\ Y_{i-1} + b_2 Y_{i-2}$ (2nd order) – similar to multiple regression

14. Measurement of forecast error

$$\text{MAD} = \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|}{n}$$

where MAD = Mean absolute deviation

Y_i = Actual value of an item

$\hat{Y}_i$ = Fitted value of an item

$||$ = Absolute value

QUESTIONS

19.1 Given below are twelve statements. Indicate in each case whether it is true or false.

(a) There are three components in a time series—trend, cyclical and seasonal.

(b) Secular trends indicate the long-term direction of a time series.

(c) Among the components of time series, cyclical is the most difficult to compute.

(d) Time series analysis is unable to help us in future uncertainties.
(e) If the time series consists of annual data, then the seasonal factor is not at all important.
(f) There are two approaches to the time series analysis, viz. additive and multiplicative.
(g) Time series analysis is confined to linear trends alone.
(h) In order to measure cyclical variations, it is necessary that the deseasonalised monthly data are adjusted for trend.
(i) There are only two methods of computing seasonal index.
(j) The seasonal indices in a monthly-series must always add to 1,200 for the whole year.
(k) In forecasting, subjective methods are equally good as the statistical methods.
(l) When the time period of a forecast is too long, a second degree equation would give better results than a linear equation.

Multiple Choice Questions (19.2 to 19.16)

19.2 Which of the following components a time series having annual data contains?
(a) Secular trend (b) Cyclical fluctuation
(c) Seasonal variation (d) All of these
(e) (a) and (b) but not (c)

19.3 Suppose that a time series analysis with annual data for 1996-2004 gives the equation $\hat{Y} = 20 + 6x + 8x^2$. On the basis of this equation, what is the forecast for 2005?
(a) 200 (b) 250 (c) 230 (d) 220

19.4 Suppose you have a time series data on quarterly basis for 2 years 2003 and 2004. The third quarter of 2004 would be coded as
(a) 3 (b) 5 (c) 6 (d) 7

19.5 When a time series has an even number of years which are coded, then each coded interval is of
(a) one month (b) one year (c) two years (d) six months

19.6 Which one of the following is *not* a quantitative method of forecasting?
(a) Exponential smoothing (b) Autoregressive model
(c) Input-output analysis (d) The Delphi method

19.7 Which of the following components is used for a short-term forecast?
(a) cyclical (b) trend (c) seasonal (d) none of these

19.8 After detrending, a multiplicative model of time series is represented as
(a) Y = TSI (b) Y = TSCI (c) Y = TCI
(d) Y = SCI (e) none of these

19.9 Which of the following methods is used for measuring seasonal variation?
(a) The method of simple averages (b) The moving average method
(c) The ratio to trend method (d) The link relative method
(e) All of these

19.10 Suppose a seasonal index for a particular month is greater than 100, then the following must be true.
(a) The seasonal index for some other month > 100.
(b) The seasonal index for some other month = 100
(c) The seasonal index for some other month < 100
(d) (b) and (c)
(e) none of the above

19.11 In time series analysis both trends and seasonal variations are studied because they
(a) describe past patterns
(b) allow projections into the future
(c) allow the elimination of the component from the series
(d) all of the above

19.12 The standard error of estimate in a forecast using regression analysis, shows
(a) time period for which the forecast is valid
(b) maximum error of the forecast
(c) overall accuracy of the forecast
(d) all of the above

19.13 Which of the following methods is used for eliminating C from TSCI
(a) Spearman analysis (b) second-degree analysis
(c) relative cyclical residual (d) none of the above

19.14 A long range forecast implies that the forecast is for the time period
(a) At least 1 year (b) Between 1 year and 3 years
(c) Broadly 5 years and more (d) Between 3 and 5 years

19.15 Before attempting any time series forecast, one must consider:
(a) extent of accuracy desired (b) data required
(c) availability of time (d) (a) and (b)
(e) (a), (b) and (c)

19.16 Which of the following methods is not a subjective method of forecast?
(a) Users' expectations (b) Jury of executives
(c) Box-Jenkins model (d) The Delphi method

19.17 What are the advantages of undertaking a time series analysis to a business firm?

19.18 Briefly explain the components of a time series.

19.19 Which of the four components of a time series you would use in the following cases:
(a) The effect of Diwali sales of textiles on a large retail outlet of readymade garments
(b) The effect of war
(c) A decline in the sale of ice-cream during the months of December to February
(d) Increasing house construction activity during the past five years
(e) Recession
(f) The decline in the death rate on account of improvement in medical facilities?

19.20 What are the limitations of time series analysis in forecasting?

19.21 What is a seasonal variation? What steps are involved in calculating a seasonal index?

19.22 One of the components in time series analysis is 'irregular variation'. Do you think that it is important? How would you measure it?

19.23 How would you isolate seasonal variation in a time series?

19.24 What are the limitations of a time series forecasting? What precautions one should take while attempting a forecast from the time series?

19.25 Explain how you would deseasonalise a time series stating the assumptions you would be making.

19.26 What are the commonly used models in a time series analysis? Discuss the underlying assumptions of each model.

19.27 What are the objectives of time series analysis? Why do we need to separate out fluctuations? Explain.

19.28 Explain briefly the additive and multiplicative models of time series.

19.29 Why is forecasting necessary for a business firm?

19.30 What are the steps involved in a forecasting process?

19.31 What are the causal methods of forecasting? Are they superior to other methods? Why or why not?

19.32 What are the 'exponential smoothing' and 'autoregression' measures in forecasting?

19.33 How would you measure errors in forecasting?

19.34 What are the assumptions involved in a forecasting model based on regression analysis?

19.35 The data given below were collected from the records of a car manufacturing company. The company sold cars from 1980 to 1987. Fit a straight line trend to the data. Estimate the production in 1987 and compare with the actual production.

Year	1980	1981	1982	1983	1984	1985	1986	1987
No. of cars sold ('000)	7.7	6.3	5.3	7.1	6.6	8.9	6.8	7.2

19.36 The following data refer to annual profits of a certain business.

Year	1991	1992	1993	1994	1995	1996	1997
Profits Rs '000	60	72	75	65	80	85	95

Find the trend values using the linear trend method. Using the trend equation, estimate the profit for 2001.

19.37 Find the 5-yearly weighted moving averages with weights 1,2,2,2,1 respectively for measuring trend of the following time series :

Year	1974	1975	1976	1977	1978	1979	1980	1981	1982
Sales (*Rs* '00000)	2	6	1	5	3	7	2	6	4

19.38 The figures of quarterly income of municipal corporation (in Rs lakh) for 2 years are given below:

Year	Q_1	Q_2	Q_3	Q_4
1995	74	56	48	69
1996	83	52	49	81

Using a four-quarterly moving average, estimate the trend values.

19.39 Assuming an additive model, apply 3-year moving averages to obtain the trend-free series for years 2 to 6 from the following data.

Year	1	2	3	4	5	6	7
Exports (*Rs lakh*)	126	130	137	141	145	155	159

19.40 Given the following ratio of observed data to moving averages, compute the seasonal indices:

Year	*QI*	*QII*	*QIII*	*QIV*
1	–	–	87	92
2	105	100	90	93
3	110	105	93	100
4	115	110	–	–

19.41 The following is a monthly trend equation:

$\hat{Y} = 20 + 2X$

[Origin January 1992, X-unit = one month, Y-unit = monthly sales (in Rs '000)].

Convert it into an annual trend equation.

19.42 Use a four-yearly moving average method to calculate trend for the following data :

Year	*Production*
1978	614
1979	615
1980	652
1981	678
1982	681
1983	655
1984	717
1985	719
1986	708
1987	779
1988	757

19.43 Given below are the figures of production (in million tonnes) of wheat :

Year	1989	1990	1991	1992	1993	1994	1995
Production	80	90	92	83	94	99	92

Fit a straight-line trend to these figures.

19.44 What is meant by seasonal fluctuations? Compute the seasonal index for the following data by using the method of moving averages:

Quarter/Year	1970	1971	1972	1973
January – March	75	86	90	100
April – June	60	65	72	78
July – September	54	63	66	72
October – December	59	80	85	93

19.45 **(a)** What is a time series? Mention its important components. Explain them briefly.

(b) The following data give the total expenditure incurred by a certain college during the respective years:

Year	1967	1968	1969	1970	1971	1972
Expenditure (Rs lakh)	1.5	1.8	2.0	2.3	2.4	2.6

Estimate the expenditure that might have been incurred by the above college during the years 1973 and 1974.

19.46 You are given the annual profit figures for a certain firm for the years 1970–76. Fit a straight line trend to the data and estimate the expected profit for the year 1977.

Year	1970	1971	1972	1973	1974	1975	1976
Profit (Rs lakh)	60	72	75	95	80	85	95

19.47 Fit a second degree parabola to the following data and forecast the sales for 2003. Comment on the suitability of the model.

Year	1995	1996	1997	1998	1999
Sales (in Rs lakh)	10	12	13	10	8

19.48 The quarterly sales data for a graphics software company are given below. Determine the seasonal components.

	Sales in Rs '000					
Quarter	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
1	500	450	350	550	550	750
2	350	350	200	350	400	500
3	250	200	150	250	350	400
4	400	300	400	550	600	650

19.49 Construct a 4-yearly centred moving average from the following data:

Year	1991	1992	1993	1994	1995	1996	1997
Sale of computer (100 units)	129	131	106	91	95	14	90

Also determine the short-term fluctuations.

19.50 Fit a second degree parabola to the following data. Estimate the value for the year 2000.

Year	1950	1960	1970	1980	1990
Production ('000 units)	6	8	9	10	12

19.51 Assuming that the trend is absent, calculate the seasonal indices for various quarters from the following data.

Year	*I Quarter*	*II Quarter*	*III Quarter*	*IV Quarter*
1996	37	41	33	35
1997	37	39	36	36
1998	40	41	33	31
1999	33	44	40	40

19.52 The following table gives the total expenditure of the government during the period 1978-1985. Fit a quadratic trend to the data.

Year	1978–79	79–80	80–81	81–82	82–83	83–84	84–85
Expenditure	177.2	185.0	224.9	254.0	304.9	359.9	438.8

19.53 From the time series data below, using
(i) last period demand, **(ii)** arithmetic average, and **(iii)** two-month moving average technique, find the forecast for period 7, and choose the best technique, using mean absolute deviation (MAD).

Month	1	2	3	4	5	6
Demand	20	30	40	30	50	58

19.54 The prices of a commodity during 1993–1998 are given below:

Year	1993	1994	1995	1996	1997	1998
Price (Rs)	100	107	128	140	181	192

Fit a parabola to the above data and estimate the price of the commodity for the year 2000.

The McGraw-Hill Companies

CHAPTER 20 NONPARAMETRIC TESTS

Learning Objectives

By the end of your work on this chapter, you should be able to

- differentiate between parametric and nonparametric tests
- understand the relevance of nonparametric tests in data analysis
- understand the procedure involved in carrying out nonparametric tests
- design and conduct some selected nonparametric tests.

Chapter Prerequisites

Before starting work on this chapter, make sure you have fully grasped the material given in Chapter 13 on Testing Hypotheses.

20.1 INTRODUCTION

Hypothesis tests can be divided into two categories: *parametric tests* and *nonparametric tests*. In Chapter 13, we discussed hypothesis test relating to the former category. Parametric tests assume that parameters such as mean, standard deviation, and so forth, exist and these are used in testing a hypothesis. These tests assume that the form of the population distribution is known and that a test concerning a parameter of a distribution is to be made. However, there are many situations where one or more assumptions that are made in the case of parametric tests cannot be met. In such cases, statisticians have developed some other techniques that are based on less stringent assumptions. These include nonparametric methods as well as distribution-free methods. When we are not concerned with the parameters of a given population, then the nonparametric methods are applied. As regards distribution-free methods, we do not make any assumptions about the population from which we are sampling. It may be noted that as this distinction between nonparametric methods and distribution-free methods is rather fine, in practice, both the methods are referred to as nonparametric methods.

Main Advantages of Nonparametric Tests There are certain advantages of the nonparametric tests; as a result they have become more important and are being increasingly used in recent years. The main *advantages* of these tests are:

1. No assumptions are required or less stringent assumptions are required in nonparametric tests as compared to parametric tests.
2. Nonparametric tests are more suitable when ranked, scaled or rated data are to be analysed.
3. Nonparametric tests do not take much time as they involve very simple calculations.

As against these advantages of nonparametric tests, there are certain *disadvantages* of these tests.

Disadvantages of Nonparametric Tests Following are the disadvantages of these tests

1. They are based on limited amount of information and do not make use of all the available information. For example, a number of figures are just replaced by ranks 1, 2, 3, and so on. Obviously, a good deal of information is lost in this way.
2. Another disadvantage of such tests is that they are less powerful than the parametric tests. This means that there is a greater risk of accepting a false hypothesis. In other words, chances of committing the Type II error are considerable.
3. Another disadvantage of such tests is that the null hypothesis is somewhat loosely formulated. In view of this, when the null hypothesis is rejected, conclusions arising therefrom are less precise as compared to the parametric tests.

We now discuss some of the frequently used nonparametric tests.

20.2 SIGN TESTS

One of the easiest nonparametric tests is the sign test. The test is known as the sign test as it is based on the direction of the plus or minus signs of observations in a sample instead of their numerical values. There are two types of sign tests: **(a)** *One-sample sign test*, and **(b)** *Two-sample sign test*.

The One-sample Sign Test

The one-sample sign test is a very simple nonparametric test applicable on the assumption that we are dealing with a population having a continuous **symmetrical** distribution. As such, the probability of getting a value less than the mean is 0.5. Likewise, the probability of getting a value greater than the mean is also 0.5. To test the null hypothesis $\mu = \mu_0$ against an appropriate alternative, each sample value greater than μ_0 is replaced by plus (+) sign and each sample value less than μ_0 with a minus (–) sign. Having done this, we can test the null hypothesis that the probabilities of getting both plus and minus signs are 0.5. It may be noted that if a sample value happens to be equal to μ_0, it is simply discarded.

To perform the actual test, we use either of the two methods. When the sample is small, the test is performed by computing the binomial probabilities or by referring to the binomial probabilities table. When the sample is large, the normal distribution is used as an approximation of the binomial distribution. Let us take an example to show how the one-sample sign test is applied.

Example 20.1 We are required to test the hypothesis that the mean value μ of a continuous distribution is 20 against the alternative hypothesis $\mu \neq 20$. Fifteen observations were taken and the following results were obtained:

18, 19, 25, 21, 16, 15, 19, 22, 24, 21, 18, 17, 15, 26 and 24.

We may use $\alpha = 0.05$ level of significance.

Solution Replacing each value greater than 20 with a plus (+) sign and each value less than 20 with a minus (–) sign, we get

– – + + – – – + + + – – – + +

Now, the question before us is whether 7 plus signs observed in 15 trials support the null hypothesis $p = 0.5$ or the alternative hypothesis $p \neq 0.5$. Using the Appendix Table 3, we find that the probability of 7 or more successes is 0.196 + 0.196 + 0.153 + 0.092 + 0.042 + 0.014 + 0.003 = 0.696* and $p = 0.5$ and since this value is greater than $\alpha/2 = 0.025$, we find that the null hypothesis will have to be accepted. We can also use normal approximation to the binomial distribution when $np \geq 5$. As here $p = ½$, the condition for the normal approximation to the binomial distribution is satisfied as $n > 10$. As such, we can use the Z statistic for which the following formula is to be used.

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X - (np)}{\sqrt{\frac{n}{4}}}$$

$$= \frac{7 - (15/2)}{\sqrt{\frac{15}{4}}} = \frac{\frac{14 - 15}{2}}{1.9365}$$

$$= \frac{-0.5}{1.9365} = -0.26$$

Since calculated $Z = -0.26$ lies between $Z = -1.96$ and $Z = 1.96$ (the critical value of Z at 0.05 level of significance), the null hypothesis is accepted.

The Two-sample Sign Test

The sign test can be applied to problems that deal with paired data. In such problems, each pair can be replaced with a plus sign if the first value is greater than the second, or a minus sign if the first value is smaller than the second. In case the two values in the pair turn out to be equal, these are discarded. There are essentially two kinds of situations: **(a)** the data are actually given as pairs and **(b)** the data comprise two independent samples that are randomly paired.

Example 20.2 Suppose we have the following table indicating the ratings assigned to two brands of cold drink X and Y by 12 consumers. Each respondent was asked to taste the two brands of cold drink and then rate them.

Table 20.1 Ratings of Brands *X* and *Y* Cold Drinks

Brand X	26	30	44	23	18	50	34	16	25	49	37	20
Brand Y	22	27	39	7	11	56	30	14	18	51	33	16
Sign	+	+	+	+	+	–	+	+	+	–	+	+

We have to apply the two-sample sign test.

Solution Row three of Table 20.1 shows + and – signs. When X's rating is higher than that of Y, then the third row shows the '+' sign. As against this, when X's rating is lower than that of Y, then it shows

* It may be noted that the Appendix Table 3 gives figures up to four decimal. However, we have used these figures up to three decimal by rounding them where necessary.

the '–' sign. The table shows 10 plus signs and 2 minus signs. Now, we have to examine whether '10 successes in 12 trials' supports the null hypothesis $p = ½$ or the alternative hypothesis $p > ½$. The null hypothesis implies that both the brands enjoy equal preferences and none is better than the other. The alternative hypothesis is that the brand X is better than brand Y. Referring to the Appendix Table 3, we find that for $n = 12$ and $p = ½$ the probability of '10 or more successes' is 0.016 + 0.003 = 0.019. It follows that the null hypothesis can be rejected at $\alpha = 0.05$ level of significance. We can, therefore, conclude that brand X is a preferred brand as compared to brand Y.

Example 20.3 To illustrate the second case, which relates to two independent samples, let us consider the following data pertaining to the downtimes (periods in which computers were inoperative on account of failures, in minutes) of two different computers. We have to apply the two-sample sign test.

Computer A	58	60	42	62	65	59	60	52	50	75	59
	52	57	30	46	66	40	78	55	52	58	44
Computer B	32	48	50	41	45	40	43	43	70	60	80
	45	36	56	40	70	50	53	50	30	42	45

Solution These data are shown in Table 20.2 along with + or – sign as may be applicable in case of each pair of values. A plus sign is assigned when the downtime for computer A is greater than that for computer B and a minus sign is given when the downtime for computer B is greater than that for computer A.

Table 20.2 Downtime of Computers A and B (Minutes)

Computer A	58	60	42	62	65	59	60	52	50	75	59
Computer B	32	48	50	41	45	40	43	43	70	60	80
Sign	+	+	–	+	+	+	+	+	–	+	–
Computer A	52	57	30	46	66	40	78	55	52	58	44
Computer B	45	36	56	40	70	50	53	50	30	42	45
Sign	+	+	–	+	–	–	+	+	+	+	–

It will be seen that there are 15 plus signs and 7 minus signs. Thus, we have to ascertain whether '15 successes in 22 trials' support the null hypothesis $p = ½$. The null hypothesis implies that the true average downtime is the same for both the computers A and B. The alternative hypothesis is $p \neq ½$.

Let us use in this case the normal approximation of the binomial distribution. This can be done since np and $n(1 – p)$ are both equal to 11 in this example. Substituting $n = 22$ and $p = ½$ into the formulas for the mean and the standard deviation of the binomial distribution, we get $\mu = np = 22(½) = 11$ and

$$\sigma = \sqrt{np(1-p)} = \sqrt{22 \cdot ½ \cdot ½} = 2.345$$

Hence, $$Z = (X - \mu)/\sigma = (15 - 11)/2.345 = 1.71$$

Since this value of 1.71 falls between $-Z_{0.025} = -1.96$ and $Z_{0.025} = 1.96$, we find that the null hypothesis cannot be rejected. This means that the downtime in the two computers is the same.

This seems to be surprising as we find that there are substantial differences. The two sample means, for example, are 55.5 for A and 48.6 for B. This example illustrates the point that at times the sign test

can be quite a waste of information. It may also be noted that had the continuity correction been used, we would have obtained

$$Z = 3.5/2.345 = 1.49$$

This would not have changed our earlier conclusion.

20.3 THE TWO-SAMPLE AND K-SAMPLE MEDIAN TESTS

In order to perform this test, let us use our previous example, which pertains to the downtimes of the two computers. The median of the **combined data** is 52, which can easily be checked. There are 5 values below 52 and 15 values above it, in case of computer A. As regards computer B, the corresponding figures are 16 and 6. All this information is summarised in Table 20.3, which also indicates the totals of the rows and columns.

Table 20.3 Classification of Downtime for Computers *A* and *B*

	Below Median	*Above Median*	*Total*
Computer A	5	15	20
Computer B	16	6	22
Total	**21**	**21**	**42**

Our null hypothesis H_0 is that there is no difference in the median downtime for the two computers. The alternative hypothesis H_1 is that there is difference in the downtime of the two computers.

We now calculate the expected frequencies by the formula ($\text{Row}_i \times \text{Column}_i$)/Grand total. Thus, Table 20.4 shows both the observed and the expected frequencies. Of course, we could have obtained these results by arguing that half the values in each sample can be expected to fall above the median and the other half below it.

Table 20.4 Calculation of Chi-square

Observed Frequencies (O)	*Expected Frequencies (E)*	$O - E$	$(O - E)^2$	$(O - E)^2/E$
5	10	–5	25	2.50
15	10	5	25	2.50
16	11	5	25	2.27
6	11	–5	25	2.27
			Total	**9.54**

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = 9.54$$

The critical value of χ^2 at 0.05 level of significance for (2 – 1) (2 – 1) = 1 degree of freedom is 3.841 (Appendix Table 5). Since the calculated value of χ^2 exceeds the critical value, the null hypothesis has to be rejected. In other words, there is no evidence to suggest that the downtime is the same in case of the two computers.

It may be recalled that in the previous example having the same data, the null hypothesis could not be rejected. In contrast, we find here that the two-sample median test has led to the rejection of the null hypothesis. This may be construed as evidence that the median test is not quite so wasteful of the information as the sign test. However, in general, it is very difficult to make a meaningful comparison of the merits of two or more nonparametric tests, which can be used for the same purpose.

The K-sample Median Test The median test can easily be generalised so that it can be applied to K-samples. In accordance with the earlier procedure, first find the median of the combined data. We then determine how many of the values in each sample fall above or below the median. Finally, we analyse the resulting contingency table by the method of chi-square. Let us take an example.

Example 20.4 Suppose that we are given the following data relating to marks obtained by students in Statistics in the three different sections of a B.Com class in a certain college. The maximum marks were 100.

Section A	46	60	58	80	66	39	56	61	81	70
	75	48	43	64	57	59	87	50	73	62
Section B	60	55	82	70	46	63	88	69	61	43
	76	54	58	65	73	52				
Section C	74	67	37	80	72	92	19	52	70	40
	83	76	68	21	90	74	49	70	65	58

Test whether the differences among the three sample means are significant.

Solution In case of such problems, analysis of variance is ordinarily performed. However, here we find that the data for Section C have much more variability as compared to the data for the other two sections. In view of this, it would be wrong to assume that the three population standard deviations are the same. This means that the method of one-way analysis of variance cannot be used.

In order to perform a median test, we should first determine the median of the combined data. This comes out to 63.5, as can easily be checked. Then we count how many of the marks in each sample fall below or above the median. Thus, the results obtained are shown in Table 20.5.

Table 20.5 Worksheet for Calculating Chi-square

	Below Median	*Above Median*	*Total*
Section A	12	8	20
Section B	9	7	16
Section C	7	13	20
Total	28	28	56

On the basis of $\left(\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}\right)$, the expected frequencies for each cell can be worked out. Thus the corresponding expected frequencies for Section A are 10 and 10, for Section B are 8 and 8, and for Section C 10 and 10. We can now obtain the value of chi-square. These calculations are shown below:

$$\chi^2 = \frac{(12-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-8)^2}{8} + \frac{(7-8)^2}{8} + \frac{(7-10)^2}{10} + \frac{(13-10)^2}{10}$$
$$= 0.4 + 0.4 + 0.125 + 0.125 + 0.9 + 0.9 = 2.85$$

Now, we have to compare this value with the critical value of χ^2 at 5 per cent level of significance. This value is 5.991 for 2 ($K - 1 = 3 - 1$) degrees of freedom (Appendix Table 5). As the calculated value of χ^2 is less than the critical value, the null hypothesis that there are no differences in the average marks, cannot be rejected. Hence, we conclude that there is no significant difference in the true average (median) marks obtained by the students in Statistics test from the three sections.

20.4 WILCOXON MATCHED-PAIRS TEST (OR SIGNED RANK TEST)

Wilcoxon matched-pairs test is an important nonparametric test, which can be used in various situations in the context of two related samples such as a study where husband and wife are matched or when the output of two similar machines are compared. In such cases, we can determine both direction and magnitude of difference between matched values, using Wilcoxon matched-pairs test.

Procedure of Wilcoxon Matched-pairs Test The procedure involved in using this test is simple. To begin with, the difference (d) between each pair of values is obtained. These differences are assigned ranks from the smallest to the largest, ignoring signs. The actual signs of differences are then put to corresponding ranks and the test statistic T is calculated, which happens to be the smaller of the two sums, namely, the sum of the negative ranks and the sum of the positive ranks.

There may arise two types of situations while using this test. One situation may arise when the two values of some matched-pair(s) is/are equal as a result the difference (d) between the values is zero. In such a case, we do not consider the pair(s) in the calculations. The other situation may arise when we get the same difference (d) in two or more pairs. In such a case, ranks are assigned to such pairs by averaging their rank positions. For instance, if two pairs have rank score of 8, then each pair is assigned 8.5 rank [(8 + 9)/2 = 8.5] and the next largest pair is assigned the rank 10.

After omitting the number of tied pairs, if the given number or matched pairs is equal to or less than 25, then the table of critical value T is used for testing the null hypothesis. When the calculated value of T is equal to or smaller than the table (i.e. critical) value at a desired level of significance, then the null hypothesis is rejected. In case the number exceeds 25, the sampling distribution of T is taken as approximately normal with mean $\mu_T = n(n+1)/\mu$ and standard deviation

$$\sigma_T = \sqrt{n(n+1)(2n+1)/24}$$

where n is taken as the number of given matched pairs— number of tied pairs omitted, if any. In such a situation, the test Z statistic is worked out as follows:

$$Z = (T - \mu_r)/\sigma_r$$

Let us now take an example to illustrate the application of Wilcoxon matched-pairs test.

Example 20.5 The management of the Punjab National Bank wants to test the effectiveness of an advertising company that is intending to enhance the awareness of the bank's service features. It administered a questionnaire before the advertising campaign, designed to measure the awareness of services offered. After the advertising campaign, the bank administered the same questionnaire to the same group of people. Both the before and after advertising campaign scores are given in the following table.

Consumer Awareness of Bank Services Offered

Consumer	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Before ad campaign	82	81	89	74	68	80	77	66	77	75
After ad campaign	87	84	84	76	78	81	79	81	81	83

Using Wilcoxon matched-pairs test, test the hypothesis that there is no difference in consumer awareness of bank services offered after the advertising campaign.

Solution

Table 20.6 Application of Wilcoxon Matched-pairs Test

Consumer	*After Ad Campn.*	*Before Ad Campn.*	*Diff.* d_i	*Rank of* d_i	*Rank (–) Sign*	*Rank (+) Sign*
1	87	82	5	6.5		6.5
2	84	81	3	4		4
3	84	89	–5	6.5	–6.5	
4	76	74	2	2.5		2.5
5	78	68	10	9		9
6	81	80	1	1		1
7	79	77	2	2.5		2.5
8	81	66	15	10		10
9	81	77	4	5		5
10	83	75	8	8		8
				Total	**– 6.5**	**+ 48.5**

Null hypothesis H_0 : There is no difference in the consumer awareness of bank services after the ad campaign.

Alternative hypothesis H_1 : There is difference in the consumer awareness of bank services after the ad campaign.

Computed 'T' value is 6.5. The critical value of T for $n = 10$ at 5 per cent level of significance is 8 (Appendix Table 8). Since the computed T value is less than the critical T value, the null hypothesis is rejected. We can conclude that after the ad campaign there is difference in the consumer awareness of the bank's services. Our conclusion that there is some difference in the consumer awareness of the bank's services needs some explanation. Had there been no difference in the awareness before and after the ad campaigns, the sum of positive and negative ranks would have been almost equal. However, if the difference between the two series being compared is larger, then the value of T will tend to be smaller as it is defined as smaller of ranks. This is the case we find in this problem. It may be noted that with this test the calculated value of T must be smaller than the critical value in order to reject the null hypothesis.

20.5 RANK SUM TESTS

The Mann-Whitney *U* Test

Although there are a number of rank sum tests, we shall confine ourselves to just two such tests—the Mann-Whitney U test and the Kruskal-Wallis test. When only two populations are involved we shall

use the former test. When more than two populations are involved, we shall use the latter test. It may be pointed out that as these tests use ranking data rather than plus and minus signs, their use will definitely be less wasteful than the sign test.

One of the most common and best known distribution-free tests is the Mann-Whitney test for two independent samples. The logical basis of this test is particularly easy to understand. Suppose we have two independent treatment groups, with n_1 observations in Group 1 and n_2 observations in Group 2. Now, we assume that the population from which Group 1 scores have been sampled contained generally lower values than the population from which Group 2 scores were drawn. If we were to rank these scores disregarding the group to which they belong then the lower ranks would generally fall to Group 1 scores and the higher ranks would generally fall to Group 2 scores. Proceeding one step further, if we were to add together the ranks assigned to each group, the sum of the ranks in Group 1 would be expected to be considerably smaller than the sum of the ranks in Group 2. This would result in the rejection of the null hypothesis.

Let us now take another situation where the null hypothesis is true and the scores for the two groups are sampled from identical populations. If we were to rank all N scores regardless of the group, we would expect a mix of low and high ranks in each group. Thus, the sum of the ranks assigned to Group 1 would be broadly equal to the sum of the ranks assigned to Group 2.

The Mann-Whitney test is based on the logic just described, using the sum of the ranks in one of the groups as the test statistic. In case that sum turns out to be too small as compared to the other sum, the null hypothesis is rejected. The common practice is to take the sum of the ranks assigned to the smaller group, or if $n_1 = n_2$, the smaller of the two sums as the test statistic. This value is then compared with the critical value that can be obtained from the table of the Mann-Whitney statistic (W_s) to test the null hypothesis.

Let us take an example to illustrate the application of this test.

Example 20.6 The following data indicate the lifetime (in hours) of samples of two kinds of light bulbs in continuous use:

Brand A	603	625	641	622	585	593	660	600	633	580	615	648
Brand B	620	640	646	620	652	639	590	646	631	669	610	619

We are required to use the Mann-Whitney test to compare the lifetimes of brands A and B light bulbs.

Solution The first step for performing the Mann-Whitney test is to rank the given data *jointly* (as if they were one sample) in an increasing or decreasing order of magnitude. For our data, we thus obtain the following array where we use the letters A and B to denote whether the light bulb was from brand A or brand B.

Table 20.7 Ranking of Light Bulbs of Brands A and B

Sample Score	*Group*	*Rank*	*Sample Score*	*Group*	*Rank*
580	A	1	625	A	13
585	A	2	631	B	14
590	B	3	633	A	15

(*Contd.*)

(*Contd.*)

593	A	4	639	B	16
600	A	5	640	B	17
603	A	6	641	A	18
610	B	7	646	B	19.5
615	A	8	646	B	19.5
619	B	9	648	A	21
620	B	10.5	652	B	22
620	B	10.5	660	A	23
622	A	12	669	B	24

As both the samples come from identical populations, it is reasonable to assume that the means of the ranks assigned to the values of the two samples are more or less the same. As such, our null hypothesis is:

H_0 : Means of ranks assigned to the values in the two groups are the same.
H_1 : Means are not the same.

However, instead of using the means of the ranks, we shall use *rank sums* for which the following formula will be used.

$$U = n_1 n_2 + [n_1(n_1 + 1)]/2 - R_1$$

where n_1 and n_2 are the sample sizes of Group 1 and Group 2, respectively, and R_1 is the sum of the ranks assigned to the values of the first sample. In our example, we have $n_1 = 12$, $n_2 = 12$ and $R_1 = 1 + 2 + 4 + 5 + 6 + 8 + 12 + 13 + 15 + 18 + 21 + 23 = 128$. Substituting these values in the above formula,

$$\begin{aligned} U &= (12)\,(12) + [12\,(12 + 1)]/2 - 128 \\ &= 144 + 78 - 128 \\ &= 94 \end{aligned}$$

From Appendix Table 9 for n_1 and n_2, each equal to 12, and for 0.05 level of significance is 37. Since the critical value is smaller than the calculated value of 94, we accept the null hypothesis and conclude that there is no difference in the average lifetimes of the two brands of light bulbs.

The test statistic we have just applied is suitable when n_1 and n_2 are less than or equal to 25. For larger values of n_1 and/or n_2, we can make use of the fact that the distribution of W_s approaches a normal distribution as sample sizes increase. We can then use the Z test to test the hypothesis.

The Normal Approximation

When both n_1 and n_2 are more than 10, the sampling distribution of the U statistic can be approximated by the normal distribution. As our problem meets this requirement, we can also apply the normal approximation to this problem. For this, we have to use the Z statistic.

1. Mean = $\mu u = [(n_1 n_2/2] = [(12 \times 12)/2] = 72$
2. Standard error = $\sqrt{\dfrac{n_1 n_2\,(n_1 + n_2 + 1)}{12}}$

$$= \sqrt{\frac{12 \times 12\,(12 + 12 + 1)}{12}}$$

$$= \sqrt{300} = 17.3$$

3. (Statistic – Mean) / Standard deviation
= (94 – 72)/17.3 = 1.27

The critical value of Z at 0.05 level of significance is 1.96. Since the calculated value of $Z = 1.27$ is smaller than 1.96, the null hypothesis is accepted. This shows that there is no difference in average lifetimes of brands A and B bulbs. The Z test is more dependable as compared to the earlier one. Both the tests give the same result. It may be noted that Mann-Whitney test required fewer assumptions than the corresponding standard test. In fact, the only assumption required is that the populations from which samples have been drawn are continuous.

The Kruskal-Wallis Test

This test is a direct generalisation of the Mann-Whitney test to the case in which we have three or more independent groups. It tests the null hypothesis that all samples came from identical populations. As against this, the alternative hypothesis is that the means of the populations are not all equal.

To perform the Kruskal-Wallis test, we have to rank all scores without regard to groups to which they belong and then compute the sum of the ranks for each group. The sums are denoted by R_i. If the null hypothesis is true, we would expect the $R_i s$ to be more or less equal.

The formula used in this test is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1)$$

where H is test statistic, $n = n_1 + n_2 + \ldots + n_k$ is the total number of observations in all samples and R_i is the sum of ranks of all items in sample i.

If the null hypothesis is true and each sample is at least of size 5, the sampling distribution of this statistic can be approximated closely with a chi-square distribution with $K - 1$ degrees of freedom. Consequently, we can reject the null hypothesis at a level of significance if H exceeds χ^2_α for $K - 1$ degrees of freedom. If the size of one or more samples is too small to use this approximation, the test will have to be based on special tables.

Let us take an example to illustrate the application of the Kruskal-Wallis test.

Example 20.7 Suppose that three groups of salesmen (being employees of a company) underwent training. The method of training used was different for each group. When training was completed, the salesmen were given a test. The marks scored by them are shown below:

Training method A	75	83	68	85	90	61	
Training method B	62	70	67	82	80	87	64
Training method C	65	71	74	63	89		

We have to use the Kruskal-Wallis Test to find out whether there was difference in the effectiveness of the three training methods.

Solution First of all, we set up the null hypothesis that there was no difference in the effectiveness of the three training methods. The alternative hypothesis is: there was difference in the effectiveness of the three training methods. As a next step, we have to rank these data taking all the three groups as if they are one. We start with the highest marks as rank one and proceed in a descending order.

Table 20.8 Ranking of Marks in Three Groups

Marks	90	89	87	85	83	82	80	75	74
Group	A	C	B	A	A	B	B	A	C
Rank	1	2	3	4	5	6	7	8	9
Marks	71	70	68	67	65	64	63	62	61
Group	C	B	A	B	C	B	C	B	A
Rank	10	11	12	13	14	15	16	17	18

The observations in the first sample are assigned the ranks 1, 4, 5, 8, 12, and 18 so that $R_1 = 48$. Observations in the second sample are assigned the ranks 3, 6, 7, 11, 13, 15, and 17 so that $R_2 = 72$. Observations in the third sample are assigned the ranks 2, 9, 10, 14, and 16 so that $R_3 = 51$. Substituting the values obtained for R_1, R_2 and R_3 together with $n_1 = 6$, $n_2 = 7$ and $n_3 = 5$ in the formula for H, we get.

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1)$$

After calculating H, it will be evaluated against the chi-square distribution with $K - 1$ degrees of freedom.

$$H = \frac{12}{18(18+1)}\left(\frac{48^2}{6} + \frac{72^2}{7} + \frac{51^2}{5}\right) - 3(18+1)$$

$$= \frac{12}{342}\left(\frac{2{,}304}{6} + \frac{5{,}184}{7} + \frac{2{,}601}{5}\right) - 57$$

$$= 0.035\,(384 + 740.571 + 520.2) - 57$$

$$= (0.035 \times 1{,}644.771) - 57 = 57.57 - 57 = 0.57$$

The critical value of $\chi^2_{0.05}$ for $K - 1 = 3 - 1 = 2$ degrees of freedom is 5.991 (Appendix Table 5). Since the calculated value of $H = 0.57$ (which can be treated as a chi-square) is less than the critical value of χ^2, the null hypothesis cannot be rejected. In other words, we can conclude that there is no difference in the effectiveness of the three methods of training imparted to salesmen.

20.6 THE ONE-SAMPLE RUNS TEST

There are many situations in which it is difficult to decide whether our assumption of randomness holds good. In particular, this is true when we have no control over the collection of data. For example, if we are asked to forecast the volume of sales of a company for a particular month, there is hardly any choice for us except to use sales data from previous years coupled with some recent information concerning the economic outlook in general. The point to note is that none of this information can be regarded as a random sample as it was not obtained with the use of random numbers or similar schemes.

There are several methods that can be applied to determine the randomness of a sample. The technique, which we are going to use here is based on the *theory of runs*. A run is a succession of identical letters (or other kinds of symbols), which is followed as well as preceded by different letters or no letters at all. An example will enable us to understand it.

Example 20.8 Consider the following arrangement of patients M (male) and F (female) who have visited an OPD in a hospital during the first few hours on a working day. The persons have come one after the other and their names have been registered accordingly. The order is as follows:

MMM FF MMMM FFF MMM FFFFF MM

We have to ascertain whether male and female patients had come in a random manner.

Solution Using braces to combine the letters, which constitute the runs, we find that there is first run of three Ms, then a run of two Fs, then a run of four Ms, then a run of three Fs, then a run of three Ms, then a run of five Fs and finally a run of two Ms. Thus, we see that there are, in all, seven runs of varying lengths. If there are only a few runs, we might suspect that there is a definite grouping or clustering, or perhaps a trend. In contrast, if the number of runs is too large, we might suspect some sort of a repeated alternating pattern. In the example given above, there seems to be a definite clustering as both the males and females seem to come in groups. We have now to determine whether it is merely by chance or it is significant.

Now, we have to use the following formula to calculate the mean of the total number of runs n_1 and n_2.

$$\mu_r = [(2n_1n_2)/(n_1 + n_2)] + 1$$

where μ_r = mean of the sampling distribution of the r statistic

n_1 = number of observations in group 1

n_2 = number of observations in group 2

Likewise, the following formula is to be used for calculating the standard deviation:

$$\sigma_u = \sqrt{\frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}}$$

It may be noted that the sampling distribution of this statistic can be approximated closely with a normal distribution subject to the condition that neither n_1 nor n_2 is less than 10. In our example, this condition is fulfilled. For smaller values of n_1 and n_2, the test will have to be based on special tables. We can now base our decision on the statistic $Z = (u - \mu_r)/\sigma_r$ having approximately the standard normal distribution. If the calculated value of Z is less than $-Z_{\alpha/2}$ or greater than $Z_{\alpha/2}$, we reject the null hypothesis and accept the alternative hypothesis that the arrangement is non-random.

Let us specify our hypotheses:

H_0: Male and female patients had come in a random manner.

H_1: Male and female patients had come in a non-random manner, that is, in groups.

Returning now to our example, we find that n_1 (number of males) = 12, n_2 (number of females) = 10, u = 7 (total number of runs) and hence

$$\mu_r = (2 \times 12 \times 10)/(12 + 10) + 1 = 11.91$$

$$\sigma_\mu = \sqrt{\frac{(2 \times 12 \times 10)\,(2 \times 12 \times 10 - 12 - 10)}{(12 + 10)^2\,(12 + 10 - 1)}}$$

$$= \sqrt{\frac{240 \times 218}{484 \times 21}}$$

$$= \sqrt{\frac{52,320}{10,164}} = \sqrt{5.1476} = 2.269$$

and $Z = (u - \mu_r)/\sigma u = (7 - 11.91)/2.269 = -4.91/2.269 = -2.16$

Since this value (–2.16) is greater than $-Z_{0.01} = -2.58$, the null hypothesis cannot be rejected at the level of significance $\alpha = 0.01$. In other words, there is an indication that both male and female patients came in a random manner.

20.7 TESTS OF RANDOMNESS: RUNS ABOVE AND BELOW THE MEDIAN

The method just used to test the randomness is not confined to a series of attributes alone as in our example of Ms and Fs. In fact, any sample comprising numerical observations can be treated in the same manner by using the letters *a* and *b* to denote, respectively, values above the median and values below the median of the sample. In case an observation is equal to the median, it is omitted. The resulting series of *a*s and *b*s (representing the data in their original order) can be tested for randomness on the basis of the total number of runs above and below the median, respectively. Let us take an example.

Example 20.9 Suppose we have the following series of 29 college students. After performing a set of sturdy exercises, increases in their pulse rate were recorded as follows:

22, 23, 21, 25, 33, 32, 25, 30, 17, 20, 26, 12, 21, 20, 27, 24, 28, 14, 29, 23, 22, 36, 25, 21, 23, 19, 17, 26 and 26.

We have to test the randomness of these data.

Solution First, we have to calculate the median of this series. If we arrange these values in an ascending order, we find that the size of $(n + 1)/2^{th}$ item, that is, 15^{th} item is 24. Thus, the median is 24. As there is one value, which is 24, we omit it and get the following arrangement of *a*s and *b*s where *a* stands for an item greater than (or above) the median and *b* stands for an item lower than (or below) the median:

bbb aaaaa bb a bbb aa b a bb aa bbbb aa

On the basis of this arrangement, we find that n_1, (i.e. *a*s) = 13, n_2, (i.e. *b*s) = 15, and $u = 12$ as there are 12 braces, we get

$$\mu_r = [(2n_1n_2)/(n_1 + n_2)] + 1$$
$$= [(2\times13\times15)/(13 + 15)] + 1 = (390/28) + 1 = 14.93$$

$$\sigma_u = \sqrt{\frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}}$$

$$\sigma_u = \sqrt{\frac{(2\times13\times15)\,(2\times13\times15 - 13 - 15)}{(13 + 15)^2\,(13 + 15 - 1)}}$$

$$= \sqrt{\frac{390\times362}{(28)^2\,(27)}}$$

$$= \sqrt{\frac{1,41,180}{21,168}} = \sqrt{6.6695} = 2.58$$

$$Z = (u - \mu_r)/\sigma_u = (12 - 14.93)/2.58 = -2.93/2.58 = -1.14$$

Since $Z = -1.14$ falls between $-Z_{0.025} = -1.96$ and $Z_{0.025} = 1.96$, the null hypothesis cannot be rejected at the level of significance $\alpha = 0.05$. We can, therefore, conclude that the randomness of the original sample cannot be questioned.

It may be noted that this test is particularly useful in finding trends and cyclic patterns in a series. If there is a trend, there would be first mostly *a*s and later mostly *b*s or vice versa. In case of a repeated cyclic pattern, there would be a systematic alternation of *a*s and *b*s and , probably, too many runs.

20.8 KOLMOGOROV-SMIRNOV ONE-SAMPLE TEST

This test is concerned with the degrees of agreement between a set of observed values and the values specified by the null hypothesis. It is similar to the chi-square test of goodness-of-fit. It is used when one is interested in comparing a set of values on an ordinal scale. Let us take an example.

Example 20.10 Suppose that a company has conducted a field survey covering 200 respondents. Apart from other questions, it asked the respondents to indicate on a 5-point scale how much the durability of a particular product is important to them. The respondents indicated as follows:

Very important	50
Somewhat important	60
Neither important nor unimportant	20
Somewhat unimportant	40
Very unimportant	30
Total respondents	**200**

We have been asked to use the Kolmogorov-Smirnov test to test the hypothesis that there is no difference in importance ratings for durability among the respondents.

Solution In order to apply the Kolmogorov-Smirnov test to the above data, *first* of all, we should have the cumulative frequency distribution from the sample. *Second*, we have to establish the cumulative frequency distribution, which would be expected on the basis of the null hypothesis. *Third*, we have to determine the largest absolute deviation between the two distributions mentioned above. *Finally*, this value is to be compared with the critical value to ascertain its significance.

Table 20.9 shows the calculations.

Table 20.9 Worksheet for the Kolmogorov-Smirnov *D*

Importance of Durability	*Observed Number*	*Observed Proportion*	*Observed Cumulative Proportion*	*Null Proportion*	*Null Cumulative Proportion*	*Absolute Difference Observed and Null*
Very important	50	0.25	0.25	0.2	0.2	0.05
Somewhat important	60	0.30	0.55	0.2	0.4	0.15
Neither important nor unimportant	20	0.10	0.65	0.2	0.6	0.05
Somewhat unimportant	40	0.20	0.85	0.2	0.8	0.05
Very unimportant	30	0.15	1.00	0.2	1.0	0.00

From Table 20.9, we find that the largest absolute difference is 0.15, which is known as the Kolmogorov-Smirnov D value. For a sample size of more than 35, the critical value of D at an $\alpha = 0.05$ is $1.36/\sqrt{n}$ (Appendix Table 10). As sample size in this example is 200, $D = 1.36/\sqrt{200} = 0.096$. As the calculated D exceeds the critical value of 0.096, the null hypothesis that there is no difference in importance ratings for durability among the respondents is rejected.

Additional Examples

Example 20.11 A Company has collected the following data that relate to the average weekly loss of man-hours on account of accidents in 8 plants over a period of six months. The data were obtained to ascertain the effectiveness of an industrial safety 'before and after' programme was put into operation:

72 and 59, 26 and 24, 125 and 120, 39 and 35, 54 and 43, 39 and 35, 13 and 15, 12 and 18

Test the null hypothesis that the safety programme is not effective, using the two-sample sign test at $\alpha = 0.05$ level of significance.

Solution

Worksheet

Before I. S prog.		*After I.S. prog.*	
Plant No.	*Loss of Manhours*	*Loss of Manhours*	*Sign*
1	72	59	+
2	26	24	+
3	125	120	+
4	39	35	+
5	54	43	+
6	39	35	+
7	13	15	–
8	12	18	–

Last column of the Table shows + and – signs. When loss of manhours is more before I. S. programme as compared to those after I. S. programme, then a + sign is shown.

Out of 8 signs, 6 signs are positive and 2 negative.

Null hypothesis H_0: There is no significant difference in the loss of manhours i.e. $p = 0.5$

H_1: There is a significant difference $p \neq 0.5$

Now, we have to examine whether six successes in 8 trials supports the null hypothesis $p = \frac{1}{2}$ or the alternative hypothesis $p \neq \frac{1}{2}$. From Appendix Table 3 and for $n = 8$, we find that the probability of getting 6 or more successes is .1094 + .0312 + .0039 = 0.1445. It follows that H_0 can be rejected which means that there is a significant difference in loss of manhours in the latter period.

Example 20.12 Two different fertitizers were used on a sample of eight plots of same size each. The farm yield from these plots are given below:

Plot No.	1	2	3	4	5	6	7	8
Fertilizer I	49	32	44	48	51	34	30	42
Fertilizer II	40	45	50	43	37	47	55	57

The researcher would like to test the hypothesis (use Median test) that the two fertilizers have the same median. You can assume the level of significance at 5 percent. Write the hypothesis and test it.

Solution We have to test the hypothesis that the two fertilizers have the same median.

First of all, we have to calculate the combined median of the two series. Let us arrange the data in an ascending order.

30	45
32	47
34	48
37	49
40	50
42	51
43	55
44	57

$$\text{Median} = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \frac{16+1}{2} = 8.5^{\text{th}} \text{ item}$$

$$= \frac{44+45}{2} = 44.5$$

Classification of Fertilizers I and II

	Below Median	*Above Median*	*Total*
Fertilizer I	5	3	8
Fertilizer II	3	5	8
Total	**8**	**8**	**16**

We now calculate the expected frequencies by the formula: $\dfrac{\text{Row}_i \times \text{Column}_i}{\text{Total number of observations}}$

Observed frequencies	*Exp. Frequencies*	$O - E$	$(O-E)^2$	$(O-E)^2/E$
5	4	1	1	0.25
3	4	–1	1	0.25
3	4	–1	1	0.25
5	4	1	1	0.25
				1.00

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = 1$$

The critical value of χ^2 for $(2-1)(2-1) = 1$ degree of freedom at 0.05 level of significance is 3.841. Since the calculated value of χ^2 is less than the critical value, the null hypothesis is accepted. This means that the two fertilizers have the same median.

Example 20.13 A company used three different methods of advertising its product in three cities. It later found the increased sales (in thousand rupees) in identical retail outlets in the three cities as follows:

City A :	70	58	60	45	55	62	89	72	
City B :	65	57	48	55	75	68	45	52	63
City C :	53	59	71	70	63	60	58	75	

Use Kruskal-Wallis method to test the hypothesis that the mean increase in sales on account of three different methods of advertising was the same in the retail outlets in A, B and C cities. Use 5 per cent level of significance.

Solution First, we set up the null hypothesis (H_0) that the mean increase in sales on account of three different methods of advertising was the same in the retail outlets in *A*, *B* and *C* cities.

We now rank the data taking all three groups as if they are one. We start with the highest value as rank one and proceed in a descending order. This is done in the table given below.

Ranking in Three Cities

Increased Sales	89	75	75	72	71	70	70	68	65
City	A	B	C	A	C	A	C	B	B
Rank	1	2.5	2.5	4	5	6.5	6.5	8	9
Increased Sales	63	63	62	60	60	59	58	58	57
City	B	C	A	A	C	C	A	C	B
Rank	10.5	10.5	12	13.5	13.5	15	16.5	16.5	18
Increased Sales	55	55	53	52	48	45	45		
City	A	B	C	B	B	A	B		
Rank	19.5	19.5	21	22	23	24.5	24.5		

On the basis of the ranking done in the above table, the total ranks for the three cities are as follows:

$R_1 = 97.5$ $R_2 = 137$ $R_3 = 90.5$

Substituting the value of R_1, R_2 and R_3 along with values of n_1, n_2 and n_3 in the following formula:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{25(25+1)} \left(\frac{(97.5)^2}{8} + \frac{(137)^2}{9} + \frac{(90.5)^2}{8} \right) - 3\,(25+1)$$

$$= \left(\frac{12}{650} \times 4297.5 \right) - 78$$

$$= 79.34 - 78$$

$$= 1.34$$

The critical value of $\chi^2_{0.05}$ for $K - 1 = 3 - 1 = 2$ degrees of freedom is 5.991. Since the calculated value of $H = 1.34$ is less than the critical value of χ^2, H_0 is accepted. Hence, there is no significant difference in the mean increase of sales in the three retail outlets.

Example 20.14 A company manufacturing electronic toys has recently been taken over by another company. Prior to the takeover of the company, certain workers were approached to ascertain their satisfaction levels. The same workers were again approached to know their satisfaction level after the takeover of the company. The two sets of data are given below.

Before	69	73	58	76	82	65	75	64	87	70
After	65	75	63	75	82	68	71	65	85	68

Using an appropriate test, find out whether there has been an improvement in the satisfaction level of workers after the takeover of their company by a new company.

Solution This problem can be solved by applying Wilcoxon Matched-Pairs Test.

Application of Wilcoxon Matched-Pairs Test

Workers	*Before takeover*	*After takeover*	*Diff.* d_i	*Rank of* d_i	*Rank (–)*	*Rank (+)*
1	69	65	–4	7.5	–7.5	
2	73	75	2	4		4
3	58	63	5	9		9
4	76	75	–1	1.5	–1.5	
5	82	82	0	—		
6	65	68	3	6		6
7	75	71	–4	7.5	–7.5	
8	64	65	1	1.5		1.5
9	87	85	–2	4	–4	4
10	70	68	–2	4	–4	
				Total	–24.5	+20.5

H_0: There is no improvement in the satisfaction level of workers after the takeover of their company.
H_1: There is improvement in the satisfaction level of workers after the takeover of their company.

We have to take the lower value as T, which is 20.5. It may be noted that in 'Diff. d_i' column one entry is zero, which should be ignored. As such for the hypothesis test n is 9 instead of 10. The critical value of T for $n = 9$ at 5 per cent level of significance is 6 (Appendix Table 8). Since the computed T value is more than the critical value, the null hypothesis is accepted. We conclude that there is no improvement in the satisfaction level of workers.

Example 20.15 An experiment is conducted to judge the effect of brand name on quality perception. Sixteen subjects are recruited for the purpose and are asked to taste and compare two samples of product on a set of scale items judged to be ordinal. The following data are obtained:

Pair	1	2	3	4	5	6	7	8	9	10
Brand A	73	43	47	53	58	47	52	58	38	61
Brand B	51	41	43	41	47	32	24	58	43	53
Pair	11	12	13	14	15	16				
Brand A	56	56	34	55	65	75				
Brand B	52	57	44	57	40	68				

Test the hypothesis, using Wilcoxon matched-pairs test, that there is no difference between perceived quality of the two samples. Use 5 per cent level of significance.

Solution

Worksheet

Pair	*Brand A*	*Brand B*	*Difference di (A – B)*	*Rank of di*	*Rank (–)*	*Rank (+)*
1	73	51	22	13		13
2	43	41	2	2.5		2.5
3	47	43	4	4.5		4.5
4	53	41	12	11		11
5	58	47	11	10		10
6	47	32	15	12		12
7	52	24	28	15		15
8	58	58	0			
9	38	43	–5	6	–6	
10	61	53	8	8		8
11	56	52	4	4.5		4.5
12	56	57	–1	1	–1	
13	34	44	–10	9	–9	
14	55	57	–2	2.5	–2.5	
15	65	40	25	14		14
16	75	68	7	7		7
				Total	**–18.5**	**101.5**

Hence, $T = 18.5$

H_0 : There is no difference between the perceived quality of the two samples.

H_1 : There is difference between the perceived quality of the two samples.

The computed '*T*' value is 18. Ignoring pair number 8 where the difference between the perceived quality of the two brands is zero, the critical value of '*T*' for $n = 15$ at 5 per cent level of significance is 25. Since the computed '*T*' value is less than the critical '*T*' value, H_0 is rejected. The conclusion is that there is difference between the perceived quality of the two brands.

Example 20.16 The personnel director of a company wishes to select applicants for advanced training without regard to sex. Let 'W' denote women and 'M' denote men and the pattern of arrival be 'M WWW MMM WW M WWWW MMMM W M W MM WWW MM W MMMM WW M WW MMMM WW M WWWW MM WW M W M WW'

Will you conclude that the applicants have arrived in a random fashion?

Solution M WWW MMM WW M WWWW MMMM W M W MM WWW MM W MMMM WW M WW MMMM WW M WWWW MM WW M W M WW

$r = 28$ Number of runs

$n_1 = 28$ Number of males

$n_2 = 30$ Number of females

H_0: The applicants have arrived in a random fashion.

H_1: The applicants have not arrived in a random fashion.

We have to use the following formula:

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$= \frac{(2)(28)(30)}{28+30} + 1$$

$$= \frac{1680}{58} + 1 = 28.97$$

Standard error of the r statistic

$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2(28)(30)(2\times 28\times 30 - 28 - 30)}{(28+30)^2(28+30-1)}}$$

$$= \sqrt{\frac{1680\times 1622}{3364\times 57}}$$

$$= \sqrt{\frac{2724960}{191748}}$$

$$= \sqrt{14.2111}$$

$$= 3.77$$

$$Z = \frac{r - \mu_r}{\sigma_r}$$

$$= \frac{28 - 28.97}{3.77}$$

$$= \frac{-0.97}{3.77}$$

$$= -0.26$$

The critical value of Z for 0.05 level of significance is ± 1.96. It is a two-tail test. Since the calculated value of $Z = -0.26$ falls in the acceptance region, H_0 is accepted. Hence, we conclude that the applicants have come in a random fashion.

Example 20.17 The following arrangement shows the rise (U) or fall (D) in the price of an equity share on 40 consecutive trading days on which its price did not remain the same:

U U D D U U U D U U U D D U D D D U D D
U D D U U U U D D D D U U D U U D D D U

Test the hypothesis that this arrangement of Us and Ds is random at $\alpha =$ 0.05 level of significance.

Solution UU DD UUU D UUU DD U DDD U DD U DD UUUU DDDD UU D UU DDD U

We have to examine whether this arrangement of Us and Ds is random.

H_0 : The arrangement of Us and Ds is random.

H_1 : The arrangement of Us and Ds is not random.

$\alpha = 0.05$ level of significance.

n_1(U) = 20

n_2(D) = 20

r (number of runs) = 19

$$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{(2)(20)(20)}{20+20} + 1 = \frac{800}{40} + 1 = 21$$

Standard error of the r statistic

$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2(20)(20)(2\times 20\times 20 - 20 - 20)}{(20+20)^2(20+20-1)}}$$

$$= \sqrt{\frac{800\times 760}{1600\times 39}}$$

$$= \sqrt{\frac{608000}{62400}} = 3.1$$

$$Z = \frac{r-\mu_r}{\sigma_r} = \frac{19-21}{3.1} = \frac{-2}{3.1} = -0.645$$

The critical value of Z for 0.05 level of significance is ±1.96. As the calculated value of Z falls in the acceptance region, H_0 is accepted. This means that the arrangement of Us and Ds was random.

Example 20.18 The sequence of occurrence of 'zeros' and 'ones' in a message sent in a digital code is shown below. Test at 5 per cent whether the sequence of '0' and '1' is random 00110 11011 00001 11100 00110 11001 11110 00011 00100 11000 11100 00011 00111 11100 00000 11111 10001 11000 10001 01110.

Solution We first set up the hypotheses.

H_0: The sequence of 0 and 1 is random.

H_1 : The sequence of 0 and 1 is not random.

where r, being the number of runs, is 20.

Let $0 = n_1$ and $1 = n_2$

$n_1 = 51$ and $n_2 = 49$

Now, we use the following formula to calculate the mean of the total number of runs n_1 and n_2,

$$\mu_r = [(2n_1n_2)/(n_1 + n_2)] + 1$$

where μ_r is the mean of the sampling distribution of the r statistic

$$\mu_r = \frac{2\times 51\times 49}{51+49} + 1$$

$$= \frac{4998}{100} + 1 = 50.98$$

$$\sigma_\mu = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

$$= \sqrt{\frac{(2\times 51\times 49)(2\times 51\times 49 - 51 - 49)}{(51+49)^2(51+49-1)}}$$

$$= \sqrt{\frac{4998\times 4898}{10000\times 99}}$$

$$= \sqrt{\frac{24480204}{990000}}$$

$$= \sqrt{24.72747879} = 4.97$$

Applying the Z test,

$$Z = \frac{\mu - \mu_r}{\sigma_\mu}$$

$$= \frac{20 - 50.98}{4.97}$$

$$= \frac{-30.98}{4.97} = -6.233$$

The critical value of Z at 0.05 level of significance is –1.96. Since the calculated $Z = -6.233$ falls in the rejection region, the null hypothesis is rejected. The conclusion is that the sequence of 0 and 1 is not random.

Example 20.19 The municipal corporation of a city is interested to know the extent to which public is satisfied with sanitation in the neighbourhood. It decides to interview couples in a chosen locality. Husbands and wives are separately interviewed. They are asked, "How far are you satisfied with the sanitation in your locality". They are asked to rate their responses on a 10-point scale, with 10 being the most satisfied. The table given below records the responses of 14 couples.

Couple No.		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rating	H	8	3	2	0	3	6	7	6	7	4	2	5	4	5
	W	6	4	1	1	3	8	5	7	8	9	3	4	6	3

Using the sign test, test the hypothesis that there is no difference in the ratings by the husbands and the wives.

Solution

We set up a table showing responses of couples.

Couple No.	*Rating*		*Sign*
	H	*W*	$X_h - X_w$
1	8	6	+
2	3	4	–
3	2	1	+
4	0	1	–
5	3	3	0
6	6	8	–
7	7	5	+
8	6	7	–
9	7	8	–
10	4	9	–
11	2	3	–
12	5	4	+
13	8	6	+
14	5	3	+

H_0 : There is no difference in the ratings by the husbands and the wives.

H_1 : There is a significant difference in the ratings by the husbands and the wives.

As the sign in one case is zero, it is to be omitted.

Hence, $n = 14 - 1 = 13$.

We see that Pluses = 6 and Minuses = 7

The criterion for choosing between hypothesis is

$\alpha = 0.05 \quad n = 13 \quad z = 1.645$

Accept H_0 if $Z_c \leq 1.645$

Accept H_1 if $Z_a > 1.645$

Calculations:

Now, m = number of minuses = 7 $\quad p = 0.5$

$$E(m) = np = (13)(0.5) = 6.5$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{(13)(0.5)(1-0.5)}$$

$$= \sqrt{3.25}$$

$$= 1.803$$

Decision:

Accept H_1 because 1.803 > 1.645. The sign test shows a significant difference between the ratings by the husbands and the wives.

Example 20.20 A company's trainees are randomly assigned to groups which are taught certain industrial inspection procedure by three different methods. At the end of the instructing period, they are tested for inspection performance quality. The following are their scores.

Method Type	*Scores*
Method A :	80, 83, 79, 85, 90, 68
Method B :	82, 84, 60, 72, 86, 67, 91
Method C :	93, 65, 77, 78, 88

Use the H-test to determine, at the 0.05 level of significance, whether there is any difference in the effectiveness of the three methods.

Solution

H_0 : There is no difference in the effectiveness of the three methods.

H_1 : There is significant difference in the effectiveness of the three methods.

We now have to rank the performance scores, taking the three methods together. The table given below shows the rankings.

Worksheet

Ranking of the Three Methods

Scores	*Method*	*Rank*
93	C	1
91	B	2
90	A	3

(*Contd.*)

(*Contd.*)

88	C	4
86	B	5
85	A	6
84	B	7
83	A	8
82	B	9
80	A	10
79	A	11
78	C	12
77	C	13
72	B	14
68	A	15
67	B	16
65	C	17
60	B	18

The formula for the H-test is

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \text{ (where } R_i \text{ is the sum of rank's of Row } i)$$

Inputing the values, we get

$$H = \frac{12}{18(18+1)} \left(\frac{53^2}{6} + \frac{71^2}{7} + \frac{47^2}{5} \right) - 3(18+1)$$

$$= \frac{12}{342} \left(\frac{2809}{6} + \frac{5041}{7} + \frac{2209}{5} \right) - 57$$

$$= 0.035(468.17 + 720.14 + 441.8) - 57$$

$$= (0.035)(1630.11) - 57$$

$$= 57.05 - 57$$

$$= 0.05$$

The critical value of χ^2 for $2df$ is 5.991.

As the calculated value of H is 0.05, which is less than the critical value, the null hypothesis cannot be rejected. In other words, the three methods are equally effective.

Example 20.21 The following are the number of misprints, counted on pages, selected at random, from three Sunday editions of a newspaper.

April 11 :	4	10	1	6	4	12
April 18 :	8	5	13	8	3	10
April 25 :	7	9	11	2	14	7

Use the H-test at the 0.05 level of significance to test the null hypothesis that the 3 samples come from identical populations.

Solution

We first set up the hypotheses:

H_0 : The three samples come from identical populations.

H_1 : The three samples are not from identical populations.

Now, we have to rank the above data taking all the three dates together as if they are one. For sake of convenience, we may take April 11 as A, April 18 as B and April 25 as C. We start with the highest number of misprints as rank and proceed in a descending order. The following table shows the rankings.

Worksheet for Ranking of Misprints of Three Sunday Editions

Misprints	*Group*	*Rank*
14	C	1
13	B	2
12	A	3
11	C	4
10	A	5.5
10	B	5.5
9	C	7
8	B	8.5
8	B	8.5
7	C	10.5
7	C	10.5
6	A	12
5	B	13
4	A	14.5
4	A	14.5
3	B	16
2	C	17
1	A	18

The formula for the H-test, is as follows

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \quad \text{(where } R_i \text{ is the sum of ranks of Row } i\text{)}$$

Substituting the values in the above formula, we get

$$H = \frac{12}{18(18+1)} \left(\frac{67.5^2}{6} + \frac{53.5^2}{6} + \frac{50^2}{6} \right) - 3(18+1)$$

$$= \frac{12}{342} (759.375 + 468.167 + 416.667) - 57$$

$$= (0.035)(1644.209) - 57$$

$$= 57.547 - 57$$

$$= 0.547$$

The critical value of χ^2 for 2 degrees of fredom, at 0.05 level of significance, is 5.991.

As the calculated values of H is less than the critical value, the null hypothesis cannot be rejected. In the other words, the three samples come from identical populations.

Conclusion

We have discussed several nonparametric tests in this chapter. As was mentioned in the beginning of the chapter, these tests are suitable when stringent assumptions about the population may not be necessary. In particular, the most common assumptions of a normal distribution as are necessary for the parametric t, Z and F tests, is not required in case of nonparametric tests. A major advantage of such tests is that they need limited information though, as a result, they are less powerful. Since there are several nonparametric tests, one has to be careful in choosing the most appropriate test for a given set of data or problem.

Before we close this chapter, it may be pointed out that the two nonparametric tests, viz. the rank correlation and the chi-square have not been discussed here because they have already been covered earlier. Chi-square has been discussed in Chapter 14 while rank correlation in Chapter 17. Although there are a number of nonparametric tests, we have presented some of the more frequently used tests in this chapter. While using these tests, we must know that the advantages we derive by limiting our assumptions may be offset by the loss in the power of such tests. However, when basic assumptions as required for parametric tests are valid, the use of nonparametric tests may lead to a false hypothesis and thus we may commit a Type II error. We have to consider this aspect very carefully before deciding in favour of nonparametric tests. It may be reiterated that such tests are more suitable in case of ranked, scaled or rated data.

GLOSSARY

Kolmogorov-Smirnov test	A nonparametric test that is concerned with the degrees of agreement between a set of observed ranks (sample values) and a theoretical frequency distribution.
Kurskal-Wallis test	A nonparametric method for testing the null hypothesis that K independent random samples come from identical populations. It is a direct generalisation of the Mann-Whitney test.
Mann-Whitney U test	A nonparametric test that is used to determine whether two different samples come from identical populations or whether these populations have different means.
Nonparametric tests	Tests that rely less on parameter estimation and/or assumptions about the shape of a population distribution.
One-Sample Runs test	A nonparametric test used for determining whether the items in a sample have been selected randomly.
Rank Sum tests	A group of nonparametric tests that are based on the sum of ranks assigned to the observations in samples.
Run	A sequence of identical occurrences that may be preceded and followed by different occurrences. At times, they may not be preceded or followed by any occurrences.
Sign test	A nonparametric test that takes into account the difference between paired observations where plus (+) and minus (–) signs are substituted for quantitative values.

Theory of runs	A theory concerned with the testing of samples for the randomness of the order in which they have been selected.
Wilcoxon Matched-pairs Test (or Signed Rank Test)	A nonparametric test that can be used in various situations in the context of two related samples.

LIST OF FORMULAE

1. U statistic in Mann-Whitney U test to measure the difference between the ranked observations of the two variables:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

where R_1 is the sum of ranks of observations of variable 1
n_1 is the number of observations in sample 1
n_2 is the number of observations in sample 2

2. U statistic in Mann-Whitney U Test:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

This formula and the formula in (1), above, can be used interchangeably. This is preferable when sample 2 has considerably less number of observations.

3. Mean of the U statistic in the Mann-Whitney U Test:

$$\mu_u = \frac{n_1 n_2}{2}$$

4. Standard error of the U statistic of the Mann-Whitney U Test:

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

5. Z Statistic in M-W Test:

$$Z = U - \mu_u / \sigma_u$$

6. H statistic in the Kruskal-Wallis test for different means among three or more populations:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n + 1)$$

where n_i = the number of observations in the i^{th} group
R_i = the sum of the ranks of all observations in the i^{th} group
K = number of samples
$n = n_1 + n_2 + \ldots + n_k$, the total number of observations in all samples

7. Mean of the sample distribution of the r statistic used in a one-sample runs test:

$$\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

Here, r stands for the number of runs in the sample being tested.

8. Standard error of the r statistic in a one-sample runs test:

$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

9. *K–S* statistic:

$$D_n = \max | Fe - Fo |$$

where D_n stands for the maximum absolute deviation of *Fe* (expected frequencies) from *Fo* (observed frequencies).

QUESTIONS

20.1 Given below are ten statements. Indicate in each case whether it is true or false.

(a) A limitation of nonparametric methods is that they do not consider a certain amount of information.

(b) Nonparametric methods give more accurate results than the parametric methods.

(c) As compared to the sign test, the Mann-Whitney U test tends to waste less data.

(d) The sequence A, A, A, B, B, A, B, A, A, B, B contains six runs.

(e) A condition in the Mann-Whitney U test is that the two samples must be of the same size.

(f) In order to measure the goodness-of-fit of a theoretical distribution, the Kolmogorov-Smirnov test can be used.

(g) In a one-sample test, the null hypothesis assumes that the sequence of observations is random.

(h) The Kruskal-Wallis test can be regarded as a nonparametric version of Analysis of Variance.

(i) To test whether two independent samples have been drawn from populations with the same distributions, we can use the one-sample runs test.

(j) The Mann-Whitney U test belongs to a group of tests known as rank-difference tests.

Multiple Choice Questions (20.2 to 20.11)

20.2 Nonparametric methods in comparison with parametric methods

(a) are easier to compute (b) are less efficient

(c) are less accurate (d) need less information

(e) (a), (b) and (d) (f) none of these

20.3 In the Kruskal-Wallis test having k samples, the appropriate number of degrees of freedom is

(a) $k - 1$ (b) k (c) $n - k$ (d) $n - k - 1$

(e) none of these

20.4 For a perfect correlation, the coefficient of rank correlation r_s would be

(a) 1 (b) –1 (c) zero (d) none of these

20.5 The sequence of A, B, A, B, A, B, A, B, A, B is likely to be rejected by a test of runs as not being truly random because

(a) The sequence has very far runs.

(b) The sequence has too many runs.

(c) The sequence has only two symbols.

(d) The pattern A, B occurs only 5 times which is too less to guarantee randomness.

(e) None of these.

Questions 20.6 and 20.7 refer to the following situation:

Assume that in a certain hospital, 5 patients from ward A and 4 patients from ward B were selected at random and their duration of stay (number of days) in the hospital was noted down as follows:

Ward A	10 7	5 2	15
Ward B	8 7	3 12	

20.6 In order to determine whether there is a significant difference between the duration of hospital stays for the two wards, a Mann-Whitney U test is to be performed. If the durations of stay are ranked in an ascending order, what is the ranking for the 10-day stay in ward A?

(a) 8 (b) 7 (c) 6 (d) 5

20.7 If the durations of stay are ranked in an ascending order, what is the value of $(R_1 - R_2)$?

(a) 4 (b) 2 (c) 9 (d) 0

20.8 Which of the following is a signed rank test?

(a) The Mann-Whitney Test
(b) The Two-sample Sign Test
(c) One-sample Sign Test
(d) Wilcoxon Matched-pairs Test
(e) None of these

20.9 Assuming that the two brands of light bulbs A and B are to be compared by using the Mann-Whitney test, which of the following steps is valid?

(a) Ranking of the given data separately for brands A and B
(b) Ranking of the given data jointly
(c) Absolute figures instead of ranking are used
(d) None of the above

20.10 Which of the following is most appropriate in a nonparametric test?

(a) F test (b) t test (c) U test (d) Z test

20.11 The formula for Spearman's correlation is

(a) $6\Sigma d^2/N^2 - 1$
(b) $6\Sigma d^2/(N - 1)$
(c) $1 - [6\Sigma d^2/N(N^2 - 1)]$
(d) $1 + [6\Sigma d^2/N(N^2 - 1)]$
(e) None of these

20.12 What do you understand by nonparametric or distribution-free methods?

20.13 What are the major advantages of nonparametric methods over parametric methods?

20.14 What are the main limitations of nonparametric tests?

20.15 Enumerate the different nonparametric tests and explain any two of them.

20.16 Random numbers are given in Appendix Table 12. Choose any four complete columns of random digits (100 digits in all). Represent each even digit by the letter E and each odd digit by the letter O. Having done this, now test for randomness on the basis of the total number of runs at $\alpha = 0.05$.

20.17 The proprietor of a small business computed his average earnings per day over a period of 12 days. For each day, an L was recorded if the earnings were less than the average, otherwise an M was recorded. These data are given below:

L L L L M M L L L L M M

Do you think that the data indicate a lack of randomness at the 5 per cent level of significance?

20.18 A pharmaceutical company has come out with a new pain-relieving lotion to relieve muscular pain in 10 minutes on overage. To test its claim, 15 people were asked to use the lotion when they felt muscular pain. As a result, the following data (in minutes) emerged:

11.0 11.3 10.1 10.2 13.1 12.1 10.7 10.3 9.1 9.2 12.2 10.6 12.0 9.5 11.1

You are asked to verify the company's claim by using the Wilcoxon signed-rank test with a 1 percent level of significance.

20.19 In a metropolitan city, a city bus service was scheduled to reach a major bus stop at 11 a.m. each day. If the bus reached that stop within 5 minutes of 11 a.m. it was considered to be on time. Over a 15-day period, an A was recorded if the bus was on time, otherwise a B was recorded. The picture that emerged after ten days was as follows:

A A B A B B A B A A B B B A A

Do you find a lack of randomness in the above data ? Use a 5 per cent level of significance.

20.20 A company is engaged in the manufacture of electrical components. A sample of 15 such components inspected at 10-minute interval. If the component did not satisfy the required specifications, a W was recorded, otherwise an R was recorded. The results of this inspection are given below:

R R R R R R W W W R R R R R R

Examine whether the process was random, using 5 per cent level of significance.

20.21 The management of a factory decided to evaluate a shift on the basis of a number of hours to complete a job. In this connection, the management set the median time to complete the job at 2.8 hours. However, the plant manager claimed that the median time of the shift was longer. A sample of ten completed jobs gave the following times:

3.4 3.3 2.0 3.2 2.5 3.7 3.5 3.1 4.1 3.6

Test the manager's claim using the sign test with a level of significance at 5 per cent.

20.22 The following data show employees' rate of substandard performance before and after a new incentive scheme. Determine whether the introduction of the new incentive scheme has reduced the substandard performance at 0.05 level of significance.

Before	7	8	5	9	10	6	5	9	6	8
After	5	6	7	6	9	7	6	6	5	7

20.23 The following data relate to the costs of building comparable lots in the two Resorts A and B (in million rupees):

Resort A :	30.9	32.5	44.3	39.5	35.0	48.9
Resort B :	53.9	61.0	36.0	42.5	40.9	47.9

The company owning the resort area A claimed that the median price of building lots was less in area A as compared to resort area B. You are asked to test this claim, using a nonparametric test with a 1 per cent level of significance.

20.24 The following table gives the sample data along with three blank columns.

Sample Data	*Difference*	*Magnitude*	*Signed Rank*
X	$D = X - 50$		
37			
55			
42			
46			
39			
33			
52			

Additional information : $Ha : Md < 50$, $\alpha = 0.05$. You are asked to (a) complete the table; and (b) use the Wilcoxon signed - rank test and complete the test.

20.25 The following are the number of workers who did not report to work in their factory on 15 consecutive days:

11, 9, 13, 15, 12, 20, 7, 12, 10, 6, 10, 14, 9, 7 and 10.

Test for randomness at 5 per cent level of significance.

20.26 The management of a company was interested to know the opinion of its workers on a new incentive scheme. It interviewed 34 workers, who either said they were in favour of the scheme (F) or against it (A). Their responses are shown below in the order in which they were received.

FFFFF AA FFF AAA FF A FFFF AA FF AA FFF AAA FF

Use an appropriate test to determine whether this arrangement of Fs and As was random at 5 per cent level of significance.

20.27 On 15 different days, A had to wait for the city bus to reach his office as shown below:

17, 12, 18, 20, 25, 30, 10, 15, 7, 10, 9, 11, 5, 11 and 20 minutes.

Use the sign test at 5 per cent level of significance to test the bus company's claim that on an average A should not have to wait for more than 15 minutes.

20.28 Appendix Table 12 gives random numbers. Choose any four complete columns of random digits (100 digits in all), represent each even digit by the letter *E*, each odd digit by the letter O, and test for randomness on the basis of the total number of runs at $\alpha = 0.05$.

20.29 The following arrangement indicates the movement of the equity share of a company in the stock exchange. The letter *R* stands for rise in the share price while the letter *D* shows decline in its price on 30 consecutive trading days, on which there was change in the share price.

D D R R R D R D D D R R R R R
R R D R R R R D R D R D D D R

Test whether this movement of *R* and *D* may be regarded as random, using α at 0.05.

20.30 A professor has two classes in Statistics, a morning class of 9 students and an afternoon class of 12 students. For a final examination, scheduled at the same time for all students, the classes received marks, as shown in the Table. Can one conclude, at the 5% significance level,

that the morning class performed worse than the afternoon class? Solve the problem using the *U* test.

Morning Class Marks	*Afternoon Class Marks*
73	86
87	81
79	84
75	88
82	90
66	85
95	84
75	92
70	83
	91
	53
	84

20.31 Three different brands of king size cigarettes were tested for tar content in a pack of 10 cigarettes. Five packs of each brand were tested. The tar content, in milligram, for the three brands is listed in the following table. Using Kruskal-Walli's test, verify, at $\alpha = 0.05$ level of significance, that there is no significant difference in the three brands of cigarettes in terms of tar content.

	X	*Y*	*Z*
1	10	16	12
2	14	13	14
3	13	11	10
4	11	14	17
5	12	10	11

20.32 A consumer panel has 14 individuals. Its is asked to rate two brands of coca-cola according to a point evaluation system, based on several criteria. The table given below reports the points assigned to each brand by the panelists. Test the null hypothesis that there is no difference in the level of ratings for the two brands of cola, at 5% level of significance, using the sign test.

Table for Point rating assigned

Panel Member	*Brand 1*	*Brand 2*
1	20	16
2	24	26
3	28	18
4	24	17
5	20	20
6	29	21
7	10	23
8	27	22
9	20	23

(*Contd.*)

(*Contd.*)

10	30	20
11	18	18
12	28	21
13	26	17
14	24	26

20.33 Two methods of instruction to apprentices are to be evaluated. A director assigns 15 randomly selected trainees to each of the two methods. Due to drop-out, 14 compete in Batch 1 and 12 compete in Batch 2. An achievement test is given to the successful candidates. Their scores are as follows:

Method 1 :	70,	90,	82,	64,	86,	77,	84,	79,	82,	89,	73,	81,	83	66
Method 2 :	86,	78,	90,	82,	65,	87,	80,	88,	95,	85,	76,	94		

Test whether the two methods have significant difference in effectiveness. Use Mann-Whitney test at 5% significance level.

20.34 A quality control engineer in an electronics plant has sampled the output of three assembly lines and recorded the number of defects observed. The samples involve the entire output of the three lines for 10 randomly selected hours, in a given week. Do the data provide sufficient evidence to indicate that at least one of the lines tend to produce more defects than others. Test at the 5% level of significance using a suitable nonparametric test.

Assembly lines	*Number of Defects*									
Line 1	6	38	3	17	11	30	15	16	25	5
Line 2	34	28	42	13	40	31	9	32	39	27
Line 3	13	35	19	4	29	0	7	33	18	24

The McGraw-Hill Companies

CHAPTER 21 INDEX NUMBERS

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the concepts and techniques of different types of fixed-base index numbers
- understand and calculate the chain index numbers
- understand and apply the time and factor reversal tests, as also the circular test, to a given index number
- carry out splicing and shifting of base
- make use of a price index to deflate a series of data, such as wages, profits, and so on
- construct index numbers of production (or quantity)
- familiarise yourself with the problems involved in constructing and using index numbers.

Chapter Prerequisites

Before starting work on this chapter, make sure you are conversant with:

1. the ideas of percentages
2. the use of the Σ sign

21.1 INTRODUCTION

In Chapter 19, we discussed at length the analysis of time series and forecasting. The time series used therein related to original data. There is another form of time series and this is in the form of price and quantity index numbers. In a way, this chapter on index numbers can be regarded as an extension of that chapter, though it may be mentioned that the scope of index numbers can be beyond time as well.

If we turn to any journal devoted to economic and financial matters, we are very likely to come across an index number of one or the other type. It may be an index number of share prices or a wholesale price index or a consumer price index or an index of industrial production. Such an index number always involves a comparison of the magnitude of a certain variable over time or space.

The objective of these index numbers is to measure the changes that have occurred in prices, cost of living, production, and so forth. With the help of such index numbers, economists and businessmen are able to describe and appreciate business and economic situation quantitatively.

Our focus in this chapter will be on price index numbers. However, towards the end of this chapter, we shall deal with the production or quantity index numbers. Another point worth emphasising here is that the discussion mostly relates to the methodology of index number construction. The scope of the chapter is rather limited and as such, it does not discuss a large number of index numbers that are presently compiled and published by different departments of the Government of India.

An index number is a number that measures the change in a variable between two points of time. For example, if a wholesale price index number for the year 2000 with base year 1990 was 170, it shows that wholesale prices, in general, increased by 70 per cent in 2000 as compared to those in 1990. Now, if the same index number moves to 180 in 2001, it shows that there has been 80 per cent increase in wholesale prices in 2001 as compared to those in 1990.

There are *two types of index numbers* that are frequently found in business and economic applications. These are *price* and *quantity index numbers*. The former measure changes in the price of a commodity or a group of commodities over time. The latter, on the other hand, measure changes in the production of a commodity or a group of commodities over time.

It may be noted that there are several methods that can be used in the construction of index numbers. These methods range from extremely simple to most complex, depending on the number as well as types of commodities covered in a particular index.

21.2 USES OF INDEX NUMBERS

1. Index numbers are a guide to business policy. In case prices are rising, there will be one set of policies, whereas falling prices need a different one. A stable price level or slowly rising price level is considered a favourable factor for the growth of business and industry.
2. Indices of industrial production are useful as they are the indicators of business environment. A rising index of industrial production shows that business firms are optimistic of the demand for their products and are coming out with higher output. In case the index of industrial production is declining, the business environment may be regarded unfavourable. This apart, an index of industrial production can be used by business firms to compare their own production.
3. At times, index numbers can be combined into one series that may be relevant to one's own business. A change in such a combined series is likely to forecast changes in one's own business.
4. Index numbers can be very helpful in finding out whether a business firm's sales are increasing in *physical volume* as opposed to *rupee value.* This is because there may be several products that the firm is selling; sale of some may be increasing while of others declining in regard to the number of units sold. Further, the unit prices of different products may be changing. In such a situation, it is an index of physical volume of firm's sales that may be used. Alternatively, the value of sales in terms of current prices may be 'deflated', by dividing them by an index of firm's prices.
5. Nowadays index numbers are frequently used to adjust wages or salaries on account of rising prices or inflation. For example, in case of government employees, the dearness allowance is enhanced on the basis of rise in consumer price index (or wholesale price index) by pre-specified points. In the absence of index numbers, it would be difficult to decide how much increase in dearness allowance should be made so as to compensate employees due to inflation.

Again, in case of long-term contracts, the usual practice is to include a clause in the agreement that provides for an increase in the contract amount on account of any increase in the price level during a specified period.

6. Index numbers can be used for providing incentive to efficient workers. For this purpose, productivity indices are constructed. Workers are told that if the overall productivity of the firm increases, they will stand to gain as they will get 'productivity bonus' for their improved performance.

21.3 PROBLEMS IN INDEX NUMBER CONSTRUCTION

In this section, we will discuss four major considerations that must be looked into while constructing an index number. These are: *selection of base*, *type of formula*, *weighting system* and *the data for index numbers*.

Selection of a Base Year The selection of a base year or period with which to compare the different index numbers does not pose difficult theoretical questions. To a large extent, the choice of the base year depends on the objective of the index. A major consideration should be to ensure that the base year is not an abnormal year. If we select, for example, a year of recession, then our price index will turn out to be defective. Another consideration is that the base year should not be too remote in the past. A more recent year needs to be selected as the base year. The use of a particular year for a prolonged period would distort the changes that it purports to measure. That is why, we find that the base year of major index numbers, such as consumer price index or index of industrial production, is shifted from time to time.

Type of Formula Another relevant issue in the construction of index numbers is: which formula should be used? There are numerous formulae that can be used, but each will give different result. While deciding the formula to be used, we should ensure that it is technically sound and suitable for the particular purpose. A lot of discussion has taken place amongst statisticians in respect of various formulae, but the discussion is more theoretical rather than practical.

Selection of Weights Selection of proper weights to be assigned to various items covered in the index is an important issue. Since the different items chosen for an index do not have the same importance, it becomes necessary to assign an appropriate weight to each item. A major consideration in choosing weights is to assess the relative importance of the items. This may be based on the proportion of expenditure incurred on that item to total expenditure incurred on all items. However, the weights assigned at one time may not remain valid at a later date on account of changes in the relative importance of the items included in the index. This means that we have to review from time to time that the weights do not become out of date. In case, we find that weights are no longer in conformity with the relative importance of the items, we have to decide on the new weights.

The Data for Index Numbers This perhaps is the most important problem in index number construction. The discussion in earlier chapters has shown that if our data are wrong or defective in some sense, then our findings would also be misleading. As such, we have to give proper attention in the collection of data. We must ensure that the data are accurate and comparable. This is possible if our sample is of adequate size as well as representative of the population. Further, data should be as recent as possible as the utility of index numbers lies in their being up-to-date. The data should be relevant keeping in mind the items or commodities used by the class of persons for whom the index is being constructed.

21.4 TYPES OF PRICE INDEX NUMBERS

In this section, we shall discuss different types of index numbers proceeding from very simple to relatively difficult ones.

Simple Index Numbers

A simple index number is based on the price or quantity of a single commodity. For example, consider the price of a commodity for the period 1995 to 2000.

Table 21.1 Price of Commodity *X*, 1995 to 2000

Year	1995	1996	1997	1998	1999	2000
Price (Rs per tonne)	3,000	3,200	3,250	3,100	3,290	3,600

To construct a simple index to indicate the relative changes in commodity X prices, first of all a base period is to be chosen. This will enable us to make comparisons in the prices of other years in relation to the base year price. In this case, we take 1995 as the base period. The next step is the calculation of the simple index number for each year. This is done by dividing that year's price by the price of the base period or year and multiplying the resultant figure by 100. Thus, the price index number for 1996 will be

$$\text{Price index number for } 1996 = \frac{3,200}{3,000} \times 100 = 106.67$$

Similarly, price index number for the year 2000 will be

$$= \frac{3,600}{3,000} \times 100 = 120$$

This shows that during the period 1995–2000, the price of commodity X increased by 20 per cent. In the same manner, index numbers for other years can be calculated.

Utility of Simple Index Numbers The use of simple index numbers enables us to compare price changes in a commodity over time. In addition, it enables comparisons in price changes in respect of two or more commodities that have different units of measurement. For example, the price of cloth is given in terms of metre whereas the price of toothpaste is in terms of grams. Changes in prices of these commodities can be easily compared over time with the help of simple index numbers provided the base period is the same.

Composite Index Numbers

The preceding discussion was confined to only one commodity. What about price changes in several commodities? In such cases, composite index numbers are used. A composite index number is an index for a time series consisting of the total price of two or more commodities.

Example 21.1 Suppose we want to compare the prices of three commodities—bread, milk and eggs for year 1 and year 2. These are shown as follows.

Table 21.2 Prices of Three Commodities for Two Years

Commodity	*Quantity*	*Price Year 1 (Rs)*	*Price Year 2 (Rs)*
Bread	Loaf	10	14
Milk	Litre	15	20
Eggs	Dozen	10	16
	Total	**35**	**50**

Solution

$$\text{Price index number for year 1} = \frac{35}{35} \times 100 = 100$$

$$\text{Price index number for year 2} = \frac{50}{35} \times 100 = 142.86 \text{ or } 143 \text{ approx.}$$

This is also called *simple aggregative price index*. What we have done is a simple comparison of price change in all the three commodities in year 2 with base year 1. The above index shows that while the prices of bread, milk and eggs moved differently, taking these commodities together, they registered a 43 per cent increase in year 2 over year 1.

A simple aggregative index has two major *limitations*: **(i)** The units of prices of the commodities will affect the price index. **(ii)** The relative importance of the commodities is not taken into consideration.

Simple Average of Price Relatives

Another method of constructing index numbers is to first calculate price relatives of the commodities covered in the index. This is followed by taking an average of price relatives. This will be clear from the following example.

Example 21.2 Using the data given in Table 21.2, calculate the price index for year 2, the base year being year 1.

Solution The last column of Table 21.3 shows the price relatives for the three commodities.

Table 21.3 Simple Average of Price Relatives

Commodity	*Quantity*	*Price Year 1 (Rs)*	*Price Year 2 (Rs)*	p_1 / p_0
Bread	Loaf	10	14	14/10 = 1.4
Milk	Litre	15	20	20/15 = 1.33
Eggs	Dozen	10	16	16/10 = 1.6

$$\text{Price index (year 2)} = \frac{1.4 + 1.33 + 1.6}{3} \times 100 = 1.44 \times 100 = 144$$

It will be seen that this method has given a marginally different result than the previous one. What we have done is to take a simple average of price relatives. The price relative for bread is 1.4, for milk 1.33 and for eggs 1.6. Taking their average gives us a figure of 1.44, which is multiplied by 100.

Weighted Relative Price Index Numbers

A crucial question in the construction of index numbers is: what kind of weights should be given to different commodities? It may be noted that weights should reflect the relative importance of commodities. The amount spent in terms of rupees in the base year may be taken as the weight showing the relative importance. This will be clear from the following example.

Example 21.3 Construct weighted relative price indices, using the data given in Table 21.4.

Table 21.4 Price and Quantity Data for Some Products

Commodity	*Base Year Values (Rs)*	*Current Year Price (Rs)*	*Current Year Quantity*
Bread	300	Rs 10 per loaf	500
Milk	500	Rs 15 per litre	600
Butter	240	Rs 100 per kg	4

Solution Thus, in the base year, the total expenditure incurred was Rs 1,000. The relative importance of bread, milk and butter may be taken in the proportion of the amount spent to the total expenditure on these three items.

If we multiply each price relative by the weight (i.e. the expenditure incurred for that commodity in the base year), we will have the current year price multiplied by the base year quantity. That is, $p_1/p_0 \times p_0q_0 = p_1q_0$. The total of three commodities will be Σp_1q_0. This is to be divided by the base year total value, Σp_0q_0, to yield an overall index.

Symbolically,

$$P_{01} = \frac{\Sigma[(p_1/p_0)v]}{\Sigma p_0q_0} \quad \text{where, } v \text{ stands for value, i.e. expenditure incurred for a commodity}$$

$$= \frac{\Sigma(p_1/p_0) \times p_0q_0}{\Sigma p_0q_0}$$

$$= \frac{\Sigma p_1q_0}{\Sigma p_0q_0}$$

In order to use this last formula, it is necessary for us to have separate figures of price and quantity for each commodity. Using these figures as given in Table 21.5, we now apply the formula to calculate an overall index.

Table 21.5 Calculation of Weighted Relative Price Indices

Commodity	p_1	q_0	p_1q_0	p_0q_0
Bread	10	50	500	300
Milk	15	50	750	500
Butter	100	3	300	240
		Total	**1,550**	**1,040**

Weighted relative price index for year 1 $= \frac{\Sigma p_1q_0}{\Sigma p_0q_0} = \frac{1,550}{1,040} = 1.49$

∴ The price index is $1.49 \times 100 = 149$

This shows that the index for the three commodities is 149 for the current year, that is, an increase of 49 per cent.

Laspeyres and Paasche Formulae

So far our discussion related to base-year weights. One can assign current-year weights to the items covered in an index. The index number based on the base-year weights is known as the Laspeyres Index whereas the index number based on the current year weights is known as the Paasche Index. The index numbers in our previous examples were all Laspeyres indices.

The Passche index, symbolically is

$$\text{Index} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$$

where price of each commodity in the current year is to be multiplied by the quantity purchased in the current year, that is, the total expenditure incurred. Further, the price in the base year is to be multiplied by the current year quantity.

This has to be carried out for all the items covered by the index. Now, the expenditure incurred in the current year is to be divided by the total just arrived at, yielding the Passche index. It may be noted that the construction of the Passche Index necessitates that a new set of quantities (q_n) has to be found for each current year that is to be compared with the base year.

The following example illustrates the construction of the Passche index.

Example 21.4 Given below are the price and quantity data for two years relating to three commodities. We are asked to calculate the Passche index.

Commodity	p_0	q_0	p_1	q_1
X	10	40	15	60
Y	15	80	20	100
Z	20	20	25	40

Solution We set up Table 21.6 for calculating the Passche index.

Table 21.6 Calculation for the Passche Index

Commodity	p_0	q_0	p_1	q_1	p_0q_1	p_1q_1	p_1q_0
X	10	40	15	60	600	900	600
Y	15	80	20	100	1,500	2,000	1,600
Z	20	20	25	40	800	1,000	500
				Total	**2,900**	**3,900**	**2,700**

$$\text{Paasche index} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$$

$$= \frac{3,900}{2,900}$$

$$= 1.345 \times 100 = 134.5$$

Apart from Laspeyres and Paasche indices, there are other indices based on some other methods of weighting. These indices are known after the names of those who propounded them. They are given below:

$$\text{Bowley's method} = \left\{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right)\Big/ 2\right\} \times 100$$

$$\text{Marshall-Edgeworth's method} = \frac{\Sigma p_1 (q_0 + q_1)}{\Sigma p_0 (q_0 + q_1)} \times 100$$

$$\text{Fisher's method} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

It will be seen that Bowley's method is an average of Laspeyres and Passche indices. The index based on Marshall-Edgeworth's method uses the total quantities for both base year and current year as the weight. Finally, Fisher's method is the square root of the Laspeyres and Paasche indices. In fact, Fisher suggested a number of formulae for the construction of index numbers. This particular formula is known as Fisher's Ideal Index and it is the geometric mean of the Laspeyres and Passche indices. It is known as the 'ideal' index because it satisfies time reversal and factor reversal tests, which are discussed below.

21.5 TIME REVERSAL, FACTOR REVERSAL AND CIRCULAR TESTS

We take up time and factor reversal tests.

Time Reversal Test

A method satisfies time reversal test if it gives $P_{01} \times P_{10} = 1$, where P_{01} is the price index number for the current year with the base year 0 and P_{10} is the index number of the base year, taking current year as the base, both the indices without the factor 100.

Factor Reversal Test

A method satisfies factor reversal test if it gives $P_{01} \times Q_{01} = (\Sigma p_1 q_1 / \Sigma p_0 q_0)$, where P_{01} is the price index for the current year (without the factor 100) and Q_{01} is the quantity index for the current year (without the factor 100). Let us examine these two tests with some examples.

Example 21.5 Suppose there is only one commodity where price in the base year is Rs 50 per unit and in the current year, Rs 80 per unit. We have to ascertain whether the time reversal test is satisfied.

Solution We find that the index in the current year is p_1/p_0 or 80/50 = 1.6. If we work backward—taking current year as the base, then p_0/p_1 = 50/80 = 0.625. Now, the time reversal test shows that $P_{01} \times P_{10} = 1$. In our example, $1.6 \times 0.625 = 1$. In other words, the forward and backward index numbers multiplied together should give unity.

We take another example, this time with two commodities.

Example 21.6 Given the price of commodity A Rs 10 and Rs 15 in the base year and the current year and that of commodity B Rs 40 and Rs 48, respectively, find out whether the time reversal test is satisfied.

Solution We set up Table 21.7 for calculating the price indices.

Table 21.7 Calculation for Time Reversal Test

Commodity	p_0 (Rs)	p_1 (Rs)	P_{01}	P_{10}
A	10	15	150	66.7
B	40	48	120	83.3
			270/2 = 135	150.0/2 = 75

$$P_{01} \times P_{10} = 1.35 \times 0.75 = 1.0125$$

Although this figure is very close to 1 but, strictly speaking, it fails to satisfy the time reversal test. If we take the geometric mean, we will see that the time reversal test is satisfied.

$$P_{01} = \sqrt{150 \times 120} = \sqrt{18{,}000} = 134.164$$

$$P_{10} = \sqrt{66.7 \times 83.3} = \sqrt{5{,}556.11} = 74.539$$

$$P_{01} \times P_{10} = 134.164 \times 74.539$$

or 1.34164×0.7454 (the above figures divided by 100 each)

$$= 1.0$$

This shows that the geometric mean satisfies the time reversal test whereas the arithmetic mean does not. Let us take a more elaborate example.

Example 21.7 The price and quantity data for two years for five commodities are given below;

Commodity	p_0	q_0	p_1	q_1
A	10	5	15	5
B	5	10	5	12
C	8	4	10	5
D	12	5	15	5
E	6	15	12	10

Calculate the Laspeyres and Paasche index numbers and ascertain whether they satisfy the time and factor reverseal tests.

Solution We set up Table 21.8 for the computation of time and factor reversal tests.

Table 21.8 Computation for Reversal Tests

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
A	10	5	15	5	50	75	50	75
B	5	10	5	12	50	50	60	60
C	8	4	10	5	32	40	40	50
D	12	5	15	5	60	75	60	75
E	6	15	12	10	90	180	60	120
				Total	**282**	**420**	**270**	**380**

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} = \frac{420}{282} = 1.4894$$

$$P_{10} = \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0} = \frac{282}{420} = 0.6714$$

These are Laspeyres price indices for the current year and the base year, respectively. The corresponding Paasche price indices are:

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \text{ and } \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1}$$

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = \frac{380}{270} = 1.4074$$

$$\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} = \frac{270}{380} = 0.7105$$

Thus, we have two sets of indices—Laspeyres and Paasche—each one for the current year and the base year. The geometric mean of the two indices will be

$$P_{01} \times P_{10} = \sqrt{\frac{\cancel{\Sigma p_1 q_0}}{\cancel{\Sigma p_0 q_0}} \times \frac{\cancel{\Sigma p_0 q_0}}{\cancel{\Sigma p_1 q_0}} \times \frac{\cancel{\Sigma p_1 q_1}}{\cancel{\Sigma p_0 q_1}} \times \frac{\cancel{\Sigma p_0 q_1}}{\cancel{\Sigma p_1 q_1}}}$$

We need not even carry out calculations as we can see that these algebraic terms cancel out each other, resulting into $\sqrt{1} = 1$.

It can be seen that if we take geometric mean of Laspeyres and Paasche indices both with the base year and the current year, we find that the time reversal test is fully satisfied. However, if the indices are taken separately and are based on arithmetic mean then these indices do not meet the requirement of the time reversal test.

A factor reversal test should lead to a value index—$P_{01} \times Q_{01} = V_{01}$. The Fisher's ideal index satisfies this test. Taking our earlier example with the data in Table 21.8, we get

$$P_{01} \times Q_{01} = \sqrt{\frac{420}{282} \times \frac{380}{270} \times \frac{270}{282} \times \frac{380}{420}}$$

$$= \sqrt{\frac{380}{282} \times \frac{380}{282}} = \frac{380}{282} = 1.347$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = V_{01}, \text{ which gives the same figure } 1.347$$

It may be noted that neither Laspeyres index nor Paasche index satisfies the factor reversal test.

Circular Test

There is another test with respect to index numbers. This is known as the circular test. It is an extension of the time reversal test. Symbolically, the circular test may be written as

$$P_{01} \times P_{12} \times P_{23} \times \ldots \times P_{n-1,\, n} \times P_{n0} = 1$$

As can be seen, this is a stronger version of the time reversal test. Any index that satisfies the circular test automatically satisfies the time reversal test. If the same method of construction of index is used for each link relative, then the circular test is met. Most people who believe in the time reversal test do not, however, believe in the circular test. An example of circular test is shown on the next page:

Table 21.9 Application of Circular Test

Year	*Price (Rs)*	*Simple Index*	*Chain* Index*	*Link Relative*
0	10	100	100	
1	15	150	150	P_{01}
2	30	300	200	P_{12}
3	45	450	150	P_{23}
4	54	540	120	P_{34}
			18.5	P_{40}

* The method of constructing chain index is discussed a little later.

$$P_{01} \times P_{12} \times P_{23} \times P_{34} \times P_{40}$$
$$= 1.50 \times 2.00 \times 1.50 \times 1.20 \times 0.185$$
$$= 0.999 = 1 \text{ approx.}$$

(As calculations were done up to two decimal places, the figure comes to 0.999, otherwise it would have been 1.)

Alternatively, $P_{01} \times P_{12} = P_{02}$

$= 1.50 \times 2.00 = 3.00$, that is, $P_{02} = 300$, which is true.

It may be noted that most of the formulae used in index numbers do not satisfy this test. Even Fisher's ideal formula does not satisfy this test. Only unweighted or fixed weighted aggregatives or index numbers that use simple geometric mean satisfy this test.

Comments on Reversal Tests

In 1920, Irving Fisher examined as many as 134 formulae, but it should be noted that no formula is 100 per cent accurate. Further, the use of a certain formula will depend largely on the availability of data and convenience. Given below are some broad comments in respect of some formulae, indicating whether they satisfy the time and factor reversal test.

1. *Unweighted average of price relatives:* It does not satisfy the reversibility test unless simple geometric mean is used to average the price relatives. The question of factor reversal test does not arise as the index number is unweighted and factor reversal test relates to weighted index numbers.
2. *Weighted average of price relatives—Laspeyres formula:* It does not satisfy the time reversal test because when the index for the base year is calculated, the quantities of the current year would be taken as weights, which are seldom available when the index is constructed. It also does not satisfy factor reversal test.
3. *Paasche formula:* It does not satisfy both the time and factor reversal tests.
4. *Bowley's formula:* As this formula is the arithmetic average of the Laspeyres' and Paasche's formulae, it does not satisfy the time reversal test as also the factor reversal test.
5. *Marshall-Edgeworth formula:* It too does not satisfy any of the two reversal tests.
6. *Fisher's 'ideal' formula:* This index alone satisfies both the time and factor reversal tests.

Limitations of Fisher's Ideal Index As we have just seen from the foregoing comments that the Fisher's 'ideal' formula satisfies the two tests. However, there are some objections to this index, which are listed on the next page:

1. It is laborious to compute.
2. It needs both current price and quantity data, but it is very difficult to get the current quantity data.
3. One is at a loss to know what the index measures except that it is the geometric mean of Laspeyres and Paasche indices.

It may be mentioned that from a theoretical viewpoint, index numbers covering groups of commodities should have the properties satisfied by price (or quantity) relatives, that is, index numbers for single commodity. An index number that has this property is said to satisfy the time reversal test. In addition, there are factor reversal test and circular test. None of the index numbers meets all these tests. Fisher's 'ideal' index number, as mentioned above, satisfies both the tests, but it fails to satisfy the circular test. Despite its limitations mentioned above, it can be said that among the index numbers currently available, it is the only index that comes very close to the most appropriate index. Hence, it is called the 'ideal' index. From a practical viewpoint, however, there are some other index numbers that are preferable to the 'ideal'index number and as such, they are very frequently used by the statisticians.

21.6 CHAIN BASE INDEX NUMBERS

So far the discussion was confined to fixed-base index numbers excluding the discussion on the circular test. However, at times, we require an index number series over a long period. In such cases, it is necessary to link together a number of separate indices (based on fixed base-years), resulting in a chain index. Since the original raw data over a long period may not be available, this is the only course available through a chain index, which is a mixture of different types of indices. In a chain index number, the base period shifts for each successive index. An example will make this point clear.

Example 21.8 Fixed base indices with 1995 = 100 are given below for the period 1995 to 2000. We are required to construct the chain base index numbers.

Year	1995	1996	1997	1998	1999	2000
Fixed-base Index	100	125	150	200	250	300

Solution Table 21.10 shows the construction of chain base index numbers.

Table 21.10 Construction of Chain Base Index Numbers

Year	*Fixed-base Index*	*Per cent Change*	*Chain Base Index*
1995	100	—	100
1996	125	+ 25	125/100 × 100 = 125
1997	150	+ 50	150/125 × 100 = 120
1998	200	+ 100	200/150 × 100 =133.3
1999	250	+ 150	250/200 × 100 = 125
2000	300	+ 200	300/250 × 100 = 120

It may be noted that we have obtained chain base index numbers by applying the formula:

$$C.B.I. = \frac{F.B.I. \text{ of the current year}}{F.B.I. \text{ of the previous year}} \times 100$$

In fact, the concept of link relatives is the basis for C.B.I. It may be recalled that this concept was used in the construction of seasonal indices in Chapter 19. When there is a single product, the link relatives become the C.B.I. as is the case in Example 21.8. When two or more products are involved, first link relatives are obtained, their average is taken and then the C.B.I. are computed.

A C.B.I. can be changed into a F.B.I. by applying the formula:

$$\text{F.B.I.} = \frac{\text{C.B.I. of the current year} \times \text{F.B.I. of the previous year}}{100}$$

If we apply this formula to the C.B.I. of Table 21.10, we shall get the same figures as are given in Col. (2) of the table.

Advantages of Chain Base Index Numbers

1. The chain base index numbers facilitate the introduction of new items as also the deletion of obsolete items in a smooth manner. Such changes do not require the recalculation of the entire series.
2. In business, often, comparisons are made in the current period with the immediately preceding period rather than any distant period in the past. The chain base index numbers have a major advantage in this respect as the link relatives obtained by such method serve this purpose.

Limitations

1. If the data for any one year are not available, the chain index number for subsequent period cannot be computed.
2. If an error in the computation of any link relative takes place, then such an error gets compounded and the entire series gives a distorted picture.

21.7 SPLICING AND SHIFTING THE BASE OF INDEX NUMBERS

In this section, we shall deal with two aspects—splicing of index numbers and shifting their base-year. Let us first deal with splicing. When two or more overlapping series of index numbers are combined into one series, then this process is known as splicing. We now take an example to explain how it can be done.

Example 21.9 Suppose we have the following two indices, Index A with base-year 1990 = 100 and Index B with base-year 1993 = 100. We are now required to complete the series in Index A.

Table 21.11 Splicing of Two Indices

Year	*Index A*	*Index B*
1990	100	
1991	110	
1992	120	
1993	130	100
1994	x_1	125
1995	x_2	150
1996	x_3	160

Solution In order to complete the series in Index A, we have to find out the corresponding values of x_1, x_2 and x_3. First of all, we calculate the chain indices for 1994, 1995 and 1996. This is shown below:

Chain indices

$$I'_{1993–1994} = 125/100 \times 100 = 125$$
$$I'_{1994–1995} = 150/125 \times 100 = 120$$
$$I'_{1995–1996} = 160/150 \times 100 = 106.67$$

The notation I' denotes that it is a chain index. Further, the first subscript of chain index indicates the base-year and the next subscript indicates the year for which the index is. Thus, $I'_{1994–1995}$ is the chain index for 1995 with the base-year 1994. The notation I indicates that it is a fixed-base index and belongs to series A.

Now, price indices x_1, x_2 and x_3 are obtained as follows:

$$I_{1994} = x_1 = I_{1993} \cdot I'_{1993–1994} \times 1/100$$
$$= 130 \times 125 \times 1/100 = 325/2 = 162.5$$
$$I_{1995} = x_2 = I_{1994} \cdot I'_{1994–1995} \times 1/100$$
$$= 162.5 \times 120 \times 1/100 = 195.0$$
$$I_{1996} = x_3 = I_{1995} \cdot I'_{1995–1996} \times 1/100$$
$$= 195 \times 106.67 \times 1/100 = 4{,}160.13/20 = 208$$

If we use simple proportionality between the two indices, we will get the same result. Thus,

Index for 1994 = $125/100 \times 130 = 162.5$
Index for 1995 = $(150 \times 130)/100 = 195$
Index for 1996 = $(160 \times 130)/100 = 208$

Example 21.10 Let us take another example. Table 21.12 gives indices with two different base-years, 1988 = 100 and 1992 = 100, but in both the cases, series are not complete.

Table 21.12 Splicing of Index Numbers

Year	*Index A (1988 = 100)*	*Index B (1992 = 100)*	*Index C (Index A: Complete Series)*
1988	100		100
1989	110		110
1990	120		120
1991	130		130
1992	150	100	150
1993		120	180
1994		150	225
1995		180	270

It is required to develop one series with 1988 = 100.

Solution Although we can take Index B (1992 = 100) backward from 1991 to 1988 and thus complete the series. However, as our requirement is to bring forth Index A (1988 = 100) up to 1995, we attempt this.

It may be noted that in Table 21.12, the year 1992 has two indices—150 (Index A) and 100 (Index B). We can work out the relationship between the two as follows:

$$\frac{\text{Index A}}{\text{Index B}} \quad \frac{150}{100} = 1.5$$

This means when the Index B is, say x, then the corresponding Index A will be 'x' $\times$ 1.5 = 1.5x. In other words, Index A is 50 per cent higher than Index B. We can now workout the indices in Index A for the years 1993 to 1995 as follows:

Index 1993 = 120 $\times$ 1.5 = 180
Index 1994 = 150 $\times$ 1.5 = 225
Index 1995 = 180 $\times$ 1.5 = 270

The last column of Table 21.12 shows the complete series with 1988 = 100 Index C.

Shifting the Base-Year

At times it is preferable to shift the base of an existing index on account of several reasons. These reasons are: **(1)** to make the base more recent, which will increase its utility; **(2)** to ensure better comparison with some other index that is available on some other base; **(3)** to splice two overlapping indices together; **(4)** to construct a chain index; and **(5)** to facilitate comparison with some date of special interest. For example, this can be the beginning of any five-year plan.

The technique of shifting the base of an index number from one period to another is also a problem of calculating proportions. Let us take an example to illustrate how the base of an index is shifted.

Example 21.11 Suppose we are given an index number series as follows:

Year	1995	1996	1997	1998	1999	2000
Index	100	120	125	150	200	225

We are asked to shift the base from 1995 to 1997.

Solution In order to shift the base from 1995 to 1997, what we have to do is to divide each year's index by the 1997 index and multiply the figure by 100. For example, index for 1998 is 150. Divide it by 1997 index, that is, 150/125. Multiply it by 100—(150/125) $\times$ 100 = 120. In the same manner, we can work out the index for 1999: (200/125) $\times$ 100 = 160 and for 2000: (225/125) $\times$ 100 = 180. We can thus calculate the indices for the remaining years—1995 and 1996. Now, for 1996 the index is 120 for the base-year 1995 = 100. To convert it to the base-year 1997 = 100, we have to divide its index 120 by 125 and multiply by 100. This gives the index as 96. For 1995, the index would be (100/125) $\times$ 100 = 80. These are shown as Index B in Table 21.13.

Table 21.13 Shifting the Base-Year from 1995 to 1997 and 1999

Year	*Index A (Existing Series)*	*Index B (Base 1997)*	*Index C (Base 1999)*
1995	100	80	50
1996	120	96	60
1997	125	100	62.5
1998	150	120	75
1999	200	160	100
2000	225	180	112.5

We can further shift the base of Index B. Suppose that we want to shift the base now from 1997 to 1999. What we have to do is to divide index for each year by 160 and then multiply it by 100. This is shown in the last column of Table 21.13 as Index C. It may be noted that the proportion between the indices of two years in Index A and Index B is maintained in Index C. If we change the base from 1995 (Index A) to 1999 (Index C), we would get the same figures as shown in the last column of Table 21.13.

21.8 DEFLATING PRICES AND INCOMES

The process of adjusting prices and incomes by a price index and expressing them in terms of base-year rupees is called deflating prices and income. An example will illustrate this.

Example 21.12

Year	*Price of Commodity A (Rs)*	*Price of Commodity B (Rs)*	*Price Level*
1990	3	6	100
2000	6	18	200

We have to deflate the price of the two commodities for the year 2000.

Solution

Deflated prices

Year	*Commodity A*	*Commodity B*
2000	3	9

The deflated prices have been calculated by dividing the year 2000 prices by the price level index of 200 and then multiplying the figure by 100. As can be seen from these figures, in 1990, two units of commodity A were equal to one unit of commodity B. But, in the year 2000, three units of commodity A were equal to one unit of commodity B. This means that there was an increase in the real value of commodity B.

Let us take another example. Table 21.14 gives wages and price index in the second and third columns, respectively.

Table 21.14 Deflation of Current Wages

Year (1)	*Wages in Rs (2)*	*Price Index (3)*	*Real Wages in Rs (4)*
1990	2,000	100	2,000
1991	2,200	110	2,000
1992	2,480	120	2,067
1993	2,900	125	2,320
1994	3,000	150	2,000
1995	3,150	160	1,969
1996	3,600	180	2,000

In column (4) of the same table, real wages have been shown. Real wage of a particular year has been arrived at by dividing the actual wage of that year by the corresponding price index and then

multiplying it by 100. For example, take the wage of 1991, which was Rs 2,200—Rs 200 more than what it was in 1990. This figure has been divided by the relevant price index (110) and then multiplied by 100. Thus,

Real wage = (Rs 2,200/110) × 100 = Rs 2,000.

This deflated wage shows that although in year 1991 the wages increased as compared to the preceding year, in real terms they remained at the same level. In the same manner, wages have been deflated by the corresponding price indices for the remaining years 1992 to 1996. It should be noted that the current wages increased from Rs 2000 in 1990 to Rs 3600 in 1996, indicating an increase of 80 per cent. But on account of the same percentage increase in the price index, the real wages in 1996 remained at the same level of Rs 2000. Thus, the importance of price index numbers becomes chear in such cases.

21.9 QUANTITY INDEX NUMBERS

So far our discussion was confined to price indices. We now turn to quantity indices. As the index of industrial production is most common among quantity indices, we confine our discussion to this index alone.

Symbolically, an index of industrial production is

$$Q_{0n} = \frac{\Sigma q_n p_0}{\Sigma q_0 p_0}$$

where Q is an index of industrial production, q is quantity of a particular product produced, p is price of that product per unit of measurement and subscripts 0 and n stand for the base-year and the current year, respectively. This index can also be written as

$$Q_{0n} = \frac{\Sigma (q_n / q_0) w}{\Sigma w}$$

where w is the weight for each commodity.

Taking the weight as the value of output in the base-year, that is, p_0q_0, we find that

$$\frac{\Sigma (q_n/q_0)\, p_0q_0}{\Sigma p_0q_0} = \frac{\Sigma q_n p_0}{\Sigma q_0 p_0}, \text{ as shown earlier}$$

Let us take an example to illustrate the calculation of a quantity index. This is done in Table 21.15.

Table 21.15 Calculations for Quantity Index

Item	p_0	q_0	p_0q_0	*Weight*	q_1	q_2	q_1p_0	q_1/q_0	q_2p_0	q_2/q_0
A	0.50	100	50	25	200	300	100	2.0	150	3
B	0.30	150	45	22.5	270	300	81	1.8	90	2
C	0.15	500	75	37.5	700	750	105	1.4	112.5	1.5
D	0.10	300	30	15	450	450	45	1.5	45	1.5
		Total	**200**	**100**			**331**		**397.5**	

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100 = \frac{331}{200} \times 100 = 165.5$$

Alternative method

$$Q_{01} = \frac{\Sigma(q_1/q_0)w}{\Sigma w} \times 100$$

$$= \frac{(2.0 \times 25) + (1.8 \times 22.5) + (1.4 \times 37.5) + (1.5 \times 15)}{100} \times 100$$

$$= \frac{50.00 + 40.50 + 52.50 + 22.50}{100} \times 100$$

$$= \frac{165.50}{100} \times 100 = 165.5$$

Similarly, quantity index for year 2 will be

$$Q_{02} = \frac{\Sigma q_2 p_0}{\Sigma q_0 p_0} \times 100 = \frac{397.5}{200} \times 100 = 198.75$$

Alternative method

$$Q_{02} = \frac{\Sigma(q_2/q_0)w}{\Sigma w} \times 100$$

$$= \frac{(3.0 \times 25) + (2.0 \times 22.5) + (1.5 \times 37.5) + (1.5 \times 15)}{100} \times 100$$

$$= \frac{75.00 + 45.00 + 56.25 + 22.50}{100} \times 100$$

$$= \frac{198.75}{100} \times 100 = 198.75$$

Uses of an Index of Industrial Production Index numbers compress many facts into a few simple figures and, in conjunction with other data, their use in economic analysis is in summarising past developments, forecasting future trends and making decision policy. Accordingly, an index of industrial production can be put to several uses both in micro and macro analyses. Some of the important uses are outlined below:

1. Individual factories may compare changes in their output with the production index of their own industry.
2. Again, such factories can compare the production index in conjunction with other series, for example, employment, prices, wages, and so on. For instance, a factory may compute its own productivity index and compare it with that of the industry of which it is a part.
3. Similarly, inter-industry variations in the level of production and productivity can also be compared and analysed.
4. In macro-economics, the index of production can be compared with the corresponding changes in national income, population, foreign trade, prices and similar other aggregates.

Limitation of Index of Industrial Production An important *limitation* of index numbers of production relates to quality aspects. An index of production fails to take into account any changes in quality of products. Failure to take account of these—especially when the qualitative changes have been considerable—presumably leads to distorted measures of quantity. Many products over the years have been improved as a result of continuing research and development. To the extent these trends have developed, production indices will show a downward bias, which are not, however, capable of being measured statistically.

21.10 VALUE INDEX NUMBERS

A value index number shows a change in the value for the current year as compared to the value in the base year. Since value is the product of price and quantity, as such a value index can be written as

$$V_{01} = \Sigma p_1 q_1 / \Sigma p_0 q_0$$

Alternatively, if we have price index and quantity index pertaining to the same data and with the same base year, then the value index will be

$$V_{o1} = P_{01} \times Q_{01}$$

An example will bring home this point clearly.

Example 21.13 The following table gives prices and quantities for the base year and the current year for three commodities:

Commodity	p_0 *Rs*	q_0 *Units*	p_1 *Rs*	q_1 *Units*
A	10	15	15	20
B	20	20	30	30
C	30	25	45	40

Calculate the value index for the current year.

Solution

Worksheet

Commodity	p_0q_0 *Rs*	p_0q_1 *Rs*	p_1q_0 *Rs*	p_1q_1 *Rs*
A	150	200	225	300
B	400	600	600	900
C	750	1200	1125	1800
Total	**1300**	**2000**	**1950**	**3000**

$$V_{o1} = (\Sigma p_1 q_1 / \Sigma p_0 q_0) \times 100$$
$$= (3000/1300) \times 100$$
$$= 230.76$$

Alternative Method

$$V_{01} = P_{01} \times Q_{01}$$

$$= \left[\frac{\Sigma p_1 q_0}{\Sigma p_0 q_o} \times \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0}\right] \times 100$$

$$= \left[\frac{1950}{1300} \times \frac{2000}{1300}\right] \times 100$$

$= [1.5 \times 1.5384] \times 100$

$= 230.76$ (Same as above)

It may be noted that the value index is not so commonly used as the price and quantity index numbers. Even so, the concept of value index should be clear that it is derived from the product of two indices, viz. price and quantity.

Additional Examples

Example 21.14 Calculate index number of prices for 1995 on the basis of 1990 from the data given below:

Commodity	*Weight*	*Price per unit 1990 (Rs)*	*Price per unit 1995 (Rs)*
A	40	16	20
B	25	40	50
C	20	12	15
D	15	2	3

If the weights of commodities A, B, C and D are increased in the ratio 1:2:3:4, what will be the increase in index number?

Solution

Commodity	*Weight*	*Price per unit 1990 (Rs)*	*Price per unit 1995 (Rs)*	*Price Relative*
A	40	16	20	20/16 = 1.25
B	25	40	50	50/40 = 1.25
C	20	12	15	15/12 = 1.25
D	15	2	3	3/2 = 1.50

Now, we have to multiply each price relative by the corresponding weight.

$$\text{Price Index} = \frac{(1.25 \times 40) + (1.25 \times 25) + (1.25 \times 20) + (1.50 \times 15)}{40 + 25 + 20 + 15} \times 100$$

$$= \frac{50 + 31.25 + 25 + 22.5}{100} \times 100$$

$$= \frac{128.75}{100} \times 100 = 128.75$$

When the weights are increased in the ratio 1: 2: 3: 4 then the price relatives will have weights A = 40 B = 80 C = 120 D = 160.

$$\text{Price Index} = \frac{(1.25 \times 40) + (1.25 \times 80) + (1.25 \times 120) + (1.50 \times 160)}{40 + 80 + 120 + 160}$$

$$= \frac{50 + 100 + 150 + 240}{400} \times 100$$

$$= \frac{540}{400} \times 100 = 135$$

Thus, the index will increase from 128.75 to 135, an increase of 6.25.

Example 21.15 Calculate **(a)** a weighted average of price relatives and **(b)** a weighted aggregate price index, both for year 2 (based on year 1), for the following commodities:

Commodity	*Weight*	*Price (Rs)*	
		Year 1	*Year 2*
A	5	215	210
B	10	250	275
C	4	1,100	1,300
D	8	950	950
E	13	170	200

Solution Price ralatives

(a) $\frac{210}{215} = 0.98$

$\frac{275}{250} = 1.1$

$\frac{1300}{1100} = 1.18$

$\frac{950}{950} = 1$

$\frac{200}{170} = 1.18$

Weighted average of price relatives

$$= \frac{(0.98 \times 5) + (1.1 \times 10) + (1.18 \times 4) + (1 \times 8) + (1.18 \times 13)}{40}$$

$$= \frac{4.9 + 11 + 4.72 + 8 + 15.34}{40}$$

$$= \frac{43.96}{40}$$

$= 1.099 \times 100 = 109.9$ or 110 approx.

(b) Weighted aggregate price index

$\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0}$, Since quantities are not given, we will use weights as quantities. The index would be $\frac{\Sigma p_1 w}{\Sigma p_0 w}$.

Hence $\frac{(210 \times 5) + (275 \times 10) + (1300 \times 4) + (950 \times 8) + (200 \times 13)}{(215 \times 5) + (250 \times 10) + (1100 \times 4) + (950 \times 8) + (170 \times 13)}$

$$= \frac{1050 + 2750 + 5200 + 7600 + 2600}{1075 + 2500 + 4400 + 7600 + 2210}$$

$$= \frac{19200}{17785} \times 100 = 1.079 \times 100 = 107.9 \text{ or } 108 \text{ approx.}$$

Example 21.16 Calculate price index number for the year 1996 with 1986 as the base-year from the following data using:
(i) Laspeyres index, **(ii)** Paasche index, and **(iii)** Fisher's index

		1986		*1996*	
Commodity	*Unit*	*Price (Rs)*	*Value (Rs)*	*Qty. Consumed*	*Value (Rs)*
A	kg	10	1,500	160	1,760
B	kg	12	1,080	100	1,300
C	Metre	15	900	60	960
D	Packets	9	450	40	480

Solution

Calculation of Laspeyres, Paasche and Fisher's Price Index Numbers

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_1q_1	p_0q_1
A	10	150	11	160	1500	1650	1760	1600
B	12	90	13	100	1080	1170	1300	1200
C	15	60	16	60	900	960	960	900
D	9	50	12	40	450	600	480	360
					3930	4380	4500	4060

Laspeyres Index Number:

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} = \frac{4380}{3930} \times 100 = 111.45$$

Paasche's Index Number:

$$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = \frac{4500}{4060} \times 100 = 110.84$$

Fisher's Index Number:

$$\sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

$$= \sqrt{1.1145 \times 1.1084} \times 100$$

$$= 1.1114 \times 100 = 111.14$$

Example 21.17 Compute Fisher's 'ideal' index from the following data and check whether the time reversal test is satisfied.

Commodities	*Base-year*		*Current year*	
	Unit price (Rs)	*Quantity (kg)*	*Unit price (Rs)*	*Quantity (kg)*
A	2	7	6	6
B	3	6	2	3
C	4	5	8	5
D	5	4	2	4

Solution

Commodities	*Base-year*		*Current year*					
	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	2	7	6	6	42	14	36	12
B	3	6	2	3	12	18	6	9
C	4	5	8	5	40	20	40	20
D	5	4	2	4	8	20	8	20
					102	72	90	61

$$p_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} = \sqrt{\frac{102}{72} \times \frac{90}{61}}$$

$$p_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{61}{90} \times \frac{72}{102}}$$

$$p_{01} \times p_{10} = \sqrt{\frac{102}{72} \times \frac{90}{61} \times \frac{61}{90} \times \frac{72}{102}} = 1$$

It will be seen that numerator values cancel out the denominator values, resulting into 1. Hence, Fisher's 'ideal' index satisfies the time reversal test.

Example 21.18 The price relatives and weights of a set of commodities are given in the following table:

Commodity	C_1	C_2	C_3	C_4
Price relative	120	127	125	119
Weight	$2W_1$	W_2	W_1	$W_2 + 3$

If the index for the set is 122 and the sum of the weights is 40, find W_1 and W_2.

Solution

Total weight = 40

or $2W_1 + W_2 + W_1 + W_2 + 3 = 40$

or $3W_1 + 2W_2 = 40 - 3 = 37$ (1)

Now, we multiply price relative for each item by the corresponding weight,

$$\frac{(120 \times 2W_1) + (127 \times W_2) + (125 \times W_1) + (119 \times (W_2 + 3))}{40} = 122$$

or $$\frac{240W_1 + 127W_2 + 125W_1 + 119W_2 + 357}{40} = 122$$

or $$\frac{365W_1 + 246W_2 + 357}{40} = 122$$

or $365W_1 + 246W_2 + 357 = 4880$

or $365W_1 + 246W_2 = 4880 - 357$

or $365W_1 + 246W_2 = 4523$ (2)

$3W_1 + 2W_2 = 37$ (1)

Multiplying (1) by 123, we get

$369W_1 + 246W_2 = 4551$ (3)

$365W_1 + 246W_2 = 4523$

$4W_1 = 28$

$\therefore \quad W_1 = 28/4 = 7$

Substituting the value of $W_1 = 7$ in equation (1), we get

$(3 \times 7) + 2W_2 = 37$

or $2W_2 = 37 - 21 = 16$ Hence, $W_1 = 7$

or $W_2 = 16/2 = 8$ $W_2 = 8$

Example 21.19 Calculate the chain base index number from the following data:

Commodity	*Price in Rs*				
	1991	*1992*	*1993*	*1994*	*1995*
A	2	3	4	2	7
B	3	6	9	4	3
C	4	12	20	8	16
D	5	7	18	11	22

Solution Taking prices for preceding year as 100, price relatives for 1992 to 1995 have been calculated. These are given below.

Commodity	*1991*	*1992*	*1993*	*1994*	*1995*
A	2	150	133	50	350
B	3	200	150	44	75
C	4	300	167	40	200
D	5	140	257	61	200
	Total	**790**	**707**	**195**	**825**

Average L.R. for 1992: $\frac{790}{400} \times 100 = 197.5$. This is the chain index for 1992. Calculations for other years are shown on the next page.

Average L.R.		197.5	176.75	48.75	206.25
1991	100				
Chain Index					
1992		197.5			

1993	$\frac{176.75 \times 197.5}{100} = 349.08$	349.1
1994	$\frac{48.75 \times 349.08}{100} = 170.18$	170.2
1995	$\frac{206.25 \times 170.18}{100} = 351.0$	351.0

Example 21.20 Given below are two sets of indices. For the purpose of continuity of records, you are required to construct a combined series with the year 1983 as the base.

Year	*I set (Price relatives)*	*II set (Link relatives)*
1980	100	
1981	120	
1982	125	
1983	150	100
1984		110
1985		120
1986		195
1987		105

Solution We are required to construct a combined series with the year 1983 as the base.

It may be noted that the year 1983 has two indices—one in I set as 150 and the other in II set as 100.

First, we have to make 1983 as the base-year = 100. The ratio between the two indices of 1983 is 100/150 or 2/3. It we multiply the indices of earlier years by this ratio, we will be able to convert those indices based on 1983 = 100.

For 1982: $125 \times \frac{2}{3} = \frac{250}{3} = 83.3$

For 1981: $120 \times \frac{2}{3} = \frac{240}{3} = 80$

For 1980: $100 \times \frac{2}{3} = \frac{200}{3} = 66.7$

Now, the complete series will be as follows:

1980	1981	1982	1983	1984	1985	1986	1987
66.7	80	83.3	100	110	120	195	105

Example 21.21 Assuming that all the products can be assigned equal weights, calculate the chain base index numbers for the years 1996 to 2000 on the basis of the following price relatives:

(*Note*: Price relative = $(p_1/p_0) \times 100$)

Year	*A*	*B*	*C*	*D*	*E*
1996	100	100	100	100	100
1997	90	525	134	118	133
1998	81	61	60	115	125
1999	112	200	80	93	140
2000	122	66	150	86	86

Solution We are given price relatives, which have to be changed into chain-base index numbers. The following table shows the calculations.

Calculation of Chain Base Index Numbers

	Products—Price Relatives						
Years	*A*	*B*	*C*	*D*	*E*	*Total of L.R.*	*Average*
1996	100	100	100	100	100	500	100
1997	90	525	134	118	133	1000	200
1998	81	61	60	115	125	442	88.4
1999	112	200	80	93	140	625	125
2000	122	66	150	86	86	510	102

			Chain Base Index
1996	100	$\frac{\overset{100}{\cancel{100}} \times 100}{\cancel{100}} = 100$	100
1997	200	$\frac{200 \times 100}{100} = 200$	200
1998	88.4	$\frac{88.4 \times 200}{100} = 176.8$	176.8
1999	125	$\frac{125 \times 176.8}{100} = 221$	221
2000	102	$\frac{102 \times 221}{100} = 225.42$	225.4

Example 21.22 Splice the two indices with base-year 1990 = 100.

Year	*Index*	
1990	100	
1991	110	
1992	120	
1993	130	(Base 1994 = 100)
1994	150	100
1995	x_1	120
1996	x_2	150
1997	x_3	180

Solution We are required to splice the two indices with base-year 1990 = 100. This means, we have to find the indices for 1995, 1996 and 1997 with 1990 base-year = 100

There is a link as 1994 with 1990 base year is 150 and it is the base-year = 100 for the subsequent three indices. We apply the rule of proportion as follows.

1995 Index $\frac{120}{100} \times 150 = 180$

1996	Index	$\frac{150}{100} \times 150 = 225$
1997	Index	$\frac{180}{100} \times 150 = 270$

Hence the series after splicing will be

1990	100
1991	110
1992	120
1993	130
1994	150
1995	180
1996	225
1997	270

Example 21.23 Given below are three indices A, B and C for different years. With the Index B and Index C, complete the series for Index A.

Year	*Index A*	*Index B*	*Index C*
1994	100		
1995	120		
1996	140	100	
1997		115	
1998		130	
1999		150	100
2000			120

Solution It will be seen that for two years 1996 and 1999, we have two indices. On the basis of this information, we can transform Indices B and C by applying the rule of proportionality.

Let us take first Index B.

For 1997 $\frac{115}{100} \times 140 = 161$

For 1998 $\frac{130}{100} \times 140 = 182$

For 1999 $\frac{150}{100} \times 140 = 210$

Now, we take Index C.

For 2000 $\frac{120}{100} \times 210 = 252$

Hence, Index A for 1994 to 2000 is as given below:

1994	1995	1996	1997	1998	1999	2000
100	120	140	161	182	210	252

Example 21.24 A textile worker in the city of Ahmedabad earns Rs 3,000 per month. The Cost of Living Index for January 2000 is given as 160. Using the following data, find out the amount spent on (i) food and (ii) rent.

Group	*Expenditure (Rs)*	*Group index*
Food	?	190
Clothing	800	181
Rent	?	140
Fuel & lighting	400	118
Miscellaneous	300	101

Solution

Food x, Rent y } $x + y = 3000 - 800 - 400 - 300$ or $x + y = 1500$

Group Index

$$\frac{(190 \times x) + (181 \times 800) + (140 \times y) + (118 \times 400) + (101 \times 300)}{x + 800 + y + 400 + 300} = 160$$

or $$\frac{190x + 144800 + 140y + 47200 + 30300}{x + y + 1500} = 160$$

or $$\frac{190x + 140y + 222300}{x + y + 1500} = 160$$

or $$190x + 140y + 222300 = 160x + 160y + 240000$$

or $$190x - 160x + 140y - 160y = 240000 - 222300$$

or $$30x - 20y = 17700$$

Dividing it by 10, we get

$$3x - 2y = 1770 \quad (2)$$

Now. we have two simultaneous equations,

$$x + y = 1500 \quad (1)$$

$$3x - 2y = 1770 \quad (2)$$

Multiplying (1) by 2 and by addition, we get

$$2x + 2y = 3000 \quad (3)$$

$$3x - 2y = 1770 \quad (2)$$

$$5x = 4770$$

$\therefore$ $$x = 954$$

Substituting the value of $x = 954$ in equation (1), we get

$$954 + y = 1500$$

$\therefore$ $$y = 1500 - 954 = 546$$

Hence, the amount spent on food: Rs 954 and on rent: Rs 546.

Example 21.25 You are given the following price index numbers with base year 1995. Calculate the new price index numbers with base year 2000.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Price Index Number	100	105	110	118	120	125	135	148	152	160

Solution In order to shift the base from 1995 to 2000, what we have to do is each index is divided by 125 and multiplied by 100. To further simplify calculations, we can multiply each price index number by the factor $\frac{100}{125} = 0.8$. Working out this way, the new series with base year 2000 for the years 1995 to 2004 will be as follows:

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Price Index Number	80	84	88	94	96	100	108	118	122	128

Example 21.26 The employees of a textile mill have approached the management for a raise in their wages. Their contention is that while prices have increased enormously, their wages have risen marginally and, as such, their "real" wages have declined. In support of their case, they have presented the following data to the management.

Year	1999	2000	2001	2002	2003	2004
Average monthly wage (Rs)	3800	3910	3990	4050	4095	5120
Price Index	110	125	155	170	189	216

You are asked to calculate the "real" wages. Also compute the additional amount necessary in 2004 to provide buying power at par with that in 1999.

Solution

Years	1999	2000	2001	2002	2003	2004
"Real" wages	3800	3441	2832	2621	2383	2607

The "real" wages have been arrived at as follows:

$$\frac{\text{Wages for the current year}}{\text{Price Index of the current year}} \times \text{Price Index for 1999}$$

Take, for example, the year 2004

$$\text{Real wage for 2004} = \frac{5120}{216} \times 110 = 2607$$

As the deficiency in real wage for 2004 is Rs 3800 – 2607 = Rs 1193. Hence, this amount is necessary in 2004 to bring it at par with the year 1999.

Example 21.27 Construct the quantity index numbers for 1991 and 1992 with base as 1990, for the following data:

Commodities	*Quantity produced*			*Price per unit in*
	1990	*1991*	*1992*	*1990 (Rs)*
A	20	24	30	40
B	15	18	20	60
C	10	15	25	50

Solution

Construction of the Quantity Index Number

Commodities	*1990*			*1991*		*1992*	
	q_0	p_0	p_0q_0	q_1	q_1p_0	q_2	q_2p_0
A	20	40	800	24	960	30	1200
B	15	60	900	18	1080	20	1200
C	10	50	500	15	750	25	1250
			2200		2790		3650

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$$

$$= \frac{2790}{2200} \times 100 = 1.268 \times 100 = 126.8$$

$$Q_{02} = \frac{\Sigma q_2 p_0}{\Sigma q_0 p_0} \times 100$$

$$= \frac{3650}{2200} \times 100 = 1.659 \times 100 = 165.9$$

Example 21.28 The price and quantity data for the years 2005 and 2006 are given below. You are required to construct index numbers of price using the following methods.

1. Laspeyres' method
2. Paasche's method
3. Bowley's method
4. Fisher's method
5. Marshall-Edgeworth's method

Commodity	*2005*		*2006*	
	Price	*Quantity*	*Price*	*Quantity*
A	3	5	6	7
B	4	6	8	8
C	6	5	9	10
D	2	10	8	12
E	5	15	10	20

Solution

Laspeyres' Method: The formula for Laspeyres' price index is

$$\text{Index for year 1: } \frac{\Sigma p_i q_o}{\Sigma p_o q_o} \times 100$$

As such, we have to get the value of $\Sigma p_1 q_0$ and $\Sigma p_0 q_0$.

Now, $\Sigma p_1 q_0 = (6 \times 5) + (8 \times 6) + (9 \times 5) + (8 \times 10) + (10 \times 15)$

$= 30 + 48 + 45 + 80 + 150$

$= 353$

And, $\Sigma p_0 q_0 = (3 \times 5) + (4 \times 6) + (6 \times 5) + (2 \times 10) + (5 \times 15)$

$= 15 + 24 + 30 + 20 + 75$

$= 164$

Hence, Laspeyres' price index for year 1 is

$(353/164) \times 100$

$= 2.1524 \times 100$

$= 215.24$

Passche's Method: The formula for Paasche's price index is,

$$\text{Index} = \frac{\Sigma p_1 q_1}{\Sigma p_o q_1}$$

Now, $\Sigma p_1 q_1 = (6 \times 7) + (8 \times 8) + (9 \times 10) + (8 \times 12) + (10 \times 20)$
$= 42 + 64 + 90 + 96 + 200$
$= 492$

And, $\Sigma p_0 q_1 = (3 \times 7) + (4 \times 8) + (6 \times 10) + (2 \times 12) + (5 \times 20)$
$= 21 + 32 + 60 + 24 + 100$
$= 237$

Hence, Paasche's price index for year 1 is
$= (492/237) \times 100$
$= 2.0759 \times 100$
$= 207.59$

Bowley's Method: The formula for Bowley's price index is

$$\left\{\left(\frac{\Sigma p_1 q_o}{\Sigma p_o q_o} + \frac{\Sigma p_1 q_1}{\Sigma p_o q_1}\right) \Big/ 2\right\} \times 100$$

Substituting the values arrived at earlier in the above formula, we get

$$\left[\left(\frac{353}{164} + \frac{492}{237}\right) \Big/ 2\right] \times 100$$

$= [(2.1524 + 2.0759)/2] \times 100$
$= (4.2283/2) \times 100$
$= 2.11415 \times 100$
$= 211.415$

Fisher's Method: The formula for Fisher's price index is,

$$\sqrt{\frac{\Sigma p_1 q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_1 q_1}{\Sigma p_o q_1}} \times 100$$

$$= \sqrt{\frac{353}{164} \times \frac{492}{237}} \times 100$$

$$= \sqrt{2.1524 \times 2.0759} \times 100$$

$$= \sqrt{4.46816716} \times 100$$

$= 2.113804 \times 100$
$= 211.3804$

Marshall-Edgeworth Method : The formula for price index, according to this method, is

$$= \frac{\Sigma p_1 (q_o + q_1)}{\Sigma p_o (q_o + q_1)} \times 100$$

The above formula can be written as,

$$= \left(\frac{\Sigma p_1 q_o + \Sigma p_1 q_1}{\Sigma p_o q_o + \Sigma p_o q_1}\right) \times 100$$

Substituting the values, we get

$$= \left(\frac{353 + 492}{164 + 237}\right) \times 100$$

$$= \frac{845}{401} \times 100$$
$$= 2.10723 \times 100$$
$$= 210.723$$

Let us compare the relative values of these indices.

Laspeyres Index	215.24
Paasche Index	207.59
Bowley Index	211.42
Fisher's Index	211.38
M-E Index	210.72

It can be seen that different methods give different price indices. The last three methods (Bowley's, Fisher's and Marshall-Edgeworth's) give results very close to each other. However, in practice, Laspeyers Index is most frequently used, followed by Paasche Index and Fisher's Ideal Index.

21.11 CAUTION IN USING INDEX NUMBERS

Index numbers are an extremely useful device that enables us to have up-to-date information on prices, production and other related aspects. That is why, they are being used increasingly by the government, business and industry. However, if the index numbers are defective on account of one or more reasons, their interpretation will lead to drawing of wrong inferences from them. Even if index numbers are constructed on sound statistical principles, but we interpret them in the wrong manner, our conclusions will be wrong. Any major decision based on misinterpretation of an index number will result in loss to the industry. It is, therefore, very necessary to ensure that index numbers are not only properly constructed but also properly interpreted.

As far as construction of index numbers is concerned, it is the domain of the statistician who has to ensure that the index numbers are based on sound statistical principles. As regards proper use of index numbers, it is the user who should ensure that he interprets index numbers properly. This is possible only if he is aware of the limitations of index numbers, so that he can guard himself against certain pitfalls that are likely to arise at certain places. As such, we give here some major limitations of index numbers.

Limitations of Index Numbers

1. Index numbers are based on sample data. In case sample size is extremely limited and its selection is faulty in the sense that the sample units have not been selected randomly, index numbers will give wrong figures.
2. At times, index numbers can be manipulated by those who are in authority. This is purposely done to support their viewpoint. While it may not be possible for the common man to know such manipulations, at least he can find out whether the base year is abnormal. A normal base year free from major or abrupt fluctuation needs to be used in index number construction.
3. A number of formulae can be used in index number construction. These will give different results. One who is using the index should know a little more about different formulae and their effect on the magnitude of the index.
4. Index numbers with the same base and items are useful for a short period. This is because of the changes in the consumption pattern of people in the current period. One has, therefore, to ensure that index does not use a very remote year as the base.

5. One who is interpreting an index must be familiar with general aspects of the economy and the factors relevant in this regard. If he possesses a reasonably good background of economic conditions, he should be in a position to interpret the indices properly.
6. As we know, our indices are of prices and quantities. The question is: does our index reflect a change in the quality of a product or item? This aspect has also been covered when we discussed the index of industrial production.
7. Apart from quality changes, there are other aspects, that are pertinent while we are interpreting index numbers. We have to ask whether the weights assigned to different items are appropriate. Although a layman cannot have any data to examine this issue he, at least, can examine whether the weights are based on a remote year in the past. As we have seen earlier, when the relative importance of items is undergoing changes, we have just to ensure that the index that we are interpreting is not based on weights, which have now ceased to be relevant.

GLOSSARY

Chain index number	An index that links up different fixed-base indices to obtain a long comparable series.
Circular test	It is an extension of time reversal test. It lays down that an index must satisfy the time reversal test through a number of intermediate years.
Composite index number	Index that compares the combined change in two or more items with respect to price or quantity over two or more periods of time.
Consumer price index	An index that measures the price changes in a representative set of consumer goods over time.
Deflating price and income	The process of adjusting prices and incomes by a price index and expressing them in terms of base-year rupees.
Factor reversal test	It requires that the product of price and quantity must show the change of values from year '0' to year n.
Fixed-weight aggregates method	A method wherein the weights of an aggregate index are based on quantities consumed during a specified period.
Index number	A ratio that measures how much a variable such as price or production has changed over time. It can be used for comparison over two or more locations as well.
Index of industrial production	An index that measures the quantitative changes in industrial output over time. It covers manufacturing, mining and utilities.
Laspeyres index	An index that uses the quantities consumed during the base-year as weights.
Paasche index	An index that uses the quantities consumed in the current year as weights.

Percentage relative	Relative change in two values expressed in percentage terms.
Price index	An index that measures the change in prices between two or more periods of time.
Quantity index	An index that measures the change in quantity (usually production) between two or more periods of time.
Simple index number	Index that compares the change in the price or quantity of a single item between two or more periods of time.
Splicing of index numbers	It means combining two or more overlapping series of index numbers into one continuous series.
Time reversal test	It requires that if an index for year n based on year '0' and another index for year '0' based on year n are multiplied, then the result must be 1.
Unweighted aggregates index	An index that gives equal importance to all the values considered. Unlike other indices, it does not assign weights to the items it covers.
Unweighted average of relatives method	A method wherein the total of price relatives of different products over time is divided by the number of products, without assigning any weights to the products.
Value index	An index that measures the change in value (i.e. price and quantity taken together) between two or more periods of time.
Weighted aggregates index	An index that assigns weights to different values on the basis of their relative importance.
Weighted average of relatives method	In contrast to unweighted average of relatives method, it assigns weights to price relatives on the basis of their relative importance.

LIST OF FORMULAE

1. Simple aggregative price index:

$$\frac{\Sigma p_i}{\Sigma p_0} \times 100$$

where p_i stands for price in the i^{th} year and p_0 for price in the base-year for a specified item.

2. Simple aggregative quantity index:

$$\frac{\Sigma q_i}{\Sigma q_0} \times 100$$

where q_i stands for quantity in the i^{th} year and q_0 for quantity in the base-year for a specified item.

3. Weighted aggregative price index:

$$\frac{\Sigma p_i q_0}{\Sigma p_0 q_0} \times 100$$

This is also known as the Laspeyres price index, where the base-year quantities are used as weights.

4. Paasche index:

$$\frac{\Sigma p_i q_i}{\Sigma p_0 q_i} \times 100$$

This is also a weighted aggregative price index, where the current year quantities are used as weights.

5. Unweighted average of relatives price index:

$$\frac{\Sigma[(p_i / p_0) \times 100]}{n}$$

where n stands for the total number of items covered in the index. Ratios of current price to base-year price are multiplied by 100 and then an unweighted average is taken.

6. Weighted average of relative price index:

$$\frac{\Sigma[(p_i / p_0 \times 100)(p_n q_n)]}{\Sigma p_n q_n}$$

In this price index, weights are assigned on the basis of total expenditure on the product in the current year.

7. Weighted average of relative quantity index:

$$\frac{\Sigma[(q_i / q_0 \times 100)(q_n p_n)]}{\Sigma q_n p_n}$$

This is the quantity index where the weights to different items are assigned on the basis of their current expenditure.

8. Fisher's formula:

$$\sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

This index is the square root of the products of Laspeyres and Paasche indices multiplied by 100.

9. Bowley's formula:

$$\left\{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right) \Big/ 2\right\} \times 100$$

This index is an average of base-year and current year price indices, multiplied by 100.

10. Marshall-Edgeworth's formula:

$$\frac{\Sigma p_1 (q_0 + q_1)}{\Sigma p_0 (q_0 + q_1)} \times 100$$

In this price index, the weights are the total quantities consumed in the base-year and the current year.

11. Time reversal test:

$$P_{01} \times P_{10} = 1$$

This shows that the price index for the current year compared to the base-year and the base-year price index with the current year as the base, if multiplied, should result 1.

12. Factor reversal test:

$$P_{01} \times Q_{01} = V_{01}, \text{ i.e. } \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

This shows that if price and quantity indices of the current year with the same base-year are multiplied, then the resultant should be the value index for the current year.

13. Circular test:

$P_{01} \times P_{12} = P_{02}$ (This can be extended to n years)

This is an extension of the time reversal test.

14. Value Index Number

(i) $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$

(ii) $P_{01} \times Q_{01}$

as shown in No. 12 above.

QUESTIONS

21.1 Given below are twelve statements. Indicate in each case whether it is true or false:

(a) The index number for a base-year is always 100.
(b) Selection of an inappropriate base-year will give wrong results.
(c) Index numbers on prices or quantities cannot be used to measure differences in several locations.
(d) A major limitation of Laspeyres indices is that they are not comparable to one another.
(e) As compared to Paasche index, the Laspeyres index is more frequently used.
(f) An unweighted aggregates index is the simplest form of a composite index.
(g) The validity of an index number with the same base-year for a prolonged period gets considerably reduced.
(h) Index numbers can only be used to measure changes in prices and production.
(i) The consumer price index number as well as the index of industrial production are value indices.
(j) An index number need not be confined to a fixed base-year.
(k) A value index gives us the combined effect of variations in prices and quantities.
(l) Index numbers are a simple device and as such are very frequently used both by business and government.

Multiple Choice Questions (21.2 to 21.14)

21.2 If one wants to measure changes in total monetary worth, then the right choice should be
(a) a quantity index (b) a value index
(c) a price index (d) none of these

21.3 Commodities which show considerable price fluctuations could be best measured by a
(a) Quantity index (b) Value index
(c) Price index (d) None of these

21.4 Which of the following can be considered as the "normal" base period?
(a) The most recent period for which the data are available.
(b) It remained free from inflation or deflation of prices.
(c) It is an average of two or three consecutive periods.
(d) It is neither at the peak nor at the trough of a fluctuation.

21.5 Which of the following is used as weights in a quantity index?
(a) Average of quantities (b) Percentages of total quantity
(c) Prices (d) None of these

21.6 The aggregate price index that uses base-year quantities as base is
(a) Paasche's index (b) Laspeyres index
(c) Fisher's index (d) None of these

21.7 The two price indices P_{01} and P_{10} when multiplied, satisfy the following test.
(a) Factor reversal test (b) Time reversal test
(c) Circular test (d) None of these

21.8 A time reversal test is satisfied if we use
(a) mode (b) arithmetic mean
(c) geometric mean (d) none of these

21.9 Which of the following index satisfies the circular test?
(a) Laspeyres index (b) Paasche's index
(c) Fisher's index (d) Bowley's index

21.10 A shift in the base-year is possible without changing the quantities used as weights provided we use
(a) The Laspeyres method (b) The Paasche method
(c) The weighted aggregates method (d) None of these

21.11 For the three years 1995, 1996 and 1997 the fixed base index numbers were 100, 125 and 150, respectively. What would be the chain base index for 1997?
(a) 160 (b) 130 (c) 120
(d) 140 (e) None of these

21.12 Which of the following is a quantity index?

(a) $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$ (b) $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$ (c) $\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times 100$ (d) $\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times 100$

21.13 Which one of the following is the Paasche's price index?

(a) $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$ (b) $\frac{\Sigma p_1 q_1}{\Sigma p_1 q_0} \times 100$ (c) $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$

(d) $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$ (e) None of these

21.14 Which one of the following is the circular test?
(a) $P_{01} \times Q_{01} = V_{01}$ (b) $P_{01} \times P_{12} = P_{02}$
(c) $P_{01} \times P_{10} = 1$ (d) None of these

21.15 What are the uses and limitations of index numbers?

21.16 On what counts are the simple aggregative index numbers inadequate?

21.17 What are index numbers? Explain briefly the various problems involved in their construction.

21.18 Discuss some of the major difficulties that you are likely to face while constructing index numbers.

21.19 "For constructing index numbers the best method on theoretical ground is not the best method from practical point of view." Discuss.

21.20 What are the advantages and limitations of a chain index number?

21.21 "The Laspeyres index has an upward bias and the Paasche index a downward bias." Explain.

21.22 Show that Fisher's 'ideal' index number satisfies the time reversal test and the factor reversal test.

21.23 Prove that Fisher's 'ideal' index lies between the Laspeyres and Paasche index numbers.

21.24 Show that Fisher's 'ideal' index is the geometric mean of the Laspeyres and Paasche index numbers.

21.25 Construct a table of the various types of index numbers, specifying in each case whether it does or does not satisfy the time reversal, factor reversal and circular tests.

21.26 What is time series deflation?

21.27 Explain the following terms:

(a) Splicing **(b)** Shifting the base

(c) Value index **(d)** Circular test

21.28 What are the three tests that are used in testing an index number? Explain them with suitable examples.

21.29 You are asked to construct an Index of Industrial Production. Explain the procedure that you would adopt in constructing such an index.

21.30 What are the uses and limitations of quantity index numbers?

21.31 Distinguish between the Bowley's formula and the Marshall-Edgeworth formula used in the construction of index numbers.

21.32 There are chances of an index being misinterpreted. What factors would you bear in mind to avoid misinterpretation of an index?

21.33 It is said that index numbers are a specialised type of averages. How far do you agree with this statement? Explain briefly time reversal and factor reversal tests.

21.34 Prove that the weighted aggregate price index with fixed weights satisfies the circular test.

21.35 What is a circular test? Do you think that Fisher's index satisfies this test? Give an example to support your answer.

21.36 Using Fisher's formula, find the price index number of the following data:

		Base-year		*Current year*	
Commodity	*Unit*	*Price (Rs)*	*Value (Rs)*	*Quantity*	*Value (Rs)*
A	kg	12.5	125	12	156
B	kg	14	112	9	135
C	Metre	11	88	9	108
D	kg	13	78	6	90

21.37 Convert the following fixed-base index numbers into chain index numbers.

Year	1995	1996	1997	1998	1999	2000
Price index	100	120	150	250	270	350

21.38 Calculate (a) a weighted average of price relatives and (b) a weighted aggregate price index, both for year 1 (based on year 1), for the following commodities:

		Price (Rs)	
Commodity	*Weight*	*Year 1*	*Year 2*
A	5	215	210
B	10	250	275
C	4	1,100	1,300
D	8	950	950
E	13	170	200

21.39 Given the following data on prices and quantities, calculate the Laspeyres, Paasche and Fisher's ideal price index numbers:

Commodities	p_0	q_0	p_1	q_1
A	2	5	3	10
B	3	10	4	12
C	5	12	6	10
D	4	8	6	15
E	6	9	5	10

21.40 Given the following data, find x_1, x_2 and x_3.

Year	*Index A*	*Link relatives*
1994	100	
1995	120	
1996	160	
1997	x_1	$I'_{96\text{-}97} = 110$
1998	x_2	$I'_{97\text{-}98} = 80$
1999	x_3	$I'_{98\text{-}99} = 140$

21.41 Calculate the chain base index number from the following data:

	Price in Rs				
Commodity	*1991*	*1992*	*1993*	*1994*	*1995*
A	2	3	4	2	7
B	3	6	9	4	3
C	4	12	20	8	16
D	5	7	18	11	22

21.42 Assuming that all the products can be assigned equal weights, calculate the chain base index numbers for the years 1996 to 2000 on the basis of the following price relatives:
(*Note*: Price relative $= (p_1/p_0) \times 100$)

Year	*A*	*B*	*C*	*D*	*E*
1996	100	100	100	100	100
1997	90	525	134	118	133
1998	81	61	60	115	125
1999	112	200	80	93	140
2000	122	66	150	86	86

21.43 An enquiry into the budgets of the middle class families in Mumbai gave the following information:

Expenses on	*Food 35 per cent*	*Rent 15 per cent*	*Clothing 20 per cent*	*Fuel 10 per cent*	*Misc. 20 per cent*
Price (1975) (Rs)	150	50	100	20	60
Price (1976) (Rs)	174	60	125	25	90

What changes in the cost of living figure of 1976 have taken place as compared to 1975?

21.44 Suppose that the average prices of four commodities (A, B, C and D) for the year 2005 were A = Rs 60, B = Rs 35.5, C = Rs 42.8 and D = Rs 87.3. Further, assume that a recent survey shows that typical monthly consumption of the commodities is A = 15, B = 18, C = 10 and D = 7. October 2007 prices of the commodities were A = Rs 82, B = Rs 41, C = 53.5 and D = Rs 92.6. What is the October 2007 price index, using the weighted aggregative method (2005 = 100)?

21.45 While calculating a certain cost of living index number, the following weights were used: Food = 25, Rent = 10, Clothing = 6, Fuel and Light = 5 and Miscellaneous = 4. Calculate the index for the period when the average percentage rise in price for the aforesaid items was 30, 42, 36, 40 and 26, respectively.

Assuming that a person was earning Rs 9800 in the base period, how much should his earning be to neutralise the price rise in the subsequent period.

21.46 The Wholesale Price Index (WPI) for selected years is given below:

Year	*WPI (2001 = 100)*
1998	96.6
1999	97.5
2000	99.3
2001	100.0
2002	103.8
2003	109.4
2004	111.6
2005	120.2
2006	123.4
2007	125.7

Calculate an index of the purchasing power of the Rupee, as measured by this index. Interpret your answer.

21.47 Answer the following short questions:

(a) In case you are called upon to compare two index numbers, is it necessary that they must have the same base period?

(b) Suppose the CPI Consumer Price Index for two different cities happens to be the same. Does this indicate that these two cities have experienced the same percentage change in prices?

(c) If prices rise (i) 30% (ii) 42%, by how much would the purchasing power of Rupee fall?

21.48 Assume that an index of industrial production is 100 in 2000, it rises 8 per cent in 2001, falls 4 per cent in 2002, rises 10 per cent in 2003, falls 2 per cent in 2004, and again rises 12 per cent in 2005, calculate the index for the six years, using 2005 as the base year.

21.49 From the chain base index numbers given below, construct fixed base index numbers (base year 2001):

Year	2001	2002	2003	2004	2005	2006	2007
Index No.	80	110	120	90	140	150	170

Convert the fixed base numbers, so derived, into chain base index numbers.

The McGraw-Hill Companies

22 CHAPTER DECISION THEORY

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the steps involved in the Decision Theory Approach
- use different criteria in decision-making under uncertainty
- understand and apply the expected value and utility as decision criteria
- understand the concept and use of Bayesian Analysis
- apply the decision tree approach in decision-making.

Chapter Prerequisites

Before starting work on this chapter, make sure you are quite conversant with the material pertaining to probability given in Chapter 9.

22.1 INTRODUCTION

In business, there arise several situations when management does not have adequate information and yet it becomes necessary to decide on a problem confronting it. In other words, decisions are made with some uncertainty. Even in other spheres too, decisions are made under uncertainty. For example, take the case of the government that is concerned with the unfavourable social environment in the country and decides major policies to bring about improvement in it in five to ten years' time. It is indeed very difficult to envisage the environmental changes in the future as the government does not have adequate information concerning the future. Despite this inherent shortcoming, the government cannot escape its responsibility towards improving the social environment.

Since decisions have to be made under uncertainty, the question is: is there any rational approach that can be followed? It is here that the role of decision theory becomes very important. In this chapter, we shall study some of the major aspects of the decision theory.

Statistical decision theory comprises a number of quantitative techniques that greatly facilitate us in analysing a decision situation. Further, they enable us to decide the best possible course from amongst the several alternatives available under given circumstances of the problem.

22.2 STEPS IN THE DECISION THEORY APPROACH

There are four steps involved in the decision theory approach:

1. The first step is to list all the possible outcomes that must be considered in the decision. These are called *states of nature* or *events*, for the decision problems. Thus, the decision-maker has no control over these outcomes or events. This apart, there is uncertainty as to which outcome or state of nature will actually occur.
2. The second step involves the listing of all the courses of action that are available to the decision maker. He has control over these in the sense that he can choose any one of them.
3. The third step is that the decision-maker has to construct a payoff table that deals with the monetary gain or loss from each possible combination of decision alternative and state of nature. Sometimes, the benefit may be in terms of the reduced cost or reduced time or any other measure of benefit but these too must be in quantitative terms.
4. Finally, the decision-maker has to choose one of the alternatives on the basis of some criterion. This is normally based on the information given in step (3). Sometimes it becomes necessary to collect additional information without which an appropriate decision cannot be made.

22.3 TYPES OF ENVIRONMENTS

There are three types of environments in which decisions are made:

Under Conditions of Certainty This environment has only one state of nature; thus, there is complete certainty about the future. This environment is usually associated with routine decisions involving minor issues. However, even in routine matters, it is usually impossible to have complete certainty about the future.

Under Conditions of Uncertainty In this environment, more than one state of nature exists and the decision-maker does not have any knowledge about the various states of nature. He is even unable to assign probabilities to the states of nature because of inadequate information with him. As the probabilities of various states of nature are unknown, the actual decisions are based on specific criteria. Several principles on the basis of which decisions can be made in such situations will be discussed below.

Under Conditions of Risk Here, more than one state of nature exists. However, as the decision-maker has some information with him, he is able to assign probabilities to each of the states of nature. As the decision maker can assign probabilities to different outcomes, this situation is better than the preceding one so far as decision making is concerned.

22.4 DECISION-MAKING UNDER CERTAINTY

It will be obvious from the above discussion that it is easy to analyse the situation and reach a good decision under conditions of complete certainty. As certainty involves only one state of nature, the decision-maker has just to choose the best payoff from amongst all the alternatives available. Let us take an example.

Example 22.1

Table 22.1 Payoff Table for XYZ Company's Expansion Decision (Payoff Expressed in Profits Earned Over Next Five Years)

		State of Nature (Demand) (Rs Lakh)			
		High	*Moderate*	*Low*	*Failure*
Options available to decision-maker	*Expand*	400	500	– 400	– 350
	Build	700	400	– 600	– 850
	Subcontract	200	100	– 50	– 80

Which of the three options the XYZ company should choose?

Solution Since our assumption here is that the management of the company XYZ is not subject to uncertainty, decision-making is not difficult here. One has to go in for the highest payoff and the corresponding action. For example, if management knows that the demand for its products is going to be moderate, then the best course of action for it is to choose 'expand' as it yields the highest payoff of Rs 500 lakh. Likewise, if management knows that the demand is going to be low, then its choice should be in favour of 'subcontract' even though it generates a loss of Rs 50 lakh. This is because it is still the best alternative given the state of nature. When the management thinks that the demand for its product is high, then it should choose 'build' as it gives the maximum payoff of Rs 700 lakh. In business, however, one comes across not one state of nature but two or more states of nature where decision becomes more difficult. As the decision-making under certainty does not pose any major problem, it is not so important for us.

22.5 DECISION-MAKING UNDER UNCERTAINTY

There are four criteria for decision-making under uncertainty. These are:

1. The *maximax* or *minimin* criterion
2. The *maximin* or *minimax* criterion
3. The *minimax* regret criterion
4. The criterion of *realism*

We shall discuss these criteria briefly. Before we do so, let us set up a payoff table with hypothetical figures:

Table 22.2 Payoff Table

	Rs Lakh			
Demand Supply	*High Demand*	*Moderate Demand*	*Low Demand*	*Failure*
Action A_1 Expand	6,500	3,000	–2,000	–5,000
Action A_2 Build	9,000	5,000	–3,000	–10,000
Action A_3 Subcontract	3,000	1,500	–1,000	–2,000

Maximax Criterion The maximax criterion for decision-making under uncertainty is an optimistic criterion. It provides an optimum solution. According to this criterion, the decision-maker has to select that course of action which maximises his maximum payoff. Table 22.2 gives the payoff in different states of nature, that is, high demand, moderate demand, low demand and failure. Likewise, three alternative actions, viz. expand, build and subcontract are shown. The figures underlined in the table show the maximum payoff possible for each of the three decision alternatives. It will be seen that the second decision alternative (A_2)—'build'—provides the maximum payoff. As such, this will be chosen.

Maximin Criterion Unlike the maximax criterion, the maximin criterion for decision-making under uncertainty is a pessimistic criterion. Here the decision-maker tries to maximise his minimum possible payoffs. First, he identifies the minimum payoff that he can get for each decision alternative. He then selects the alternative within this group, which results in the maximum payoff. Table 22.2 is reproduced as follows:

Table 22.3 Payoff Table

	Rs Lakh			
Demand Supply	*High Demand*	*Moderate Demand*	*Low Demand*	*Failure*
Action A_1 Expand	6,500	3,000	–2,000	–5,000
Action A_2 Build	9,000	5,000	–3,000	–10,000
Action A_3 Subcontract	3,000	1,500	–1,000	–2,000

It may be noted that the minimum payoff in each of the three decision alternatives is in the last column of the table. All these figures have been underlined to give them a separate identification. On the basis of the maximin criterion, the maximum payoff is in A_3, that is, subcontract where the payoff amounts to Rs –2,000 lakh over the next five years.

The Minimax Regret Criterion On the basis of this criterion, the decision-maker selects a decision alternative that offers the least amongst the maximums of regrets in each.

In order to elaborate this criterion, let us assume that the decision-maker had earlier chosen A_3—subcontracting—thinking that the demand for his product would be low. But later on it turned out that the demand was high. The profit he would make from subcontracting in case of high demand would be Rs 3,000 lakh. Had the decision-maker known that there would be high demand for his product, he would have chosen the second decision alternative of building a new plant and that would have given him a profit of Rs 9,000 lakh. The difference between Rs 9,000 lakh (which is the optimum profit) and Rs 3,000 lakh (the payoff which actually realised) is Rs 6,000 lakh. This amount of Rs 6,000 lakh is known as the *regret* resulting from a decision.

It may be noted that if this criterion is to be used by the decision-maker, he should first calculate the regret associated with all the twelve combinations of decision alternatives and states of nature.

Table 22.4 shows the regret in each case. The regret values are obtained by subtracting every entry in the original payoff Table 22.2 from the largest entry in its column. Now, the decision-maker has to indicate the maximum regret for each decision alternative. These regrets have been underlined in the table. He has now to choose the minimum of these regret values. In this particular case, the minimum regret value happens to be Rs 3,000 lakh and it is associated with the decision alternative A_1, that is, 'expand'.

Table 22.4 Regrets for each Combination of Decision Alternative and State of Nature

	Rs Lakh			
Demand Supply	*High Demand*	*Moderate Demand*	*Low Demand*	*Failure*
Action A_1 Expand	2,500	2,000	1,000	3,000
Action A_2 Build	0	0	2,000	8,000
Action A_3 Subcontract	6,000	3,500	0	0

The Criterion of Realism or Hurwitz Criterion This criterion for decision-making under uncertainty is in between the two criteria—maximax and maximin. It may be recalled that the maximax criterion is an optimistic criterion while the maximin is a pessimistic criterion. Leonid Hurwitz evolved a formula for combining optimistic and pessimistic criteria. In order to use this criterion, the decision maker has to specify a coefficient or an index of optimism, which is denoted by the notation α (the Greek letter alpha), which can take any value between 0 and 1. Before applying this criterion, the decision-maker has to determine both the maximum and the minimum payoff for each decision alternative. Table 22.5 shows these payoffs, which are reproduced from Table 22.2 (maximum payoffs and Table 22.3 minimum payoffs).

Table 22.5 Maximum and Minimum Payoffs for Each Decision Alternative

	Rs Lakh		
Decision Alternative	*Maximum Payoff*	*Minimum Payoff*	*Measure of Realism*
A_1 Expand	6,500	–5,000	4,200
A_2 Build	9,000	–10,000	5,200
A_3 Subcontract	3,000	–2,000	2,000

It is now necessary to assign a certain value to α. Let us assume that the decision-maker is very confident that there would be quite a high demand for his product and, therefore, assigns 0.8 value to α. We have to multiply each value given in the first column of Table 22.5 by 0.8. Similarly, each value given in column 2 of Table 22.5 is to be multiplied by $1 - \alpha$, that is, 0.2. The resultant two figures are to be added to obtain the measure of realism. The final choice of the decision maker will obviously be in favour of the highest number amongst the three values. These calculations are given below:

Expand: $(6{,}500 \times 0.8) + (-5{,}000 \times 0.2) = \text{Rs } 4{,}200 \text{ lakh}$
Build: $(9{,}000 \times 0.8) + (-10{,}000 \times 0.2) = \text{Rs } 5{,}200 \text{ lakh}$
Subcontract: $(3{,}000 \times 0.8) + (-2{,}000 \times 0.2) = \text{Rs } 2{,}000 \text{ lakh}$

These three figures have now been entered in the last column of Table 22.5. Since decision alternative (A_2) 'build' gives the highest measure of realism, the decision-maker will choose it. The point to note is that under this criterion the decision-maker has been able to use his own assessment of the market by choosing $\alpha = 0.8$.

Jacob Bernoulli Method or the Principle of Insufficient Reason

We have seen in the criterion of realism that the decision maker uses his own assessment of a given problem and then on that basis he decides a certain value of α to ascertain the best possible outcome.

But at times it may be difficult for him to show any preference to one decision alternative over the other. In such cases, the Jacob Bernoulli method or the principle of insufficient reason becomes more relevant. In the absence of any information to the contrary, the decision maker assigns equal probabilities to the different states of nature.

Taking our previous example (Table 22.5), we can assign 0.5 probability to each maximum payoff and minimum payoff for each decision alternative. Based on this criterion, the calculations will be as follows:

Expand: $(6{,}500 \times 0.5) + (-5{,}000 \times 0.5) = \text{Rs } 750 \text{ lakh}$
Build: $(9{,}000 \times 0.5) + (-10{,}000 \times 0.5) = \text{Rs } -500 \text{ lakh}$
Subcontract: $(3{,}000 \times 0.5) + (-2{,}000 \times 0.5) = \text{Rs } 500 \text{ lakh}$

It will be seen that among the three decision alternatives the maximum payoff is for expansion: Rs 750 lakh. As such, the decision-maker should choose to expand his production instead of building a new plant or subcontracting. It may also be noted that assignment of equal probabilities to maximum and minimum payoffs has changed the decision from 'build' to 'expand', that is, from building a new plant to expanding the existing one.

22.6 DECISIONS UNDER RISK

The decision in situations wherein the decision-maker considers several possible outcomes and assigns probabilities to their occurrence are called decision under risk. The decision-maker is in a position to assign probability of occurrence of each outcome.

Expected Monetary Value (EMV) or Expected Payoff (EP) This criterion requires the decision-maker to calculate the expected monetary value for each alternative decision. This is done by **(i)** multiplying each decision alternative by the probability assigned to the state of nature that can occur and **(ii)** aggregating the values thus arrived at. This is illustrated by the following example.

Example 22.2 ABC Company is a leading firm of dry cleaners with branches in Delhi and the large towns in Western UP. The company has been thinking of setting up a new shop in Agra, which already has a large number of dry cleaning shops, one of which is a branch of a national company at par with the ABC. ABC is concerned about the reaction of customers and potential customers to the opening of its branch in Agra, and is of the opinion that one of the three possibilities can take place: (i) increased market share, (ii) no change in the present operations and (iii) a reduced market share. The company assigns a value and probability to each of these outcomes as shown in Table 22.6.

Table 22.6 Value and Probability to Three States of Nature

States of Nature	*Value in Rs Lakh*	*Probability*
Increased market share	3	0.2
No change	1	0.5
Reduced market share	–2	0.3

On the basis of the information given in Table 22.6, determine the expected monetary value of the decision to open the branch.

Solution The expected monetary value can be calculated as follows:

$$\begin{aligned}\text{EMV (Expected Monetary Value)} &= (\text{Rs } 3{,}00{,}000 \times 0.2) + (\text{Rs } 1{,}00{,}000 \times 0.5) + (\text{Rs } -2{,}00{,}000 \times 0.3)\\ &= \text{Rs } 60{,}000 + \text{Rs } 50{,}000 - \text{Rs } 60{,}000\\ &= \text{Rs } 50{,}000\end{aligned}$$

Thus, the expected monetary value of the decision to open the branch is Rs 50,000. It may be noted that a little change in these probabilities would change the EMV considerably. A more pessimistic view of the venture might reverse a profitable venture into one resulting in loss. In other words, much depends on the probabilities that are assigned for decision-making. Another point to note is that the decision depends entirely on the EMV. Sometimes, one may be more interested in the utility as a yardstick rather than monetary consideration. This aspect we will see later in this chapter.

Expected Opportunity Loss (EOL) In the preceding section, we discussed concept of regret or opportunity loss associated with each combination of decision alternative and the state of nature.

An alternative approach to maximise EMV is to minimise expected opportunity loss (EOL). This concept represents the amount of profit that the decision-maker has lost as he has not taken the most profitable course of action. We can calculate EOL if we know the conditional opportunity loss (COL), which is zero for the optimal act. For other states of nature, COL is the difference between the payoff of the optimal act and the payoff of each combination of decision alternatives and the state of nature. This will always be positive. When we replace payoffs by the corresponding opportunity losses, then the table thus obtained is known as the *Loss Table* instead of the *Payoff Table*. An example will make the concept clear.

Example 22.3 A group of volunteers of a service organisation raise money each year by selling gift articles outside a stadium after a football match between teams A and B. They can buy any of the three different articles from a dealer. Their sales are mostly dependent on the team that wins the match. A conditional payoff table is as under:

Payoff Table

	Type of gift articles		
	I	*II*	*III*
Team A wins	Rs 1,500	Rs 1,000	Rs 700
Team B wins	Rs 700	Rs 500	Rs 900

(i) Construct the opportunity loss table.

(ii) Which type of gift article should the volunteers buy if the probability of team A's wining is 0.7?

Solution In order to construct an opportunity loss table, we have to take the difference between the highest payoff and each state of nature. Since in the category 'team A wins', the highest payoff of Rs 1,500, from which we have to take the differences. Likewise, in case team B wins, the highest payoff of Rs 900 is for gift article type III. Hence, the opportunity loss is to be calculated from this figure. These calculations are given in Table 22.7.

Table 22.7 Opportunity Loss Table

	Type of Gift Articles		
	I	*II*	*III*
Team A wins	Rs 1,500 – Rs 1,500 = 0	Rs 1,500 – Rs 1,000 = 500	Rs 1,500 – Rs 700 = 800
Team B wins	Rs 900 – Rs 700 = 200	Rs 900 – Rs 500 = 400	Rs 900 – Rs 900 = 0

Since the probability of team A winning is 0.7, the probability of team B wining is 1 – 0.7= 0.3. Taking these probabilities, the expected opportunity loss from buying and selling different gift articles would be as shown below:

Table 22.8 Calculation of Expected Opportunity Loss

Type of Gift Article	*Expected Opportunity Loss*
I	$(0.7 \times 0) + (0.3 \times 200)$ = Rs 60
II	$(0.7 \times 500) + (0.3 \times 400)$ = Rs 470
III	$(0.7 \times 800) + (0.3 \times 0)$ = Rs 560

As the optimal strategy is the one which minimises the expected loss, it becomes evident from the above calculations that the expected opportunity loss of Rs 60 is the lowest for type I gift article. As such, the volunteers should purchase this gift article in preference to the other two types.

Let us take another example for calculating expected opportunity loss.

Example 22.4 Calculate expected opportunity loss from the following payoff table:

Payoff Table

Event \ *Action*	A_1 (Rs)	A_2 (Rs)	A_3 (Rs)	*Event probabilities*
E_1	50	–10	–80	0.2
E_2	400	500	600	0.5
E_3	600	900	800	0.3

Solution We have to follow the same procedure as was used in the earlier example, to calculate the opportunity loss. The calculations are shown in the table given below:

Table 22.9 Opportunity Loss Table

	A_1 (Rs)	A_2 (Rs)	A_3 (Rs)
E_1	50 – 50 = 0	50 – (–10) = 60	50 – (–80) = 130
E_2	600 – 400 = 200	600 – 500 = 100	600 – 600 = 0
E_3	900 – 600 = 300	900 – 900 = 0	900 – 800 = 100

As the probabilities of three events are different, we have to work out the values of the preceding table along with the respective probabilities. These calculations are shown in Table 22.10.

Table 22.10 Expected Opportunity Loss

Events	*Probabilities*	A_1 *(Rs)*	A_2 *(Rs)*	A_3 *(Rs)*
E_1	0.2	0	$0.2 \times 60 = 12$	$0.2 \times 130 = 26$
E_2	0.5	$0.5 \times 200 = 100$	$0.5 \times 100 = 50$	0
E_3	0.3	$0.3 \times 300 = 90$	0	$0.3 \times 100 = 30$
	Total	**190**	**62**	**56**

The minimum expected loss of Rs 56 is in A_3. As such, we should prefer action A_3.

Expected Value of Perfect Information (EVPI) Another criterion commonly used in decision-making is the expected value of perfect information (EVPI). This is the difference between the expected profit with perfect information (EPPI) and the expected profit (EP) of the optimal decision without perfect information. It may be noted that EVPI is the maximum amount that the decision-maker can spend to obtain additional information relating to the states of nature. Another point to note is that EVPI is always equal to EOL of selecting the optimum action under uncertainty.

Symbolically,

$$EVPI = EOL$$
$$EVPI = EPPI - EP$$
$$EPPI = EP + EOL$$

It is necessary to know EVPI in order to decide whether additional information should be collected or not. Since the collection of any additional information will add to the existing cost, unless EVPI is higher than the cost of additional information, there is no point in obtaining it. In contrast, when EVPI is higher than the cost of collecting additional information, then it is advisable to collect it.

Let us take an example to illustrate how EP, EPPI and EVPI can be calculated.

Example 22.5 The following table shows three states of nature (events) and three actions. The amount that a person will gain in each combination of state of nature and action is shown in the table. Probabilities for three states of nature are also given.

Action / *State of Nature (Event)*	A_1 *(Rs)*	A_2 *(Rs)*	A_3 *(Rs)*
S_1	25,000	–7,000	–18,000
S_2	30,000	50,000	–8,000
S_3	40,000	25,000	60,000
State of nature	**S_1**	**S_2**	**S_3**
Probability	0.3	0.5	0.2

(i) Find out values of EP, EPPI and EVPI.

(ii) A research firm has agreed to conduct a survey to provide the management with additional information regarding the states of nature. It will charge Rs 12,000 for undertaking this survey. Do you think that the survey should be conducted?

Solution

(i) We have to first compute EP. However, this is possible only when we know the expected payoffs for each action under uncertainty. These calculations are given below:

$$EP(A_1) = Rs\ (25{,}000 \times 0.3) + (30{,}000 \times 0.5) + (40{,}000 \times 0.2)$$
$$= Rs\ 7{,}500 + 15{,}000 + 8{,}000$$
$$= Rs\ 30{,}500$$
$$EP(A_2) = Rs\ (-7{,}000 \times 0.3) + (50{,}000 \times 0.5) + (25{,}000 \times 0.2)$$
$$= Rs\ -2{,}100 + 25{,}000 + 5{,}000$$
$$= Rs\ 27{,}900$$
$$EP(A_3) = Rs\ (-18{,}000 \times 0.3) + (-8{,}000 \times 0.5) + (60{,}000 \times 0.2)$$
$$= Rs\ -5{,}400 - 4{,}000 + 12{,}000$$
$$= Rs\ 2{,}600$$

From these calculations, it is clear that the highest payoff is in the combination of Event 1 Action 1. Hence, one has to choose this payoff (EP), which is under uncertainty.

Having computed EP, we have to find out the highest payoff for each action under certainty. This is done under the assumption that perfect predictor is available. When the state of nature S_1 is known to us, then we will take action A_1 because it gives the maximum payoff. When we know that S_2 prevails, then we will take action A_2, which gives the highest payoff. Finally, when S_3 prevails then A_3 will be preferred. On this basis, we can now calculate EPPI as follows:

$$EPPI = Rs\ (25{,}000 \times 0.3) + (50{,}000 \times 0.5) + (60{,}000 \times 0.2)$$
$$= Rs\ 7{,}500 + 25{,}000 + 12{,}000$$
$$= Rs\ 44{,}500$$

The expected value of perfect information (EVPI) = EPPI – $EP(A_1)$

$$= Rs\ 44{,}500 - 30{,}500 = Rs\ 14{,}000$$

(ii) Since the survey will cost Rs 12,000, which is less than the expected value of perfect information (EVPI), it is advisable to conduct the survey.

Let us take another example.

Example 22.6 The following payoff table gives the most accurate estimates of payoffs for three alternative courses of action and four possible states of nature under conditions of uncertainty.

A Payoff Table

(Rs lakh)

Courses of Action	*States of Nature*			
	S_1	S_2	S_3	S_4
A_1	125	100	70	20
A_2	70	90	100	50
A_3	150	120	60	–10

The probabilities are—S_1 : 0.4, S_2 : 0.25, S_3 : 0.2 and S_4 : 0.15. Two questions arise:

1. What course of action would be recommended in this situation?
2. What is expected value of perfect information?

Solution On the basis of the above information, expected payoff of each course of action can be computed as follows:

$S(A_1) = (125 \times 0.4) + (100 \times 0.25) + (70 \times 0.2) + (20 \times 0.15)$
$= 50 + 25 + 14 + 3 = \text{Rs } 92 \text{ lakh}$
$S(A_2) = (70 \times 0.4) + (90 \times 0.25) + (100 \times 0.2) + (50 \times 0.15)$
$= 28 + 22.5 + 20 + 7.5 = \text{Rs } 78 \text{ lakh}$
$S(A_3) = (150 \times 0.4) + (120 \times 0.25) + (60 \times 0.2) + (-10 \times 0.15)$
$= 60 + 30 + 12 - 1.5 = \text{Rs } 100.5 \text{ lakh}$

Since the expected payoff (EP) of A_3 is the highest, A_3 should be recommended.

In order to find the expected value of perfect information (EVPI), we assume that if we have the perfect information then we would select the best action in each case. For example, if we come to know that S_1 will prevail, then our choice will be A_3 as it gives the highest value of Rs 150 lakh. In this way, our preference would be in favour of that action that gives the highest value in other states of nature. Now, we can obtain the expected payoff with perfect information (EPPI).

$EPPI = (150 \times 0.4) + (120 \times 0.25) + (100 \times 0.2) + (50 \times 0.15)$
$= \text{Rs } 60 + 30 + 20 + 7.5 = \text{Rs } 117.5 \text{ lakh}$

In order to calculate the expected value of the perfect information, the value of optimal EP which is $S(A_3)$ is to be deducted from EPPI.

$EVPI = EPPI - EP$
$= \text{Rs } 117.5 \text{ lakh} - \text{Rs } 100.5 \text{ lakh}$
$= \text{Rs } 17 \text{ lakh}$

22.7 UTILITY AS A DECISION CRITERION

So far our discussion in this chapter was confined to expected value criterion for decision-making. However, there are certain situations where the expected value criterion may not be appropriate. Suppose that a choice is given between the two alternatives, A_1 and A_2. In A_1 a certain gift of Rs 20,000 is given while in A_2, a gift of Rs 50,000 is given if a coin, when tossed, shows the head up and nothing if it shows up the tail. When we apply the expected value criterion, then we find the calculations as follows:

$A_1 = \text{Certainty of Rs } 20{,}000$
$A_2 = (\text{Rs } 50{,}000 \times 0.5) + (\text{Re } 0 \times 0.5) = \text{Rs } 25{,}000$

Although the expected value criterion shows that A_2 is preferable as it yields a higher figure, but as it involves some risk, a person may prefer A_1. This example makes it clear that expected value criterion may not be used always while making a decision. An alternative criterion, which may be applicable in such a case as given above, should be more valid. Such an alternative has been suggested by Von Neumann and Morganstern. They hold that decisions are made so as to maximise expected utility rather than expected monetary value. In the above case where a choice is to be made between A_1 and A_2, according to them, the decision maker derives greater utility from A_1 as compared to A_2 and hence a decision to choose the former. If the expected utility for A_1 and A_2 is the same, then we can say that the decision-maker is indifferent in respect of the two options. We may say that utility is the pleasure or displeasure that one would get from certain outcomes.

In order to use the utility criterion for decision-making, it is necessary for us to first establish the relationship between money and utility. In other words, we have to find out the amount of utility that

can be derived from a given amount of money. Once the monetary values have been transformed in utility terms, an appropriate decision can be made.

Different Utilities*

It may be noted that there would be different utilities for different persons for the same situation. This is because utility is the combined product of one's psychological make-up, one's expectations about the future and the particular decision or act which is being evaluated. Some persons, for example, would prefer investments where high risk is involved. In such investments, they either gain a huge profit or sustain a heavy loss. Such investors are generally with considerable net worth and are prepared to bear a heavy loss. In contrast, there are investors who are very cautious in their investments. They would generally be persons with moderate net worth and would prefer investments where they have minimum risk of losing money. They would obviously prefer those investments where they expect that the outcome would be positive.

Figure 22.1 shows three different utility curves depicting three types of investors, viz. risk-seekers, risk-averse and risk-neutral.

It will be seen from Fig. 22.1 that investor A is cautious and conservative investor. His utility increases only very slightly after the zero-profit point, while a move to the left of the zero-profit point decreases his utility rapidly. In contrast, investor B has a different utility curve. Here, a move to the right of the zero-profit point increases the utility much more than the loss of the same amount decreases it. This shows that B is not averse to high-risk investment. He thinks that a huge profit would be quite rewarding while a setback in terms of loss would not make things much worse than what they are at present. In between these two investors is C who would not sustain a heavy loss or gain a huge profit. In fact, his utility curve is linear unlike the other two investors. This also indicates that he can use the expected value instead of utility as the criterion for decision-making.

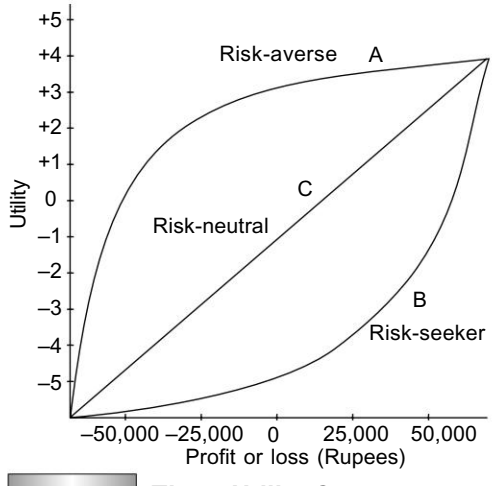


Fig. 22.1 Three Utility Curves

It is in the interest of both A and B to use utility as the criterion for decision-making. However, their approaches will be different. While A will demand a high expected value for the outcome but B may act when the expected value is negative.

Example 22.7 A business firm has two options before it:

(a) To enter into a contract with another company to supply raw material to the company. This would give a certain profit of Rs 50,000.

(b) Instead of supplying raw material, it may itself bring out a new product. The likely profit or loss along with the expected probabilities are given below. In addition, the utility values associated with different profit/loss levels are given below:

* Based on Levin, Richard I. and Charles A. Kirpatrick: *Quantitative Approaches to Management,* Tokyo, McGraw-Hill Kogakusha, Ltd., 1978, pp. 143–146.

Profit/Loss (Rs)	–50,000	0	50,000	1,00,000	2,00,000
Probability	0.1	0.1	0.4	0.3	0.1
Utility	–0.50	0	0.40	0.50	1.10

You are asked to determine which option the manager would prefer in case he wanted to maximise **(a)** the EMV, and **(b)** the expected utility.

Solution In order to ascertain the EMV and the expected utility, we set up the following table:

Table 22.11 Calculation of EMV and EU

Profit/Loss (Rs) (1)	*Probability (2)*	*Utility (3)*	*Expected Profit (1) × (2) (4)*	*Expected Utility (2) ×(3) (5)*
–50,000	0.1	–0.50	–5,000	–0.05
0	0.1	0	0	0.00
50,000	0.4	0.40	20,000	0.16
1,00,000	0.3	0.50	30,000	0.15
2,00,000	0.1	1.10	20,000	0.11
		Total	**65,000**	**0.37**

(i) The company has a certain profit of Rs 50,000 and expected profit of Rs 65,000 as given in Table 22.11. Now, the manager has to decide between the two. On the basis of these calculations, the manager should go in for the second option and undertake the manufacture of the new product.

(ii) The expected utility, EU, for option (b) is 0.37 as shown in Table 22.11 while the utility associated with the profit of Rs 50,000 in respect of alternative (a) is 0.40. Since the latter figure is higher, the manager should choose alternative (a). In other words, he should enter into a contract with another company for the supply of raw material.

This table shows how the decision based on EMV for the manufacture of a new product gets reversed on the basis of EU criterion.

22.8 DECISION TREES

The discussion so far was confined to single stage problems wherein the decision-maker was required to select the best course of action on the basis of information available at a point of time. However, there are problems with multiple stages wherein a sequence of decisions is involved. Each decision leads to a chance event which, in turn, influences the next decision. In such cases, a new approach known as the *decision tree* is used. It may be recalled that in Chapter 9 it was mentioned that in understanding probability problems, tree diagrams can be helpful. However, our focus here is to illustrate how the decision-tree approach can be helpful in decision making too.

A decision tree is a graphical device depicting the sequences of action-event combinations. All possible sequences of action-event combinations are shown in a systematic manner in a decision tree.

Figure 22.2 shows a simple decision tree.

This figure shows that a choice is to be made between two alternatives, viz. whether or not to open a new branch of a particular dry cleaning company. The square node shows a decision point and a circular node shows states of nature. Branches from this represent each possible outcome over which

the decision-maker has no control. As can be seen from Fig. 22.2, three branches are emanating from the circular node, which represent the three possible outcomes or states of nature that can result. These are: increased market share, no change in market share, and reduced market share.

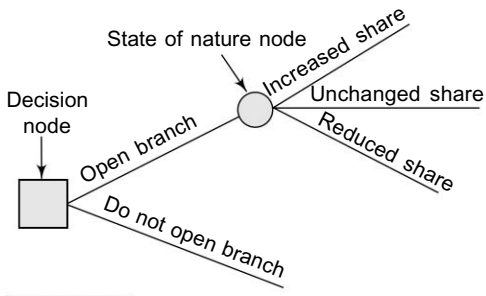


Fig. 22.2 An Example of Decision Tree

Method for Constructing a Decision Tree

Having given a broad idea of the decision tree, we should now look into the method of constructing such a tree. The following five steps are involved in constructing a decision tree:

1. Identify all the possible courses of action.
2. List the possible results—'states of nature' of each course of action specified in (1) above.
3. Calculate the payoff of each possible combination of courses of action and results. The payoff is normally in monetary terms.
4. Assign probabilities to the different possible results for each given course of action. The probability indicates the likelihood of occurrence of a particular result or event.
5. Finally, select the course of action that gives the maximum payoff.

Let us take an example to show how these steps can be used to construct a decision tree.

Example 22.8 A marketing manager has to decide between advertising his product on a national level and not advertising it. If he advertises the product and it is successful, his company will gain Rs 4 lakh, but if he advertises and the product fails, the company will lose Rs 4 lakh. No loss or gain is attached to his not taking action. He thinks that there is 0.6 probability that the advertising campaign will be successful.

(a) Construct a decision tree to help analyse this problem.

(b) What action the marketing manager should take?

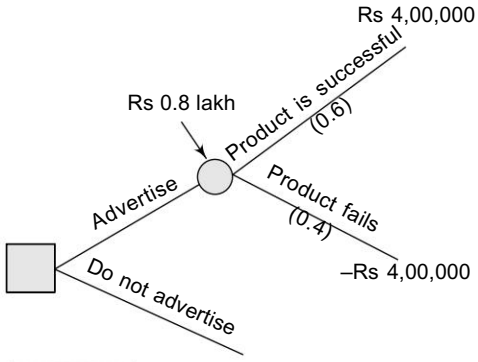


Fig. 22.3 A Decision Tree for Marketing Problem

Solution

$$\begin{aligned} \text{EMV (UC)} &= \text{Rs } (4 \text{ lakh} \times 0.6) + \text{Rs } (-4 \text{ lakh} \times 0.4) \\ &= \text{Rs } 2.4 \text{ lakh} - \text{Rs } 1.6 \text{ lakh} \\ &= \text{Rs } 0.8 \text{ lakh} \end{aligned}$$

This shows that the expected monetary value under uncertainty is Rs 0.8 lakh as also shown in Figure 22.3. As such, the marketing manager should advertise the product.

Let us take another example.

Example 22.9 A few friends who have just obtained their postgraduate degree in medicine are considering to start their private clinic. If the medical demand is high indicating a favourable market for the proposed clinic, the doctors could realise a net profit of Rs 8,00,000. If the market is not favourable, they could lose Rs 2,00,000—preliminary expenses needed to set up the clinic. Of course, the doctors do not have to proceed at all, in which case they have not to incur any expenditure. In the absence of any market data, the doctors think that there is a 50-50 chance that their clinic will be successful.

On the basis of the above information, construct a decision tree to facilitate the analysis of this problem. What should the doctors do?

Solution The fresh medical graduates should set up their clinic. The calculations are shown below.

$$\begin{aligned}\text{EMV (UC)} &= \text{Rs } (8,00,000 \times 0.5) + \\ &\qquad \text{Rs } (-2,00,000 \times 0.5) \\ &= \text{Rs } 4,00,000 - \text{Rs}1,00,000 \\ &= \text{Rs } 3,00,000\end{aligned}$$

Figure 22.4 is the tree diagram pertaining to this problem. It may be mentioned here that this tree diagram can be further expanded if we introduce the extent of favourableness of the clinic. Thus, we can further branch off from favourable clinic to **(i)** highly favourable, **(ii)** moderately favourable, and **(iii)** marginally favourable and assign probabilities and corresponding values to these three likely outcomes.

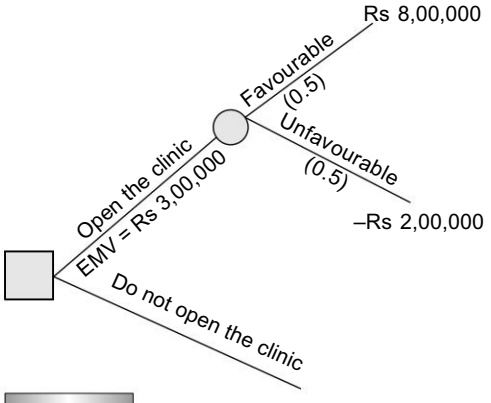


Fig. 22.4 A Decision Tree Regarding Setting up of a Clinic

Advantages of the Decision Tree Analysis As there are a number of advantages of the decision tree approach, it is now being increasingly used by management in decision-making. The main advantages are:

1. The decision tree approach structures the decision process and thus helps one in making a decision in a systematic manner.
2. The approach necessitates the decision-maker to consider all possible outcomes regardless of whether they are desirable or not.
3. The decision tree is helpful in communicating the decision-making process to others in a very succinct manner, clearly indicating the assumptions used.
4. While considering all the alternatives, attention can be focused on each individual financial figure, probability, as also the underlying assumption, one at a time.
5. The decision tree can be used with a computer, which means that different sets of assumptions can be used to ascertain their influence on the final outcome.

Limitations of the Decision Tree Analysis The decision tree analysis is not free from limitations. Let us see what they are.

1. Decision trees need time and money to complete. As such, they are unsuitable for minor decisions where their cost may exceed the benefit to be derived from them.
2. As the information is presented in a quantitative form, there is a risk that it may be taken as exact. It is necessary to ensure that the information used in the decision tree is reliable.
3. The information required for this approach may not be available because of particular decision was not taken before and hence there is no evidence on which the probability can be assumed.
4. Non-quantitative factors such as people's attitudes, government policy, and so on, may be more important but these do not enter into a decision tree.

22.9 BAYESIAN ANALYSIS

Sometimes managers find that prior probabilities regarding certain states of nature are no longer applicable. For example, a firm manufacturing readymade garments finds that certain garments are not selling in the market on account of their colour. It, therefore, has to use a different colour combination. In the meanwhile, it has to revise its prior probabilities. It may be noted that prior probabilities are changed after getting some additional information. It is here that the Bayesian analysis is used. Section 9.7 of Chapter 9 deals with the Bayesian analysis involving the revision of probabilities. It gives two examples—one with two elementary events and the second with three elemantary events. Reference may be made to that section as well. Here we discuss in greater detail how Bayes' theorem is used in handling additional information, which is more helpful in decision-making. It also enables us to determine the value of additional information and to decide whether it is advisable to get it.

There can be three types of analysis while using the Bayesian approach. These are: *prior* analysis, *posterior analysis* and *pre-posterior analysis*.

Prior Analysis

While deciding which course of action should be chosen, the decision-maker uses prior probabilities only. These probabilities are prior to the receipt of any new information.

Posterior Analysis

This involves the use of posterior or revised probabilities while deciding on the course of action. Prior probabilities are revised by the decision-maker on receiving new information on the states of nature.

Pre-posterior Analysis

This analysis deals with the strategic question of whether new information should be obtained and, if so, how much before making a final or terminal decision.

If the decision-maker is willing to make certain probability assessments, pre-posterior analysis will enable him to ascertain the value of alternative research studies prior to undertaking the research.This value is known as the expected monetary value of sample information EVSI. From this if cost of information (CI) is subtracted, the expected monetary gain of sample information EGSI can be obtained.

This can be shown as

$$EGSI = EVSI - CI$$

Steps in Pre-posterior Analysis The following steps are involved in a pre-posterior analysis:

1. Identify the possible research outcomes and calculate their unconditional or marginal probabilities.
2. Assume that each research outcome, in turn, has been obtained. Now, for each research outcome: **(i)** calculate posterior probabilities, **(ii)** calculate the expected monetary value of each course of action under consideration, **(iii)** select that course of action which yields the highest expected monetary value, and **(iv)** multiply the highest expected monetary value by the marginal probability of the research outcome.
3. Add the products of step 2(iv) to obtian the expected monetary value of the strategy that includes commissioning of research before taking the final decision.

4. Calculate the expected monetary value of sample information (EVSI).
5. Calculate the expected monetary gain of sample information (EGSI).
6. Decide in favour of that strategy which yields the highest expected gain from sample information (EGSI) provided there is at least one strategy that gives a positive EGSI. In case there is no strategy with a positive EGSI, decide in favour of the strategy that gives the highest EMV.

An example will make these steps clear. It covers all the three analyses—prior, posterior and pre-posterior.

Example 22.10 Suppose a marketing manager of a soft drink manufacturing company is seriously considering whether to undertake a special promotion. The two options before him are **(i)** run a special promotion, and **(ii)** do not run a special promotion. The following table gives the probabilities assigned by the marketing manager to the three possible outcomes, viz. very favourable consumers' reaction, favourable consumers' reaction and unfavourable consumers' reaction. These probabilities are known as prior probabilities.

Table 22.12 An Example of Prior Analysis

Possible consumer reaction	*Alternative courses of action*		*Probabilities of consumer reactions*
	A_1 *(Rs)*	A_2 *(Rs)*	
Very favourable	1,00,00,000	0	0.7
Favourable	10,00,000	0	0.1
Unfavourable	–50,00,000	0	0.2

Prior Analysis On the basis of this information, prior analysis will give the expected monetary value. This will be

$$\begin{aligned}
\text{EMV}(A_1) &= \text{Rs } 1,00,00,000 \times 0.7) + (\text{Rs } 10,00,000 \times 0.1) + (-\text{Rs } 50,00,000 \times 0.2) \\
&= \text{Rs } 70,00,000 + \text{Rs } 1,00,000 - \text{Rs } 10,00,000 \\
&= \text{Rs } 61,00,000 \\
\text{EMV}(A_2) &= \text{Rs } 0 \\
\text{EMV } (C) &= \text{Rs } 1,00,00,000 \times 0.7) + (\text{Rs } 10,00,000 \times 0.1) \\
&= \text{Rs } 71,00,000 \\
\text{EMV (UC)} &= \text{Rs } 61,00,000 \text{ same as EMV}(A_1) \\
\text{EVPI} &= EMV(C) - \text{EMV(UC)} \\
&= \text{Rs } 71,00,000 - \text{Rs } 61,00,000 \\
&= \text{Rs } 10,00,000
\end{aligned}$$

This indicates that the marketing manager should decide to run the special promotion. This is known as prior analysis as the expected monetary value is based on the assignment of probabilities by the marketing manager without any additional information.

Posterior Analysis Suppose in the foregoing example the marketing manager wishes to revise his prior probabilities on the basis of additional information received by him. Posterior analysis uses both present and additional information. In this analysis, posterior or revised probabilities are used to ascertain the expected monetary value.

The following table gives the posterior probabilities on the basis of additional information.

Table 22.13 An Example of Posterior Analysis

	Probabilities			
Outcome S_i	*Prior $P(S_i)$*	*Conditional $P(R/S_i)$*	*Joint $P(R \text{ and } S_i)$*	*Posterior $P(S_i/R)$*
(1)	*(2)*	*(3)*	*(4)*	*(5)*
S_1 : Very favourable	0.7	0.6	0.42	0.894
S_2 : Favourable	0.1	0.3	0.03	0.064
S_3 : Unfavourable	0.2	0.1	0.02	0.042
Totals	**1.0**	**1.0**	**0.47**	**1.000**

R shows very favourable pre-test result.

In the first column of the table, three possible outcomes have been identified. The second column gives prior probability of each of the three possible outcomes. Now suppose a pre-test result on this promotion was favourable. In view of this, the manager must assess the conditional probability of getting a favourable pre-test given the various possible outcomes. Column 3 of the table gives such probabilities. The joint probability of R and S_i i.e., $P(R \text{ and } S_i)$ is obtained by multiplying the probabilities in columns 2 and 3. The joint probabilities for all the three possible outcomes are shown in column 4. The last column gives the posterior probabilities.These have been calculated by applying Bayes' rule.

Symbolically,

$$P(S_i/R) = \frac{P(R \text{ and } S_i)}{P(R)}$$

These are the probabilities of various outcomes given the test results. Thus, the posterior probability of very favourable outcomes (S_1) given favourable test result (R) = 0.42/0.47 = 0.894.
Similarly, the posterior probabilities of S_2 and S_3 can be calculated.

The posterior EMV(A_1) is calculated as follows:

$$\begin{aligned}\text{Posterior EMV(A1)} &= (\text{Rs } 1{,}00{,}00{,}000)(0.894) + (\text{Rs } 10{,}00{,}000 \times 0.064) \\ &\quad + (\text{Rs} - 50{,}00{,}000 \times (0.042) \\ &= \text{Rs } 89{,}40{,}000 + 64{,}000 - \text{Rs } 2{,}10{,}000 \\ &= \text{Rs } 90{,}04{,}000 - \text{Rs } 2{,}10{,}000 \\ &= \text{Rs } 87{,}94{,}000 \\ \text{Posterior EVPI} &= \text{EMV (C)} - \text{EMV (UC)} \\ &= \text{Rs } 90{,}04{,}000 - \text{Rs } 87{,}94{,}000 \\ &= \text{Rs } 2{,}10{,}000 \\ \text{Prior EVPI} &= \text{Rs } 10{,}00{,}000\end{aligned}$$

It will be seen that prior EVPI was higher (Rs 10,00,000) than posterior EVPI (Rs 2,10,000). This makes sense as it shows that on account of the new information, the degree of uncertainty has declined. Consequently the value of additional information has reduced.

Pre-posterior Analysis Having discussed posterior analysis, we now turn to pre-posterior analysis which is helpful in evaluating the worth of research before it is undertaken.

Suppose that in our earlier example the marketing manager was conducting a test market for a special soft drink promotion. The test would cost Rs 80,000. Figure 22.5 presents the structure of the

problem in the form of a decision tree. It may be noted that the tree diagram shows the decision problems in proper sequence. The first problem is whether to have a test market or not. This is followed by the market test (if it is undertaken); then the decision alternatives about the promotion and finally the possible outcomes of these decision alternatives. What follows is a step-by-step procedure with pre-posterior analysis.

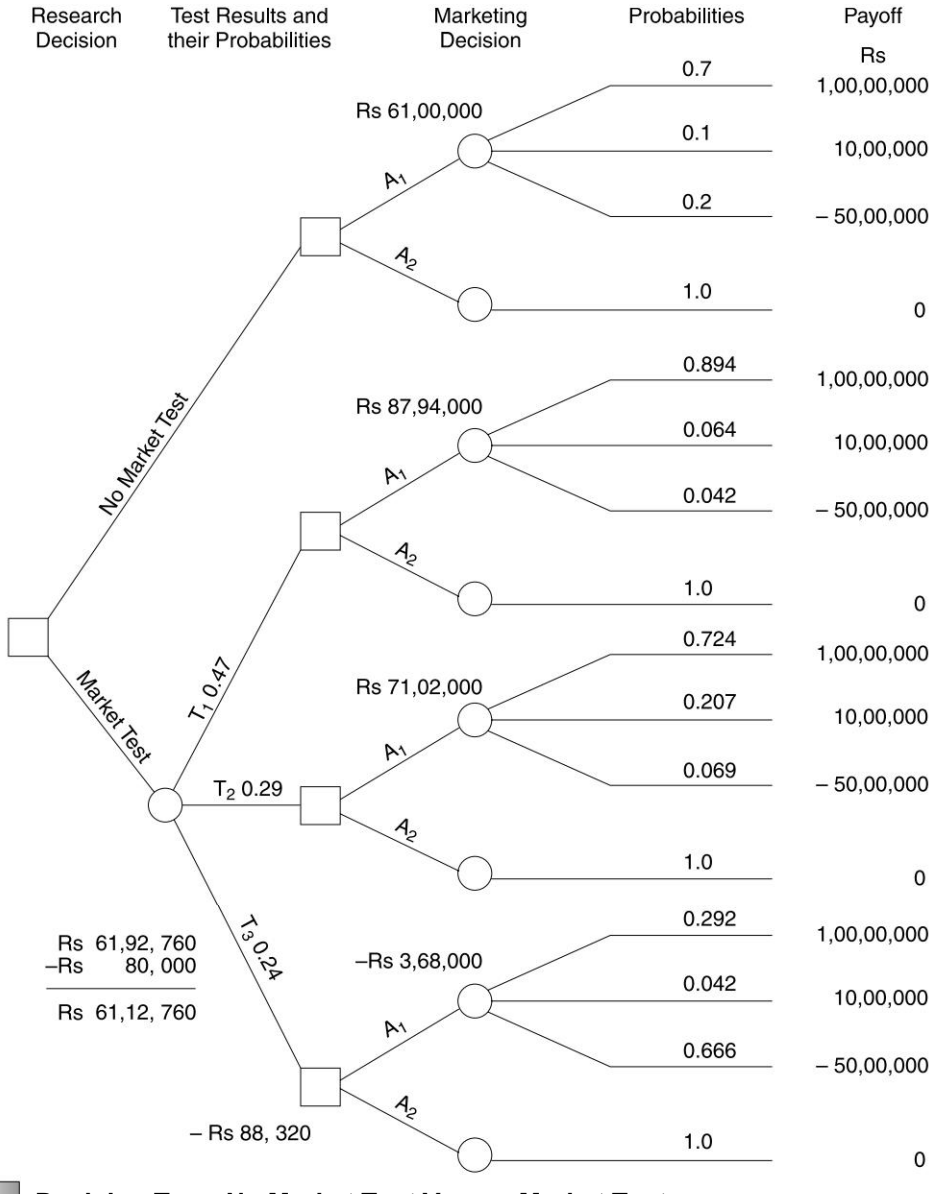


Fig. 22.5 Decision Tree: No Market Test Versus Market Test

Step 1 The marketing manager thinks that there are likely to be three test market outcomes: **(i)** a 15 per cent increase in sales (T_1) **(ii)** a 5 per cent increase in sales (T_2), and **(iii)** no increase in sales (T_3). He has now to obtain conditional probabilities of test market results as shown in Table 22.14.

Table 22.14 Conditional Probabilities of Test Market Results

States of nature	*Test results*		
	T_1 (+15%)	T_2 (+5%)	T_3 (±0%)
S_1 (Very favourable)	0.6	0.3	0.1
S_2 (Favourable)	0.3	0.6	0.1
S_3 (Unfavourable)	0.1	0.1	0.8

The conditional probability indicates that if the test market is run as proposed and receives a favourable reaction of consumers, there will be certain probability of occurrence of T_1, T_2 and T_3. Thus, the probability of having a +15 per cent test result given a very favourable test market $P(T_1/S_1)$ is 0.6.

These conditional probabilities $P(T_i/S_{is})$ are multiplied by the prior probabilities $P(S_i)$, to give joint probabilities. For example, $P(S_1 \text{ and } T_1) = P(S_1) \times P(T_1/S_1) = 0.6 \times 0.7 = 0.42$. All these joint probabilities are shown in Table 22.15. The marginal probability of each T_1 can be obtained by adding the joint probabilities where T_i occurs. Thus, the marginal probability of T_i, $P(T_1)$ is 0.47; of T_2, $P(T_2)$ is 0.29; and T_3, $P(T_3)$ is 0.24. These probabilities are also shown in Table 22.15.

Table 22.15 Joint Probabilities of States and Test Market Results

States of nature	*Test results*			*Marginal probabilities*
	T_1 (+15%)	T_2 (+5%)	T_3 (±0%)	
S_1 (Very favourable)	0.42	0.21	0.07	0.7
S_2 (Favourable)	0.03	0.06	0.01	0.1
S_3 (Unfavourable)	0.02	0.02	0.16	0.2
Marginal probabilities	0.47	0.29	0.24	1.0

Step 2 The prior probabilities of the possible outcomes can now be revised. These probabilities are revised by using Bayes' rule. For example,

$$P(S_1/T_1) = \frac{P(S_1 \text{ and } T_1)}{P(T_1)} = \frac{0.42}{0.47} = 0.894$$

$$P(S_2/T_1) = \frac{P(S_2 \text{ and } T_1)}{P(T_1)} = \frac{0.03}{0.47} = 0.064$$

$$P(S_3/T_1) = \frac{P(S_3 \text{ and } T_1)}{P(T_1)} = \frac{0.02}{0.47} = 0.042$$

Similar calculations can be easily made for the S_i's given T_2 and T_3. These results are shown on the decision tree. The expected pay-off of each alternative is calculated again using the posterior or revised probabilities. On the basis of these calculations, the best courses of action and pay-offs are

T_1 = Rs 87,94,000; T_2 = Rs 71,02,000; and T_3 = Rs 0. These three best pay-offs are then multiplied by their associated $P(T_i)$.

Step 3 The sum of the best outcomes, each multiplied by its respective $P(T_i)$, comes to T_1 = Rs 41,33,180; T_2 = Rs 20,59,580 and T_3 = Rs 0.

Step 4 Now, it is possible to calculate the expected monetary value of sample information (EVSI). This is shown below:

EVSI = EMV (with test market) – EMV (without test market)
= Rs 61,92,760 – 61,00,000
= Rs 92,760

This is the maximum amount that can be paid for this research.

Step 5 The expected monetary gain can also be calculated as shown below:

EGSI = EVSI – CI
= Rs 92,760 – Rs 80,000
= Rs 12,760

Step 6 Since the expected monetary gain is positive, it is advisable to undertake the test market.

It may be pointed out that one may consider alternative designs for test market, each having different costs and conditional probabilities. In view of these factors, the EGSI will also be different for each alternative design. In the same manner, we have to calculate EGSI for each research design or test market. Obviously our choice will be in favour of that design which gives the highest EGSI.

Advantages of Bayesian Analysis It is felt in some quarters that the Bayesian analysis is an abstruse method and is more an academic exercise than a realistic decision method. It is true that such a decision method involves considerable effort. At the same time, the Bayesian analysis covers more information than traditional methods and, as such, it should enable us to arrive at correct decisions more frequently than would otherwise be possible. This apart, a major advantage of the Bayesian approach is that it enables us to carry out the analysis in a sequential fashion. Data obtained from one sample survey can be used as the prior information when new information becomes available. Then by using Bayes' theorem, the two sources can be combined. Again, this being posterior distribution, it can be used as prior information when further new data become available. Thus, the sequence continues.

Additional Examples

Example 22.11 The following table gives payoff for actions A_1, A_2, and A_3 corresponding to states of nature S_1 and S_2, whose chances are 0.6 and 0.4, respectively:

States of nature	*Actions (Rs lakh)*		
	A_1	A_2	A_3
S_1	16	20	18
S_2	19	15	12

Find decisions under (i) Maximin criterion, and (ii) EMV criterion.

Solution

(i) Minimum value for each action

A_1	16
A_2	15
A_3	12

Maximum of these values is 16, i.e., A_1. Hence, under Maximin criterion, A_1 should be chosen.

(ii) EMV Criterion

$S_1 = (16 \times 0.6) + (20 \times 0.6) + (18 \times 0.6)$

$= 9.6 + 12 + 10.8$

$S_2 = (19 \times 0.4) + (15 \times 0.4) + (12 \times 0.4)$

$= 7.6 + 6 + 4.8$

A_1	A_2	A_3
9.6	12	10.8
7.6	6	4.8
17.2	18	15.6

Since maximum *EMV* of 18 is against A_2, the decision should be in favour of A_2.

Example 22.12 Apply **(a)** Maximax **(b)** Maximin and **(c)** Minimax regret to the following payoff matrix.

Decision alternatives	*States of nature*		
	A	*B*	*C*
X	7	5	3
Y	10	– 6	– 2
Z	– 9	7	5

Solution

(a) Maximax X : 7, Y : 10 and Z : 7

These are the maximum payoffs for each decision alternative. The decision should be in favour of *Y* as it is the maximum.

(b) Maximin criterion: Minimum payoff in each case

X : 3 Y : –6 Z : –9

Of these the maximum is to be chosen. Hence the decision should be in favour of *X*.

(c) Minimax regret criterion:

Regrets are shown in the following table.

	A	*B*	*C*
X	3	2	2
Y	0	13	7
Z	19	0	0

It will be seen that the minimum value of regrets among the maximum regrets happens to be 3 in favour of decision *X*.

Example 22.13 Three types or souvenirs can be sold outside the stadium. From the following conditional table, construct the opportunity loss table. (Sales are dependent on the winning team.)

	Types of souvenirs		
	I (Rs)	*II (Rs)*	*III (Rs)*
Team A wins	1,200	800	300
Team B wins	250	700	1,100

Point out which type of souvenir should be bought if probability of Team A's winning is 0.6.

Solution

(i) Opportunity Loss Table

	Type of Souvenir		
	I (Rs)	*II (Rs)*	*III (Rs)*
Team *A* wins	1200 – 1200 = 0	1200 – 800 = 400	1200 – 300 = 900
Team *B* wins	1100 – 250 = 850	1100 – 700 = 400	1100 – 1100 = 0

(ii) P (Team A wins) = 0.6
P (Team B wins) = 1 – 0.6 = 0.4

Souvenir	*Expected opportunity loss*
I	$(0.6 \times 0) + (0.4 \times 850)$ = Rs 340
II	$(0.6 \times 400) + (0.4 \times 400) = 240 + 160$ = Rs 400
III	$(0.6 \times 900) + (0.4 \times 0)$ = Rs 540

Since the expected opportunity loss of Rs 340 is the lowest in case of souvenir I, it should be bought.

Example 22.14 The following table gives the payoffs of the acts A, B, C and the states of nature P, Q, R. The probabilities of the states of nature are 0.2, 0.3 and 0.5, respectively. Calculate the expected monetary value (EMV) and select the best act:

States of nature	*Acts (Rs lakh)*		
	A	*B*	*C*
P	200	–100	–300
Q	250	300	–500
R	300	500	600

Solution Expected Monetary Value (EMV) for the three Acts is worked out as follows:

Act A

$$(200 \times 0.2) + (250 \times 0.3) + (300 \times 0.5)$$
$$= 40 + 75 + 150$$
$$= \text{Rs } 265 \text{ lakh}$$

Act B

$$(-100 \times 0.2) + (300 \times 0.3) + (500 \times 0.5)$$
$$= -20 + 90 + 250$$
$$= \text{Rs } 320 \text{ lakh}$$

Act C

$$(-300 \times 0.2) + (-500 \times 0.3) + (600 \times 0.5)$$
$$= -60 - 150 + 300$$
$$= \text{Rs } 90 \text{ lakh}$$

It will be seen from these calculations that Act B is best as it gives the highest EMV.

Example 22.15 Company A has to decide whether or not to introduce a new product, depending on the strategy of the rival Company B. The marketing manager of Company A has prepared the following conditional payoff table.

	Company B's strategy	
	C_1	C_2
Probability →	*0.7*	*0.3*
Strategy S_1	Rs 8 million	Rs 12 million
Strategy S_2	Rs 5 million	Rs 16 million

The notations are: C_1 = Company B introduces new product; C_2 = Company B does not introduce new product; S_1 = Company A decides to introduce new product; S_2 = Company A decides not to introduce new product.

Which of the two strategies Company A should choose?

What is the expected value of perfect information?

Solution We have to first find out the EMV of the two strategies.

Strategy S_1

$$EMV = (\text{Rs } 8 \text{ million} \times 0.7) + (\text{Rs } 12 \text{ million} \times 0.3)$$
$$= \text{Rs } 5.6 + 3.6 \text{ million}$$
$$= \text{Rs } 9.2 \text{ million}$$

Strategy S_2

$$EMV = (\text{Rs } 5 \text{ million} \times 0.7) + (\text{Rs } 16 \text{ million} \times 0.3)$$
$$= \text{Rs } 3.5 + 4.8 \text{ million}$$
$$= \text{Rs } 8.3 \text{ million}$$

Since EMV in respect of strategy 1, i.e. S_1 is higher than that of S_2, the Company *A* should choose S_1.

In order to calculate EVPI, it is presumed that the Company A is in possession of complete information.The calculation will be as follows :

$$EVPI = EMV\ (C) - EMV\ (UC)$$
$$= (\text{Rs } 8 \text{ million} \times 0.7) + (\text{Rs } 16 \text{ million} \times 0.3) - \text{Rs } 9.2 \text{ million}$$
$$= \text{Rs } 5.6 + 4.8 - 9.2 \text{ million}$$
$$= \text{Rs } 1.2 \text{ million}$$

Example 22.16 A manufacturing company is considering whether or not to launch a special promotion campaign for making its newly developed product more popular. It thinks that there could be three different types of consumer reactions along with the possible gain or loss, as follows: very favourable Rs 12,00,000; moderately favourable Rs 5,00,000; and unfavourable (–)Rs 3,00,000. It assigns probabilities to these three situations as 0.4, 0.3 and 0.3, respectively. Of course, if the company decides not to launch a special promotion campaign, it neither gains nor loses.

(i) Should the company be advised to launch a special promotion campaign?
(ii) What is the expected value of perfect information?

Solution Three different types of consumer reactions

		Probability
Very favourable	Rs 12,00,000	0.4
Moderately favourable	Rs 5,00,000	0.3
Unfavourable	–Rs 3,00,000	0.3

$12,00,000 \times 0.4 = Rs\ 4,80,000$
$5,00,000 \times 0.3 = Rs\ 1,50,000$
$-3,00,000 \times 0.3 = -Rs\ 90,000$

EMV (UC) = Rs 4,80,000 + Rs 1,50,000 – Rs 90,000
= Rs 5,40,000

EMV (C) = Rs 4,80,000 + Rs 1,50,000
= Rs 6,30,000

EVPI = EMV (C) – EMV (UC)
= Rs 6,30,000 – 5,40,000
= Rs 90,000 (Rs 0.9 lakh)

(i) It is advisable for the company to launch a special promotion campaign as it gives a high expected pay off of Rs 5.4 lakh.
(ii) EVPI would be Rs 90,000 or Rs 0.9 lakh.

Example 22.17 A manufacturing company is faced with the problems of choosing four products for manufacturing. The potential demand for each product may turn out to be good, satisfactory and poor. The probabilities, estimated for each type of demand, are given below:

	Probabilities of types of demand		
Products	*Good*	*Satisfactory*	*Poor*
A	0.60	0.20	0.20
B	0.75	0.15	0.10
C	0.60	0.25	0.15
D	0.50	0.20	0.30

The estimated profit or loss under the different states of demand in respect of each product may be taken as:

Products	*Good Rs*	*Satisfactory Rs*	*Poor Rs*
A	40,000	10,000	1,100
B	40,000	20,000	–7,000
C	50,000	15,000	–8,000
D	40,000	18,000	15,000

Prepare the expected value table and advise the company about the choice of the product to manufacture.

Solution

Expected Value Table

	Products			
State of nature	*A (Rs)*	*B (Rs)*	*C (Rs)*	*D (Rs)*
Good	40000 × 0.60 = 24,000	40000 × 0.75 = 30000	50000 × 0.60 = 30000	40000 × 0.50 = 20000
Satisfactory	10000 × 0.20 = 2000	20000 × 0.15 = 3000	15000 × 0.25 = 3750	18000 × 0.20 = 3600
Poor	1100 × 0.20 = 220	–7000 × 0.10 = –700	–8000 × 0.15 = –1200	15000 × 0.30 = 4500
Total	**Rs 26,220**	**Rs 32,300**	**Rs 32,550**	**Rs 28,100**

From the foregoing calculations, it is clear that the expected payoff of product *C* is the highest. As such the company's choice should be in favour of product *C*.

Example 22.18 A soft-drink manufacturing company is considering whether or not to launch a special promotion campaign. Its decision is summarised in the table given below:

Possible consumer reactions	*Alternative courses of action*		*Probabilities of consumer reactions*
	Launch a special promotion campaign (Rs)	*Do not launch a special promotion campaign (Rs)*	
Very favourable	9,00,000	0	0.4
Favourable	4,00,000	0	0.3
Unfavourable	–3,00,000	0	0.3

(i) What is the expected value of perfect information?

(ii) If the probabilities of very favourable, favourable and unfavourable consumer reactions were 0.5, 0.3 and 0.2, respectively; and the loss in case of unfavourable reaction of consumers was Rs 4,00,000, what would be the expected value of perfect information?

Solution

(i) Calculation of EMV under uncertainty

EMV (UC) = Rs (9 × 0.4) + (4 × 0.3) + (–3 × 0.3) lakh
= Rs 3.6 + 1.2 – 0.9 lakh
= Rs 3.9 lakh

Calculation of EMV under certainty

EMV (C) = Rs (9 × 0.4) + (4 × 0.3) lakh
= Rs 3.6 + 1.2 lakh = Rs 4.8 lakh

EVPI = EMV (C) – EMV (UC)
= 4.8 – 3.9 = 0.9 lakh rupees or Rs 90,000

(ii) EMV (UC) = Rs (9 × 0.5) + (4 × 0.3) + (–3 × 0.2) lakh
= Rs 4.5 + 1.2 – 0.6 lakh
= Rs 5.7 – 0.6 lakh = Rs 5.1 lakh
EMV (C) = Rs (9 × 0.5) + (4 × 0.3) lakh
= Rs 4.5 + 1.2 lakh
= Rs 5.7 lakh
EVPI = EMV (C) – EMV (UC)
= Rs 5.7 lakh – Rs 5.1 lakh = 0.6 lakh or Rs 60000.

It may be noted that as uncertainty in the second case has reduced, we find that EVPI has also reduced from Rs 90,000 to Rs 60,000.

Example 22.19 Under a sales promotion programme, it is proposed to allow sale of newspapers on buses during off-peak hours. The vendor can purchase the newspaper at a concessional rate of Rs 1.25 per copy against the selling price of Rs 1.50. The vendor has estimated the following probability distribution for the number of copies demanded.

Number of copies demanded	15	16	17	18	19	20
Probability	0.04	0.19	0.33	0.26	0.11	0.07

How many copies should he order so that his expected profit is maximum?

Solution

We set up the following table giving the necessary calculations relating to different number of copies demanded.

Number of copies	*Cost @ Rs.1.25 (Rs)*	*S.P @ Rs.150 (Rs)*	*Profit (Rs)*	*Probability*	*Expected profit (Rs)*
15	18.75	22.50	3.75	0.04	0.15
16	20.00	24.00	4.00	0.19	0.76
17	21.25	25.50	4.25	0.33	1.40
18	22.50	27.00	4.50	0.26	1.17
19	23.75	28.50	4.75	0.11	0.52
20	25.00	30.00	5.00	0.07	0.35

The last column of the table shows the expected profit, which is the highest against 17 copies. Hence, he should order 17 copies of the paper.

Example 22.20 An investment company is considering four investment proposals for a client, shares, bonds, real estate and saving certificates. These investments will be held for one year. The post data regarding the four alternatives give the following inference:

Shares: There is a 20 per cent chance that shares will decline by 10 per cent, 30 per cent chance that they will remain stable and 50 per cent chance that they will increase in value by 10 per cent (The shares do not pay any dividends).

Bonds: There is a 40 per cent chance that the bonds will increase in value by 5 per cent, and 60 per cent chance that they will remain stable and they will yield 12 per cent return.

Real estate: This has 20 per cent chance of increasing 30 per cent in value, 25 per cent chance of increasing 20 per cent in value, 40 per cent chance of increasing 10 percent in value, 10 per cent chance of remaining stable and 5 per cent chance of losing 5 per cent of the value.

Saving certificate: They yield 8.5 per cent with certainty.

Evaluate the alternate proposals and advise the client.

Solution

The client wants to invest Rs 1,00,000 for the period of one year.

Evaluation of the four options are given below:

Shares

20% chance that they will decline by 10 %.	Value = Rs 90,000
30% chance that they will remain stable	Value = Rs 1,00,000
50% chance that they will increase by 10%.	Value = Rs 1,10,000

Hence, overall return will be:

Rs (90,000 × 0.2) + (1,00,000 × 0.3) + (1,10,000 × 0.5)

= Rs 18,000 + 30,000 + 55,000

= Rs1,03,000.

Gain of Rs 3,000.

Bonds

40% chance of increase by 5%.	Value = Rs 1,05,000
60% chance of remaining stable 4%.	Value = Rs 1,00,000

Yield @ 12% on Rs 1,00,000 is Rs 12,000

Hence, overall return will be Rs (1,05,000 × 0.4) + (1,00,000 × 0.6) + 12,000

= Rs 42,000 + 60,000 + 12,000

= Rs 1,14,000

Gain Rs 14,000.

Real estate

20% chance of increase by 30%.	Value = Rs. 1,30,000
25% chance of increase by 20%.	Value = Rs. 1,20,000
40% chance of increase by 10%.	Value = Rs. 1,10,000
10% chance of remaining stable.	Value = Rs. 1,00,000
5% chance of loss in value 5%	Value = Rs. 95,000

Hence, overall return will be:

Rs (1,30,000 × 0.2) + (1,20,000 × 0.25) + (1,10,000 × 0.4) + (1,00,000 × 0.1) + (95,000 × 0.05)

= Rs 26,000 + 30,000 + 44,000 + 1,000 + 4,750

= Rs 1,05,750 Gain Rs 5,750

Saving certificates

Yield 8.5%

Hence, return will be: Rs $\frac{8.5}{100}$ × 1,00,000 = Rs 8,500

Hence, Gain = Rs 8,500

The position that emerges from the above calculations is as follows:

Options for Investment	*Return on Rs.4,00,000-investment*
Shares	Rs. 3,000
Bonds	Rs.14,000
Real estate	Rs. 5,750
Saving certificates	Rs. 8,500

The calculations show that out of the four options for investment, bonds will give the maximum return to the client. Therefore, he should be advised to invest in bonds.

Conclusion

We have discussed some methods that are used in decision analysis. In order to perform decision analysis properly, we should have a rudimentary understanding of probability and of expected values. In some cases, we may find that the information is inadequate. In such cases, additional information may have to be collected either by conducting a sample survey or by some other method. This additional information is then used along with the earlier information for using Bayes' theorem.

At the end, it may be emphasized that decision analysis offers an excellent aid in arriving at a final decision. For instance, when a company is interested to introduce a new product in the market, decision analysis can be very helpful in deciding whether it should be introduced. When the answer is in favour of introducing the product, decision analysis can also help in deciding the scale of operation. Likewise,when a company is interested to invest a large amount, it should consider all the possible outcomes and the chance of their occurrence. Thus, there can be numerous situations where the use of decision analysis can provide a better understanding of the problem as well as of the possible outcomes and their respective expected payoffs. Armed with all this information, a company is in a far better position for coming to a right decision.

GLOSSARY

Bayesian analysis	A statistical procedure for incorporating revised probabilities and for determining cost of additional information in decision-making.
Certainty	The decision environment that has only one state of nature. As such it has no alternative course of action.
Conditional profit	The profit based on a certain combination of decision alternative and state of nature.
Decision point	A stage in decision-making process where a decision has to be made.
Decision tree	A graphic device that shows decision environment, indicating decision alternatives, states of nature along with their probabilities and conditional benefits and losses.
Expected profit	The aggregate of the conditional profits for a given decision alternative, each weighted by the probability of its occurrence.

Expected profit with perfect information	The expected value of profit resulting from the complete information. This implies that no additional information can be made available.
Expected value of perfect information (EVPI)	This is the difference between the expected profit with perfect information and expected profit under conditions of risk.
Expected-value criterion	A criterion for decision-making that is based on the expected value for each decision alternative.
Jacob Bernoulli method	The principle of insufficient reason whereby the decision-maker assigns equal probabilities to different states of nature.
Maximax criterion	The principle of selecting the best payoff from amongst the various alternatives that yield different profits.
Maximin criterion	The principle of selecting the best (the maximum) from the set of worst (the minimum) profits.
Minimax criterion	The principle of selecting the minimum cost from amongst the maximum costs of different states of nature.
Node	A point on the decision tree indicating a chance event or a decision taking place.
Opportunity loss	The amount of payoff foregone by the decision-maker by not adopting the optimal course of action.
Payoff	The amount of benefit that results from a certain combination of a decision alternative and a state of nature.
Rollback	A method whereby the decision-maker proceeds from right to left in the decision tree to calculate an optimal alternative.
State of nature	An event that is likely to occur in future over which the decision-maker has no control.
Utility	The value of a certain outcome or payoff to the decision-maker; the pleasure or displeasure he gets from an outcome.

LIST OF FORMULAE

1. EVPI = EMV(C) – EMV(UC)
where EVPI = Expected value of perfect information
EMV(C) = Expected monetary value under certainty
EMV(UC) = Expected monetary value under uncertainty

2. EP + EOL = EPPI
where EP = Expected payoff of the optimal decision without perfect information
EOL = Expected opportunity loss
EPPI = Expected payoff of perfect information

3. EU = Utility × Probability
where EU = Expected value of utility

4. Formula (1) can also be written as EVPI = EPPI – EP

5. EVPI = EOL

6. EGSI = EVSI – CI

where EGSI = Expected gain with sample information
EVSI = Expected value of sample information
CI = Cost of information

7. EVSI = EP (with sample information) – EP (without sample information)

QUESTIONS

22.1 Given below are twelve statements. Indicate in each case whether it is true or false:

(a) A major advantage of decision-tree approach is that it considers every outcome regardless of its being desirable or undesirable.

(b) Whatever may be the events, none is beyond the control of the decision-maker.

(c) We should not wait for the availability of perfect information as it is rarely available.

(d) Payoffs are only the positive benefits of different alternatives available to the decision-maker.

(e) A circle on a decision tree indicates a decision point.

(f) Decision theory enables us to select the best possible outcome from amongst the several alternatives available.

(g) The maximax criterion for decision-making under uncertainty is an optimistic criterion.

(h) The criterion of realism lies between maximax and maximin criterion.

(i) Jacob Bernoulli method assigns equal probabilities to the different states of nature.

(j) A decision tree depicts the sequences of action-event combinations but not in a systematic manner.

(k) If a trader can earn Rs 900 per day with perfect information, then EVPI = Rs 900.

(l) Decisions are always made to maximise the expected monetary value and not the expected utility.

Multiple Choice Questions (22.2 to 22.12)

22.2 Decision theory deals with

(a) Decisions that are exclusively quantity-oriented
(b) Making decisions under conditions of uncertainty
(c) The worth of additional information to the decision-maker
(d) (a) and (b)
(e) (b) and (c)

22.3 Decision-making is done under conditions of

(a) certainty (b) uncertainty
(c) risk (d) (b) and (c)
(e) (a), (b) and (c)

22.4 Which of the following cretiria is *not* used for decision-making under uncertainty?

(a) maximax (b) minimize expected loss
(c) minimax (d) maximin

22.5 A person who is reluctant to take risk, prefers

(a) Those situations which have high expected values

(b) To take large risks to get large gains
(c) To act any time when the expected value is positive
(d) None of these

22.6 For a particular decision, the total benefit of a new plant is Rs 25,350,000. If the expected net benefit of this plant is Rs 18,280,000, what is the cost of the plant?
(a) Rs 70,00,000 (b) Rs 75,00,000
(c) Rs 70,70,000 (d) Cannot be determined from the information given

22.7 EPPI is equal to
(a) EOL (b) EMV(C) (c) EVPI
(d) EP + EOL (e) None of these

22.8 EVPI is equal to
(a) EMV(C) – EMV(UC) (b) EOL
(c) EPPI – EP (d) All of these
(e) None of these

22.9 The decision tree is useful in decision-making because
(a) It guides the analyst to approach the problem in an orderly manner
(b) It examines all desirable options
(c) It can be used with a computer
(d) It is a graphical device that facilitates in understanding the problem
(e) All of these
(f) (a), (c) and (d)

22.10 An investor whose utility curve is linear is
(a) cautious and conservative (b) risk-seeker
(c) risk-neutral (d) none of these

22.11 Which quality of the decision maker leads to a good decision?
(a) High intelligence (b) Long experience
(c) Sound quantitative understanding (d) Strong intuition
(e) All of these (f) (a), (b) and (c), but not (d)

22.12 For decision-making, a coefficient or an index of optimism (α) is used as a criterion in
(a) Maximax (b) Maximin (c) Minimax (d) Realism
(e) None of the above

22.13 What are the main elements of a decision problem?

22.14 Explain the following terms giving a suitable example in each case:
(a) Payoff table
(b) Opportunity loss
(c) State of nature
(d) Action space

22.15 Describe some methods that are useful for decision-making under uncertainty. Illustrate each by an example.

22.16 What is the 'expected value' approach to decision-making?

22.17 Explain the following terms giving a suitable example in each case:
(a) The minimax principle

(b) The maximin principle

(c) Expected value of perfect information

22.18 What is meant by a 'tree diagram'? How is it used in estimating the expected value of information?

22.19 What are the advantages and limitations of the tree-diagram approach for decision-making?

22.20 What is a 'payoff table'? What are its uses?

22.21 Distinguish between payoff table and regret table, giving suitable example.

22.22 Explain briefly the 'statistical decision theory'.

22.23 Explain the criteria of maximum and minimum regret in the context of decision theory.

22.24 Define EVPI. Explain, giving a hypothetical example, as to how it is calculated.

22.25 What is the importance of utility as a basis of decision-making? State the assumptions involved in the theory of utility.

22.26 What do you understand by the criterion of realism? Which one of the three products you would choose for the data given below? You may choose the coefficient of optimism $\alpha = 0.7$.

Product	*Maximum value (Rs lakh)*	*Minimum value (Rs lakh)*
A	50	15
B	35	25
C	70	–10

22.27 It costs Rs 600 to test a machine. If a defective machine is installed, it costs Rs 12,000 to repair the damage resulting to the machine. Is it more profitable to install the machine without testing if it is known that 3 per cent of all machines produced are defective? Show the calculation.

22.28 A producer of boats has estimated the following distribution of demand for a particular kind of boats:

Likely demand	0	1	2	3	4	5	6
Probability	0.14	0.27	0.27	0.18	0.09	0.04	0.01

Each boat costs him Rs 15,000 and he sells it for Rs 20,000. Any boat left unsold at the end of the season must be disposed off at half the price, that is, Rs 10,000 each.

(a) How many boats should he stock so as to maximise his expected payoff?

(b) What will be the EVPI?

22.29 Suppose that a decision maker faced with three decision alternatives and four states of nature. Given, the following profit payoff table:

(Figures in '000 Rs)

States of nature / *Acts*	S_1	S_2	S_3	S_4
A_1	16	10	12	7
A_2	13	12	9	9
A_3	11	14	15	14

Assuming that he has no knowledge of the probabilities of occurrence of the states of nature, find the decisions to be recommended under each of the following criteria:

(i) Maximin, **(ii)** Maximax, **(iii)** Minimax regret

22.30 A contractor has to choose between two jobs. The first job promises a profit of Rs 2,00,000 with a probability of 0.7 or a loss of Rs 40,000 (due to strike and other delays) with a probability of 0.3. The second job would give him a profit of Rs 3,00,000 or a loss of Rs 50,000 with a probability of 0.5 in each case.

(a) Which job should the contractor choose if he wants to maximise his expected profit?

(b) Which job would the contractor probably choose if his business is in fairly bad shape and he will go broke unless he can make a profit of Rs 50,000 on his next job?

22.31 A retailer has shelf space for four highly perishable items, which are destroyed at the end of the day if they are not sold. The unit cost of the item is Rs 3, the selling price is Rs 5 and the profit thus is Rs 2 per item sold. Assuming that the probabilities of the demand for 0, 1, 2, 3, or 4 items are, respectively, 0.1, 0.2, 0.3, 0.3 and 0.1, determine how many items should the retailer stock so as to maximise his expected profits.

22.32 Referring to the question No. 22.31 given above, suppose that the retailer has no idea about the potential demand for the item. How many of the items should he stock so as to minimise the maximum losses to which he may be exposed? Discuss the reasonableness of this criterion in a problem of this kind.

22.33 Suppose that we are asked to predict what proportion of the seniors of a very large high school will fail a college admission test, and all the information we have is that the proportion is not less than 0.06 and not greater than 0.14. What prediction would be 'best' if we wanted to minimise the maximum error?

22.34 A physician purchases a particular vaccine on Monday of each week. The vaccine must be used in the current week, otherwise it loses its utility and has to be discarded. The vaccine costs Rs 3 per dose and the physician charges Rs 5 per dose. In the past 30 weeks, the physician has administered the vaccine in the following quantities:

Doses per week	16	20	30	40
Number of weeks	5	10	8	7

How many doses the physician buys every week?

22.35 The marketing staff of a company has worked out the following payoff table concerning a proposal depending upon the rate of technological development in the next five years:

Technological development	*(Figures in Rs million)*	
	Accept proposal	*Reject proposal*
Considerable	3	2
Moderate	6	4
None	–2	3

The probabilities are 0.25, 0.45 and 0.30 for considerable, moderate, and none technological development, respectively. Find out which decision would be most appropriate for the company.

22.36 Given the following payoff matrix,

States of nature	*Probability*	*Acts*		
		X ('000 Rs)	*Y ('000 Rs)*	*Z ('000 Rs)*
P	0.3	–120	–80	100
Q	0.5	200	400	–300
R	0.2	260	–260	600

Using the expected monetary value criterion, decide which act can be chosen as the best.

22.37 Given the following matrix, find the best alternative using **(i)** Minimax criterion, **(ii)** Minimin criterion, and **(iii)** Minimax regret criterion:

Alternatives	*States of nature (Rs lakh)*			
	S_1	S_2	S_3	S_4
A_1	1	3	8	5
A_2	2	5	4	7
A_3	4	6	6	3
A_4	6	8	3	5

22.38 Based on the following payoff (profit) matrix:

(*Rs million*)

States of nature	*A*	*B*	*C*	*D*
P	5	10	18	25
Q	8	7	8	23
R	21	18	12	21
S	30	22	19	15

Determine the alternative to be chosen under:

(i) Minimax, **(ii)** Maximin, and **(iii)** Minimax regret criteria.

22.39 Given the following payoff matrix:

Acts	*Payoff ('000 Rs) States of nature*	
	Cold weather	*Hot weather*
Sell cold drinks	50	100
Sell hot drinks	120	40

and given the probability of weather being hot is 0.8, set up the opportunity loss table and compute opportunity loss of each action. Select the best act.

22.40 Mr. Engineer, head of a small engineering firm, must decide how many engineers to hire as full-time engineering consultants for the next year. (He has decided that he will not bother with any part-time employees.) He knows from experience that the probability distribution of the number of consulting jobs his firm will get each year is represented by the number given below:

Consulting jobs	3	6	9	12
Probability	0.3	0.2	0.4	0.1

He also knows that each engineer hired will be able to handle exactly 3 consulting jobs per year. The salary of each engineer is Rs 60,000. Each consulting job is worth Rs 30,000 to his firm. Each consulting job that the firm is awarded but cannot complete costs the firm Rs 10,000 in future business lost. How many engineers should Mr. Engineer hire using EMV criterion? Write down the methodology and carry on the computations.

22.41 The research department of M/s HL has recommended that the marketing department launches shampoos of three different types. The marketing manager has to decide the type of shampoo to be launched under the following estimated pay off for various levels of sales:

	Estimate levels of sales		
Types of shampoo	*15,000*	*10,000*	*5,000*
Egg shampoo	30	10	10
Clinic shampoo	40	15	5
Deluxe shampoo	55	20	30

Decide, based on Maximin, Minimax, Laplace and Regret.

22.42 The expected pay off of an ordering situation is summarised below:

	Demand				
Order size	*50*	*100*	*150*	*200*	*250*
75	950	1200	575	–675	–1425
150	50	1700	2000	2250	1600
225	–850	800	2550	3550	4525
300	–1800	600	1800	2000	5000

Find the best decision using

(a) Maximum criterion

(b) Regret criterion

(c) Hurwicz criterion

22.43 The daily net profit for each combination of order size and demand is shown in the following table. Obtain the best action, based on the Hurwicz criterion. Also, compare the above action with the best action, based on minimax criterion.

	Demand				
Order size	*100*	*200*	*300*	*400*	*500*
100	100	1800	800	100	–500
300	250	1750	3250	2500	1750
500	–500	1000	2500	4000	4500
700	–1200	250	1750	3250	4750

22.44 Find the Laplace criterion for the problem where the pay off for various strategies of various nature are given below:

	Strategies		
Status nature	S_1	S_2	S_3
N_1	7,00,000	5,00,000	3,00,000
N_2	3,00,000	4,50,000	3,00,000
N_3	1,50,000	0	3000

22.45 Consider the following payoff (profit) matrix.

	State of Nature				
Strategy	N_1	N_2	N_3	N_4	N_5
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 criteria.

22.46 A decision matrix with cost data is given below:

	States of Nature			
Alternatives	S_1	S_2	S_3	S_4
a_1	1	3	8	5
a_2	2	5	4	7
a_3	4	6	6	3
a_4	6	8	3	5

Find the best alternative using (i) Minimax criterion (ii) Minimin criterion, and (iii) Minimax regret criterion.

22.47 A group of students raises 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 types of souvenirs from a supplier. Their sales are mostly dependent on which team wins the match. A conditional payoff table is given below:

	Type of Souvenir		
Winner's Team	*I*	*II*	*III*
	Rs.	Rs.	Rs.
Team A	1500	700	400
Team B	600	700	900

(i) Construct the opportunity loss table.

(ii) Which type of souvenir should the students buy if opportunity of Team A's winning is 0.6?

(iii) Find the cost of uncertainty.

22.48 A manufacturer is thinking of installing new machinery in his factory at the cost of Rs 10,00,000. He expects returns over a period of one year. The market conditions would be as follows:

State of Market	*Returns (in Rs)*	*Probability*
High	18,00,000	0.2
Medium	12,00,000	0.35
Low	8,00,000	0.3
Poor	3,00,000	0.15

Assist the manufacturer to reach a decision.

22.49 A company is trying to decide what size plant to build in a certain area. Three alternative plants with capacity of 20,000, 30,000 and 40,000 units, respectively, are being considered. Demand for product is uncertain, but management has assigned the probabilities listed below to five levels of demand. The table below also shows the profit for each alternative and each possible level of demand (output may exceed rated capacity).

Demand (units)	*Probability*	*Profit (in Rs crore) for different capacity plants*		
		20000 units	*30000 units*	*40000 units*
10,000	0.2	–4.0	–6.0	–8.0
20,000	0.3	1.0	0.0	–2.0
30,000	0.2	1.5	6.0	5.0
40,000	0.2	2.0	7.5	11.0
50,000	0.1	2.0	8.0	12.0

What size of plant should the company build?

22.50 The probability distribution of demand for cakes is as follows:

Number of cakes	0	1	2	3	4	5
Probability	0.05	0.10	0.25	0.30	0.20	0.10

If the cost per cake is Rs 30 and selling price is Rs 40, how many cakes should the baker make to maximise his profit. Assume that if the cake is not sold at the end of the day, its value is zero.

The McGraw-Hill Companies

23 CHAPTER QUALITY CONTROL

Learning Objectives

By the end of your work on this chapter, you should be able to

- understand the concept and importance of quality control
- set up different types of control charts to keep the process under control
- understand the concept and importance of total quality management
- understand the concepts of acceptance sampling, single, double and multiple sampling plans, and select the most appropriate sampling plan.

Chapter Prerequisites

Before starting work on this chapter, make sure you are fully conversant with

1. the use of the Binomial, Poisson and Normal distributions
2. the idea of sampling distribution and standard error
3. the concept of hypothesis testing

23.1 INTRODUCTION

The subject of 'quality control' has assumed considerable importance in recent years in the wake of globalisation of economies world over. As a result, there has been tremendous increase in competition amongst business enterprises both within and outside the country. Companies are now talking about TQM— total quality management, whose sphere has not remained confined to manufacturing industries alone but has extended to other sectors such as service industries.

What is Quality?

Perhaps the best way to begin our discussion on quality control is to understand what quality is. We often come across in conversation that a particular product is of 'top quality' or it is an 'export quality' product. This means that this particular product is distinctive. In some manufacturing companies, quality means that their product conforms to some predetermined physical characteristics, with particularly rigid specifications. However, quality must be defined in such a way that it recognises the requirements of the customer.

Quality has been defined in different ways by different experts but almost all those definitions emphasise that quality must meet the requirements of the customer. While quality is very vital for providing satisfaction to the customer, it goes far beyond this. For industrial and commercial organisations, quality is not only central to profitability but crucial to business survival. This aspect has assumed considerable importance in today's tough and challenging business environment. If quality is ignored or overlooked by these organisations then their continued existence is in danger.

Quality vs. Luxury A point worth noting at this stage is that quality should not be confused with luxury. Quality and luxury are two different concepts. Thus, an item or product may be cheap and yet it may have quality to meet the customer's requirements. In contrast, a luxury item may not satisfy the customer's requirements. The concept of quality, as we shall see in this chapter, indicates consistency, reliability, and lack of errors and defects in the product or service.

Quality Suffers from Variability It appears that the main factor that affects quality is variability in the process. This variability does not allow a factory to provide consistently a standard quality product. Prior to mass production, an individual worker or a few of them produced by hand, checking frequently if the product manufactured is coming out as they had conceived it. If it was distorted, they would again check where the fault lay, measure, and rework on it. However, when goods began to be manufactured on a mass scale, it became apparent that individual items could not be identical. It is almost impossible to eliminate variability completely. Such a situation poses a major problem in that the parts that are supposed to fit together would not fit. This shows that variability is the cause of poor quality.

Controlling Variability: Inspection vs Prevention

From the preceding discussion, it should be clear that some way should be found to reduce or rather control variability in the product. The various causes of variation in the product may be classified into two categories:

(a) Scientific and Identifiable
(b) Random and Chance

The first category comprises such causes as the use of defective raw material, poor equipment, poor workmanship, and so on. While the second category contains causes that do not have any bearing on the production process. The main purpose of our quality control exercise is to segregate specific and identifiable causes from the chance or random causes.

In the early days of mass production, inspection of the product and sorting out the defective ones was the chief method used for quality control. It was thought that the rejection of defective items would not cost much as the marginal cost of each unit was small. But gradually it became apparent that the costs of defective items were much higher than supposed earlier. This is because a number of people had to be employed to inspect the product besides losing the goodwill of the customers.

This realisation laid emphasis on doing things right at the very first time, focussing on the concept of *zero defects*. This means that efforts must be made to prevent defects at each stage of manufacturing a product or delivering a service. In order to achieve this, workers engaged in the production are given the responsibility to check their output rather than to pass it on for a final inspection. One major benefit of this approach is that workers feel a sense of pride and satisfaction for the responsibility given to them.

23.2 STATISTICAL PROCESS CONTROL

Statistical process control (SPC) is the application of appropriate statistical tools to processes to ensure continuous improvement in quality of products, services and productivity in the workforce. As far back as in 1920, Walter A. Shewart created a system for tracking variation and identifying its causes. His system of SPC was further developed by W. Edwards Deming, a one-time colleague of Shewart. Let us see what is the essence of the theory of statistical process control. It is nothing else but a differentiation of the causes of variation during the operation of any process. The basic approach to statistical process control is to identify a parameter that is easy to measure and is relevant to ascertain whether the quality is being maintained. For this purpose, control charts are used.

Control Charts

Control charts show a step-by-step approach to statistical process control. These are 'road maps' that are very helpful in solving the problems pertaining to quality. The underlying feature of such a chart is that there are certain SPC techniques that are most appropriate in each step.

Figure 23.1 gives a schematic control chart. It will be seen from Fig. 23.1 that the control chart has three zones. These are: *stable zone*, *warning zone* and *action zone*.

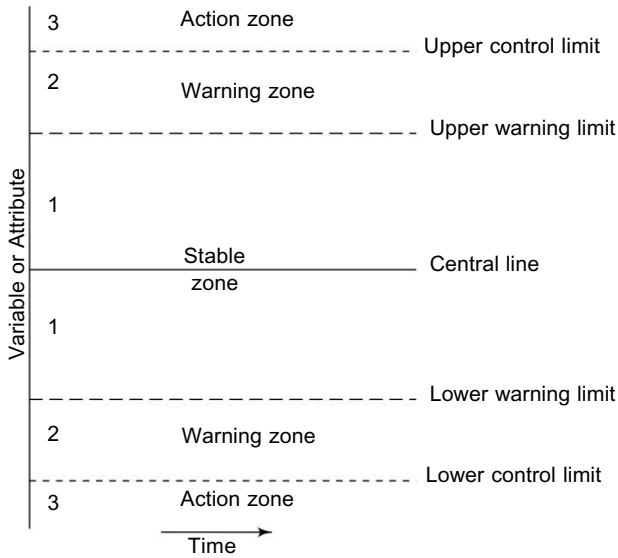


Fig. 23.1 **A Specimen of Control Chart**

Possibilities of Action in Control Chart Zones The action required depends on the zones in which the results fall. The possibilities are:

1. Nothing needs to be done in case of stable zone wherein variation occurs due to common causes only.
2. In respect of warning zone, there seem to be special causes of variation. There is a need for collecting more information and having a watchful eye on the process.
3. Action zone suggests that special causes of variation in the process are present. The situation demands further investigation and where appropriate the process needs to be adjusted.

These three situations can be compared to traffic lights, which signal 'stop', 'caution' or 'go'.

Let us examine in some more detail major parts of a control chart.

Major Parts of a Control Chart

A control chart generally includes the following four major parts, which are shown in Fig. 23.2.

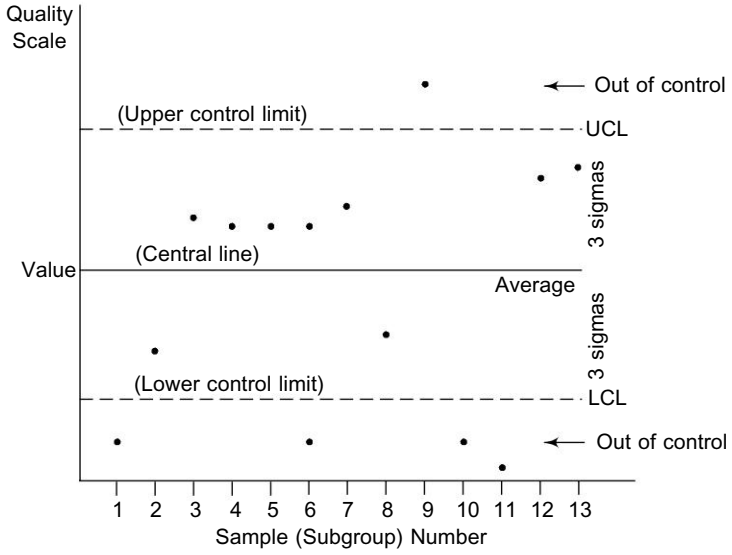


Fig. 23.2 **Major Parts of a Control Chart**

Quality Scale This is a vertical scale, which is marked as per the chosen quality characteristic (either in variables or attributes) of each sample.

Plotted Samples The control chart does not show the qualities of individual items of a sample. Instead, the quality of the entire sample represented by a single value (a statistic) is shown. The single value plotted on the chart is in the form of a dot (or sometimes a small circle or a cross).

Sample Numbers The samples, which are also referred to as subgroups in SQC, on a control chart are numbered individually and are shown on a horizontal line. The line is usually shown at the bottom of the chart. It may be noted that the utility of the control chart technique depends to a great extent on the proper grouping of items into samples. The grouping should be such that variation in quality among items within the same sample is small, while variation between one sample and another is large. Such a sample is regarded as 'rational subgroup'.

The Horizontal Lines The central line represents the average quality of the samples plotted on the chart. The line above the central line shows the upper control limit (UCL), which is commonly obtained by adding 3 sigmas (σ) to the average, that is, mean + 3 standard deviation. Similarly, the lower control limit (LCL) is given below the central line. It is obtained by subtracting 3 sigmas (σ) from the average, that is, mean –3 standard deviation. The upper and lower control limits are usually drawn as dotted lines.

Why 3-Sigma Limits?

We have just said that upper and lower control limits are set at 3σ limits. One may ask the reason for this approach. It may be noted that the 3σ limits were first proposed by Shewart for his control charts.

On the basis of probability consideration, if variable X is normally distributed, the probability that a random observation on the variable will lie between $\mu \pm 3\sigma$ (where μ is the mean and σ the standard deviation of X) is 0.997, which is extremely high. It may be recalled that in Chapter 7, Fig 7.1 shows that the area of the normal curve between $\mu \pm 3\sigma$ is 99.73 per cent.This means that the probability of a random observation going beyond these limits is nearly 0.003. This means that the variable quality characteristic is assumed to be normally distributed and that the probability of a sample point going outside 3σ limits when the process is in control is very small. If a sample point goes beyond this limit, it is highly likely that the normality assumption of the process is not applicable.

In order to set up a sound quality control mechanism, the concerned organisation must be keenly interested. It must take the following steps.

First, it must select the quality characteristics, which need to be kept under control. Besides, both their upper and lower limits within which variation can be tolerated, should be fixed up. *Second*, the production process must be analysed so that the possible causes of variation can be determined. *Finally*, it must lay down as to how the inspection data will be collected and recorded as also how they will be subdivided. Depending on the type of inspection data available, any one of the following types of control charts can be used.

1. Control charts for $\bar{x}$
2. Control chart for σ or R alone
3. Control chart for C
4. Control chart for p or p_n

23.3 $\bar{x}$-CHARTS: CONTROL CHARTS FOR PROCESS MEANS

In order to ascertain whether the process is in control or out of control, $\bar{x}$-charts are constructed. In regard to the process output, there is an assumption of normality where μ and σ are known, though in many situations this assumption may not hold good. We know from Chapter 11 that the sample means have a sampling distribution with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

The construction of $\bar{x}$-chart needs the values of μ and σ and also a sample size n. There are three lines in a $\bar{x}$ control chart, viz. the centre line indicating $\mu_{\bar{x}}$, the upper control limit (UCL), with value $\mu_{\bar{x}} + 3\sigma$ and the lower control limit (LCL), with value $\mu_{\bar{x}} - 3\sigma$. In addition to the control limits, there are warning limits, which are determined by 1.96 σ on either side of the centre line. Thus, the upper warning limit (UWL) = $\mu + (1.96)/\sqrt{n}$ and the lower warning limit (LWL) = $\mu - (1.96)/\sqrt{n}$. Figure 23.1 shows these two warning limits. However, the control charts do not normally show the warning limits.

Let us take an example to illustrate the procedure used in constructing $\bar{x}$ control charts.

Example 23.1 A company is engaged in the manufacture of battery cells in its plant. The process is said to be under control if the mean life of battery cells is 1,200 hrs with a standard deviation of 75 hrs. Considering these values to be the process average and process dispersion, you are required to determine the 3-sigma control limits for $\bar{x}$-chart for samples of size 16.

Solution Given are μ = 1,200 hrs, σ = 75 hrs and n = 16.

As the estimates of process average and process dispersion are based on a large sample, the desired control limits can be obtained by the following formula:

$$\mu \pm 3\ \sigma/\sqrt{n}$$

Substituting the values in the above formula,

$$\text{UCL} = \mu + 3(75/\sqrt{16})$$
$$= 1{,}200 + 56.25 = 1{,}256.25$$
$$\text{LCL} = \mu - 3(75/\sqrt{16})$$
$$= 1{,}200 - 56.25$$
$$= 1{,}143.75$$

$\bar{x}$-Chart When μ and σ are not Known

The preceding discussion has given us some basic ideas on $\bar{x}$-chart. The question is that when population mean and population standard deviation are not known to us, then how to construct $\bar{x}$-charts. In such cases, we use sample information to estimate unknown parameters. Let us take first the estimation of μ. This can be done by taking the mean of the sample means ($\bar{\bar{x}}$). This can be calculated by the following formula:

$$\bar{\bar{x}} = \Sigma x/n \times k = \Sigma \bar{x}/k$$

where, n = number of observations in each sample
k = number of samples taken

Another question is: how should we estimate σ? It may be recalled that in Chapter 13—'Estimation'—we used s, the sample standard deviation, to estimate σ. However, in respect of control charts, it has become customary to use $\bar{R}$ as an estimate of σ. $\bar{R}$ signifies the average of the sample ranges. It is a biased estimator of σ, and d is the correction factor. The values for d_2 are given in Appendix Table 11. Thus, the upper and lower control limits (UCL and LCL) for an $\bar{x}$-chart are computed with the following formulas:

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$
$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}}$$

In the above formula, d_2 stands for control chart factor as given in Appendix Table 11. These limits are often calculated as $\bar{\bar{x}} \pm A_2\,\bar{R}$ where $A_2 = 3/(d_2\sqrt{n})$. Appendix Table 11 also gives the values of A_2.

By using these formulas, we can now plot the three lines—CL (central line), UCL (upper control line) and LCL (lower control line). Let us take an example to show how these formulas can be used.

Example 23.2 Suppose we are given the following information:
$n = 20$, $\bar{\bar{x}} = 75$ and $\bar{R} = 15$. We are asked to find the CL, UCL and LCL for a $\bar{x}$ control chart.

Solution It is obvious that CL is the grand mean, that is, 75.

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$
$$= 75 + \frac{3(15)}{3.735 \times \sqrt{20}} \text{ (The value of } d_2 \text{ has been obtained from Appendix Table 11.)}$$
$$= 75 + \frac{45}{16.70}$$
$$= 77.69$$

$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= 75 - \frac{3(15)}{3.735 \times \sqrt{20}}$$

$$= 75 - \frac{45}{16.70}$$

$$= 72.31$$

Example 23.3 A company manufactures tyres. A quality control engineer is responsible to ensure that the tyres turned out are fit for use up to 40,000 km. He monitors the life of the output from the production process. From each of the 10 batches of 900 tyres, he has tested 5 tyres and recorded the following data, with $\bar{x}$ and $\bar{R}$ measured in thousands of km.

Batch	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
$\bar{x}$	40.2	43.1	42.4	39.8	43.1	41.5	40.7	39.2	38.9	41.9
$\bar{R}$	1.3	1.5	1.8	0.6	2.1	1.4	1.6	1.1	1.3	1.5

Construct an $\bar{x}$-chart using the above data. Do you think that the production process is in control? Explain.

Solution

$$\bar{\bar{x}} = \frac{\Sigma\bar{x}}{k} = \frac{410.8}{10} = 41.08$$

$$\bar{R} = \frac{\Sigma R}{k} = \frac{14.2}{10} = 1.42$$

$$\text{CL} = 41.08$$

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= 41.08 + \frac{3(1.42)}{2.326 \times \sqrt{5}}$$ (The value of d_2 has been obtained from Appendix Table 11.)

$$= 41.08 + \frac{4.26}{2.326 \times 2.24}$$

$$= 41.08 + 0.82 = 41.9$$

$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= 41.08 - \frac{3(1.42)}{2.326 \times \sqrt{5}}$$

$$= 41.08 - \frac{4.26}{2.326 \times 2.24}$$

$$= 41.08 - 0.82 = 40.26$$

The production process is in control in respect of only 3 batches as is indicated in Fig. 23.3. The production process is in respect of batches 1, 4, 8 and 9 has gone out of control so also batches 2, 3 and 5.

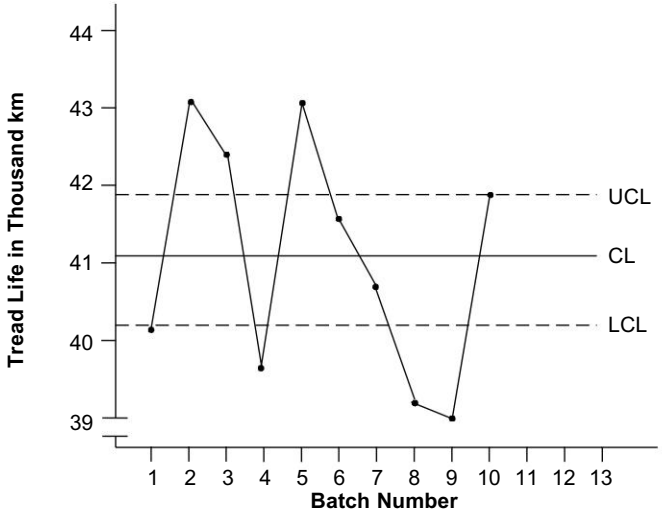


Fig. 23.3 $\bar{x}$**-Chart for the Data Given in Example 23.3**

23.4 *R*-CHARTS: CONTROL CHARTS FOR PROCESS VARIABILITY

We have seen that $\bar{x}$-charts set the control limits on the extent of a variability that can be tolerated in our sample means. However, the problem of quality is addressed to individual observations. In Chapter 11 on 'Sampling and Sampling Distributions', we had seen that variability in sample means was less than that in individual observations. For this purpose, we use the following equation:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In order to monitor the variability in the individual observations, the above equation cannot be used. Instead we use another control chart that is known as an *R*-chart. In this chart, the value of the sample range for each of the samples is plotted. The central line for *R*-charts is placed at $\bar{R}$. Now, we have to decide the control limits for which we need some additional information regarding the sampling distribution of *R*, in particular its standard deviation σ_R. For this purpose, the following formula is used.

$$\sigma_R = d_3\sigma$$

where

σ = *population standard deviation*

d_3 = *another factor depending on n*

The values of d_3 are given in Appendix Table 11. Now, σ can be substituted by $\bar{R}/d_2$ as was done in an earlier equation so that the control limits for *R*-charts can be computed.

Control Limits for an *R*-Chart

$$\text{UCL} = \bar{R} + \frac{3d_3\,\bar{R}}{d_2} = \bar{R}\left(1 - \frac{3d_3}{d_2}\right)$$

$$\text{LCL} = \bar{R} - \frac{3d_3\,\bar{R}}{d_2} = \bar{R}\left(1 - \frac{3d_3}{d_2}\right)$$

It may be noted that these limits are often calculated as:

$$\text{UCL} = \bar{R}\,D_4, \text{ where } D_4 = 1 + \frac{3d_3}{d_2}$$

$$\text{LCL} = \bar{R}\;D_3, \text{ where } D_3 = 1 - \frac{3d_3}{d_2}$$

The values of D_3 and D_4 can also be obtained from Appendix Table 11.

Example 23.4 We have to determine the UCL and the LCL by applying the above formulae to the data given in Example 23.3.

Solution The UCL and the LCL are calculated as follows:

$$\text{UCL} = \bar{R}\left(1 + \frac{3d_3}{d_2}\right)$$

$$= 1.42\left(1 + \frac{3(0.864)}{2.326}\right)$$

$$= 1.42\ (\ 1 + 1.11) = 2.996 \text{ or } 3 \text{ approx.}$$

$$\text{LCL} = \bar{R}\left(1 - \frac{3d_3}{d_2}\right)$$

$$= 1.42\left(1 - \frac{3(0.864)}{2.326}\right)$$

$$= 1.42 \times -0.11 = -0.156 \text{ (to be taken as zero)}$$

Some explanation is needed for the zero value of LCL. A sample range is always a non-negative number (because it is the difference between the largest and smallest observations in the sample). However, when $n \leq 6$, the LCL computed by the above equation will be negative. Although in this case n is 10, yet the calculation shows a negative value. As such, we set the value of LCL at zero.

A major limitation of R-chart arises from the characteristic of range itself. As we know that the range considers only the highest and the lowest values in a distribution, it may ignore the nature of variation in the remaining observations. Further, it is influenced by extreme values, which may significantly differ from one sample to the other. In view of these limitations, R-chart is only a convenient device for examining variability of the process.

23.5 CONTROL CHART FOR C (NUMBER OF DEFECTS PER UNIT)

So far we have considered the control charts for attributes in those cases wherein a random sample of definite size is selected and examined in some way. However, there are certain situations where the number of events, defects, errors can be counted, but there is no information about the number of events, defects or errors that are not present. So far we have considered defectives where each item is classified in one of the two categories—defective or non-defective. In such cases, we know the number of defects, say, number of holes in a fabric but we do not know the number of non-defects present. In such cases, the Poisson distribution is to be applied.

The central lines of the control chart for C is $\bar{C}$ *and the 3-sigma control limits are*

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$

This formula is based on a normal curve approximation to the Poisson distribution. The use of the *C*-chart is appropriate if the occasions 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 the *p*-chart (discussed just after the *C*-chart) provides a better choice.

The equation for the Poisson distribution is:

$$P(x) = e^{-c}\,(\bar{C}^2/x\,!)$$

where e = exponential constant, 2.7183 and $\bar{C}$ *= average number of defects per unit being produced by the process.*

Suppose we have to find the probability of having four bubbles in a windscreen from a process which is producing them on an average of one bubble present. The probability will be: $P(4) = e^{-1}\,(1^4/4!) = 0.01533$. As with the *np*-chart it is not necessary to calculate probability in this way to determine control limits for the *C*-chart. Once again the UCL is set at 3 standard deviation above the average number of events.

Example 23.5 Fifteen pieces of cloth from different rolls contained respectively 1, 5, 3, 2, 7, 6, 3, 2, 6, 5, 4, 3, 4, 6, and 3 imperfections. Draw a control chart using these data and state whether the process is in a state of statistical control.

Solution

$$\bar{C} = (1 + 5 + 3 + 2 + 7 + 6 + 3 + 2 + 6 + 5 + 4 + 3 + 4 + 6 + 3)/15 = 60/15 = 4$$

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$
$$= 4 + 3\sqrt{4} \quad = 4 + 6 = 10$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$
$$= 4 - 3\sqrt{4} \quad = 4 - 6 = -2$$

Since the number of defectives cannot be negative, the lower control limit will be taken as zero. Figure 23.4 shows both the control limits. The chart clearly shows that all the imperfections in cloth are

Fig. 23.4 **Control Chart for *C***

within the control limits, that is, no point lies outside the control limits. This suggests that the process is in a state of statistical control.

Let us take another example.

Example 23.6 The following table gives the number of defects observed in 8 woollen carpets passed as satisfactory. Construct the control chart for the number of defects.

Number of carpets	1	2	3	4	5	6	7	8
Number of defects	3	5	8	4	3	5	7	5

Solution $\overline{C} = (3 + 5 + 8 + 4 + 3 + 5 + 7 + 5)/8 = 40/8 = 5$

$$\text{UCL} = \overline{C} + 3\sqrt{\overline{C}}$$
$$= 5 + 3\sqrt{5} = 5 + 6.708 = 11.708$$

$$\text{LCL} = \overline{C} - 3\sqrt{\overline{C}}$$
$$= 5 - 3\sqrt{5} = 5 - 6.708 = -1.708$$

Here too, we find that the lower control limit is negative. Since number of defects cannot be negative, LCL will be taken as zero. Accordingly, the UCL is 11.7 and the LCL is zero. Figure 23.5 shows that the process is under control.

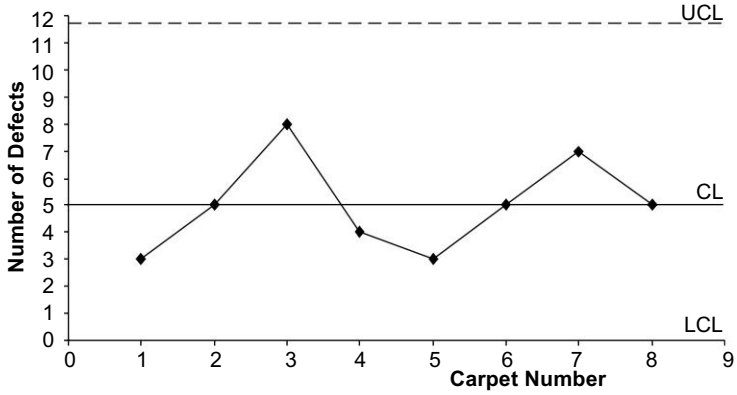


Fig. 23.5 **Control Chart for Data Given in Example 23.6**

23.6 *p*-CHARTS: CONTROL CHARTS FOR ATTRIBUTES

The control chart for attributes is known as the *p*-chart. Such a chart is used to control the proportion or percentage of defectives per sample. It may be noted that there is an assumption that the items are produced by Bernoulli process, which implies the following three assumptions: (i) There are only two outcomes—acceptable or defective. (ii) The outcomes occur randomly. (iii) There is no change in the probability of either outcome for each trial.

As we have seen earlier that the *C*-chart is concerned with the number of defectives, it can be easily converted into proportion by dividing the number of defectives by the sample size. Thus, we can use the *p*-chart in place of the *C*-chart. In order to draw the *p*-chart, we have to follow the following procedure:

1. Calculate the average fraction defective ($\bar{p}$) by dividing the number of defective units by the total number of units inspected.
2. The value of $\bar{p}$ is now used to draw a horizontal line.
3. The upper and lower control limits are to be obtained by using the following formulas:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$$
$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$$
where $\bar{q} = (1 - \bar{p})$

Any sample point falling outside the UCL and the LCL indicates that the process is not in control. It is preferable to set up the chart to express 'per cent defective' to 'fraction defective'.

Example 23.7 The following figures give the number of defects in 10 samples, each containing 200 items: 40, 44, 22, 34, 24, 32, 28, 32, 34 and 30.

Calculate the values for central line and the upper and lower control limits for *p*-chart. Draw the *p*-chart and comment if the process can be regarded in control.

Solution

Table 23.1 Worksheet for Calculating the Values for *p* -Chart

Sample No.	*No. of Defectives*	*Fraction Defectives*
1	40	0.20
2	44	0.22
3	22	0.11
4	34	0.17
5	24	0.12
6	32	0.16
7	28	0.14
8	32	0.16
9	34	0.17
10	30	0.15
Total	**320**	

$$\bar{p} = \frac{\text{No. of units defective}}{\text{Total no. of units inspected}} = \frac{320}{2{,}000} = 0.16$$

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.16 + 3\sqrt{\frac{0.16(1-0.16)}{200}}$$
$$= 0.16 + 0.07776 = 0.2378$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.16 - 3\sqrt{\frac{0.16(1-0.16)}{200}}$$
$$= 0.16 - 0.07776 = 0.0822$$

It will be seen from Fig. 23.6 that all the units fall within the upper and lower control limits. On the basis of this chart, we can say that the process is well under control. It may be noted that we have plotted the percentage defective instead of fraction defective in the above chart.

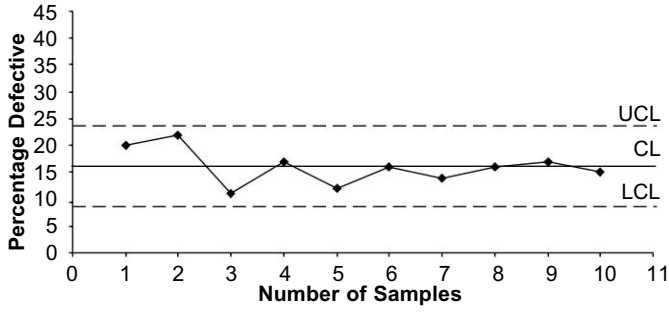


Fig. 23.6 ***p*-Chart for Data Given in Example 23.7**

Example 23.8 On inspection of a large sample of a product, it was found that the proportion of the defective product in the sample output was 0.2. Compute the upper and lower control limits for a lot size $n = 400$.

Solution Since $\bar{p}$ is 0.2, $\bar{q} = 1 - \bar{p} = 0.8$. The control limits are obtained as:

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$$

$$\text{UCL} = 0.2 + 3\sqrt{\frac{0.2 \times 0.8}{400}}$$

$$= 0.2 + 0.06 = 0.26$$

$$\text{LCL} = 0.2 - 3\sqrt{\frac{0.2 \times 0.8}{400}}$$

$$= 0.2 - 0.06 = 0.14$$

Hence, the upper and lower control limits are 0.26 and 0.14, respectively.

23.7 ADDITIONAL EXAMPLES

Example 23.9 An inspection of large sample of a product revealed a fraction defective of 0.35. Taking the sample size $n = 100$, compute the upper and lower control limits.

Solution With $\bar{p} = 0.35$, $\bar{q} = (1 - \bar{p}) = 0.65$

The control limits are obtained as follows:

$$\bar{p} \pm 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$$

Substituting the values,

$$L.C.L.\ \bar{p} = \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$$

$$= 0.35 - 3\sqrt{\frac{(0.35)(0.65)}{100}}$$

$$= 0.35 - 0.143$$

$$= 0.207$$

$$U.C.L.\ \bar{p} = \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$$

$$= 0.35 + 0.143$$

$$= 0.493$$

Example 23.10 Based on 15 subgroups each of size 200 taken at intervals of 45 minutes from a manufacturing process, the average fraction-defective was found to be 0.068. Calculate the values of central line and the control limits for a *p*-chart.

Solution

$$\bar{p} = 0.068$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= (0.068) + 3\sqrt{\frac{0.068(1-0.068)}{200}}$$

$$= 0.068 + 3\sqrt{\frac{0.063376}{200}}$$

$$= 0.068 + 3\sqrt{0.00031688}$$

$$= 0.068 + 3 \times 0.0178$$

$$= 0.068 + 0.0534$$

$$= 0.1214$$

$$LCL = 0.068 - 0.0534$$

$$= 0.0146$$

We can write 0.068 ± 0.0534

Example 23.11 Ten samples, each of size 5, are drawn at regular intervals from a manufacturing process. The sample means ($\bar{x}$) and their ranges (R) are given below:

Sample No.	1	2	3	4	5	6	7	8	9	10
$\bar{x}$	49	45	48	53	39	47	46	39	51	45
R	7	5	7	9	5	8	8	6	7	6

What are the upper and lower control limits in respect of both $\bar{x}$-chart and R-chart? Do you think that the process is under control?

Solution

$$\bar{\bar{x}} = \frac{\Sigma \bar{x}}{k} = \frac{462}{10} = 46.2 \quad \bar{R} = \frac{\Sigma R}{k} = \frac{68}{10} = 6.8$$

$$C.L = 46.2$$

$$UCL = \bar{\bar{x}} + 3\frac{\bar{R}}{d_2\sqrt{n}}$$ [Value of d_2 taken from Appendix Table 11]

$$= 46.2 + \frac{3(6.8)}{2.325\sqrt{n}}$$

$$= 46.2 + 3.92$$

$$= 50.12$$

$$LCL = 46.2 - 3.92$$

$$= 42.28$$

For R-Chart

$$C.L. = 6.8$$

$$UCL = \bar{R} + \frac{3d_3\bar{R}}{d_2}$$ [Value of d_3 taken from Appendix Table 11]

$$= 6.8 + \frac{3 \times 0.864 \times 6.8}{2.326}$$

$$= 6.8 + 7.58 = 14.38$$

$$LCL = 6.8 - 7.58 = 0$$ As it cannot be negative.

In case of $\bar{x}$-chart , three points do not fall within limits, indicating that the process is not in control.

Example 23.12 The following table gives the number of defects observed in 7 carpets passed as satisfactory. Construct the control chart for the number of defects.

No. of carpet	1	2	3	4	5	6	7
No. of defects	2	3	5	4	5	4	5

Solution

$$\bar{C} = (2 + 3 + 5 + 4 + 5 + 4 + 5)/7$$

$$= 28/7 = 4$$

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$

$$= 4 + 3\sqrt{4}$$

$$= 4 + (3 \times 2)$$

$$= 10$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$
$$= 4 - (3 \times 2)$$
$$= 4 - 6 = -2 \quad \text{It is to be taken as zero.}$$

As the number of defectives cannot be negative, LCL will be taken as zero.

Hence, $UCL = 10$ and LCL = Zero.

Example 23.13 The following figures give the number of defectives in 20 samples, each containing 2,000 items:

425	430	216	341	225	322	280	306	337	305
356	402	216	264	126	409	193	326	280	389

Calculate the values for central line and the control limits for *p*-chart (fraction defective chart). Comment if the process can be regarded as in control or not.

Solution As the question suggests, we have to use a *p*-chart.

The total number of defectives is:

$$425 + 430 + 216 + \ldots + 389 = 6148$$

Since there are 20 samples and each sample contains 2000 items. This means the total number is $2000 \times 20 = 40000$.

Hence, $$\bar{p} = \frac{6148}{40000} = 0.1537$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= 0.1537 - 3\sqrt{\frac{0.1537 \times 0.8463}{2000}}$$
$$= 0.1537 - (3 \times 0.0081)$$
$$= 0.1294$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$
$$= 0.1537 + (3 \times 0.0081)$$
$$= 0.178$$

$$0.1537 \pm 0.0243$$

Example 23.14 A machine is said to deliver the packets of a given weight. Ten samples of size five were examined and the following results were obtained:

Sample	1	2	3	4	5	6	7	8	9	10
Mean	15	17	15	18	17	14	18	15	17	16
Range	7	7	4	9	8	7	12	4	11	5

Calculate the values for the central line and the control limits for the mean chart and range chart. Comment on the state of control.

Solution We have to first calculate the mean of sample means $\bar{\bar{x}}$

$$\bar{\bar{x}} = \frac{\Sigma \bar{x}}{k} = \frac{162}{10} = 16.2$$

Similarly, mean of the range $\bar{R}$ is to be calculated

$$\bar{R} = \frac{\Sigma R}{k} = \frac{74}{10} = 7.4$$

Calculations for the $\bar{x}$-chart

$$CL = \bar{\bar{x}} = 16.2$$

$$UCL = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= 16.2 + \frac{(3)(7.4)}{2.326\sqrt{5}} = 16.2 + 4.27 = 20.47$$

$$LCL = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = 16.2 - 4.27 = 11.93$$

Calculations for the R-chart

$$CL = \bar{R} = 7.4$$

$$UCL = \bar{R} + \frac{3d_3\bar{R}}{d_2}$$

$$= 7.4 + \frac{3 \times 0.864 \times 7.4}{2.326}$$

$$= 7.4 + 8.25 = 15.65$$

$$LCL = \bar{R} - \frac{3d_3\bar{R}}{d_2}$$

$= 7.4 - 8.25 = 0$ (Being negative, its value is zero)

Comment

It may be noted that none of the 10 values lies beyond the control limits. As such, the process is in control both in respect of average and variability in the quality.

23.8 BENEFITS OF STATISTICAL PROCESS CONTROL

There are several benefits of SPC approach and these include:

1. SPC can be applied to any type of problem selected and process originally tackled will result into improvement.
2. This approach eliminates the 'emotion' factor and the decisions are based on facts rather than on opinions.
3. As the workers are directly involved in the improvement process, their 'quality awareness' increases.

4. The knowledge and experience potential of those involved in the process is released in a systematic way through the investigative approach. They increasingly realise that their role in problem solving is collecting and communicating the relevant facts on which decisions are made.
5. Managers and supervisors solve problems methodically instead of in haphazard manner. Thus, the approach to the problem becomes unified in place of an individualistic approach earlier.
6. In case of any inquiry from the government or any other appropriate authority, the quality can be defended on the basis of statistical process control.
7. Since the firm strictly adheres to the SPC, the users of the product may rely on it and may not resort to check the quality themselves.

23.9 LIMITATIONS OF STATISTICAL PROCESS CONTROL

Despite the above mentioned advantages of the SPC, it may be noted that it is unable to solve all the problems arising in quality improvement. There are several highly complex problems where SPC may not be in a position to contribute much towards reduction of variability. This apart, at times, managers use SPC mechanically and construct control charts without going into the depth of the problem. As a result, statistical methods have been criticised at times. It has been argued that continuous improvement in quality can be attained by studying all parts of an organisation and not merely one part, viz. production process. This consideration has led to the concept of Total Quality Management, which is discussed below.

23.10 TOTAL QUALITY MANAGEMENT*

The concept of total quality management (TQM) is basically very simple. Each part of an organisation has customers, the need to identify what the customer requirements are and then attempt to meet them. This forms the core of the total quality approach. This needs three conditions to be fulfilled: a good management system, tools such as statistical process control (SPC), and teamwork. In many ways, these are complimentary and they share the same requirement for an uncompromising commitment to quality.

Total quality management involves consideration of processes in all the major area: marketing, design, procurement, operations, distribution, and so forth. Each of these areas requires considerable expansion and thought but if attention is given to them in accordance with the concept of TQM then very little will be left to chance. The point to emphasise is that there must be a well-operated management system that alone can be a sound foundation for the successful application of SPC techniques and teamwork. In contrast, a poor management system cannot undertake this task.

TQM, like SPC, requires that the process should be improved continually by reducing its variability. This can be attained by studying all aspects of the process using the basic question: could we do the job more consistently and on target (i.e. better)? An attempt to answer this question will drive the search for improvements. Once improvements have been identified, management can implement the desired changes to achieve continuous improvement.

We discuss below two techniques that are very useful in TQM. These are *Fishbone* and *Pareto* diagrams.

---

* Based on Oakland, John S, *Statistical Process Control* 4th ed. (Oxford: Butterworth-Heinemann, 1999), pp 14-17.

Fishbone Diagram

Cause and effect analysis is a technique that comprises usage of cause-and-effect diagrams and brainstorming. The cause-and-effect diagram is also known as the *Ishikawa diagram* or the *Fishbone diagram* (after its appearance). It shows the effect at the head of the central 'spine' with the causes at both the ends of the 'ribs' that branch from it. The basic form (Fig. 23.7) is shown below:

The main causes are listed first and then reduced to their sub-causes, and if necessary, sub-sub-causes. This process continues until all the conceivable causes have been covered. This is followed by a critical analysis of causes and their probable contribution to the effect. The factors selected as most likely causes of the effect are then subjected to experimentation. This helps to determine how far the causes selected are valid in a given situation. Such an experimentation is repeated until the true causes are identified.

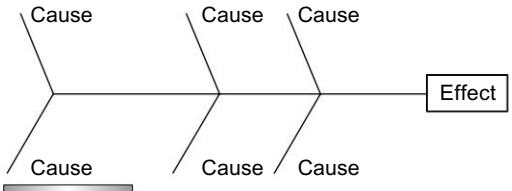


Fig. 23.7 **Basic Form of Cause-and-Effect (Fishbone) Diagram**

The question is: how to construct the fishbone diagram? A brainstorming session is an essential technique for a fishbone diagram as it brings out ideas on causes, from a group of people entrusted with this job.

Steps for Constructing Fishbone Diagram There are five steps involved in the construction and analysis of such a diagram. These are briefly explained below.

Identify the Effect or Problem This is very important that the effect or the problem is formulated in clear and concise terms. This will help avoid the situation where the causes are identified and eliminated.

Establish Goals In order to carry out problem-solving activity properly, it is necessary to set up realistic and meaningful goals. A goal should be set up in some terms of measurement related to the problem. It should also give a time limit to achieve the goal.

Construct the Diagram Framework The diagram framework on which causes are listed can considerably facilitate the thinking process. In this connection, Oakland has found that the five Ps of production management are very useful in the construction of cause-and-effect diagrams. The five Ps are:

- **Production,** including service, materials and any intermediates
- **Processes** or methods of transformation
- **Plant,** that is, the building and equipment
- **Programmes** or timetables for operations
- **People,** operators, staff and managers.

These are shown on the main ribs of the fishbone diagram with the effect at the end of the spine of the diagram. This is shown in Fig. 23.8.

A point worth noting is that the grouping of the subgroups under the five Ps can be very helpful in a subsequent analysis of the diagram.

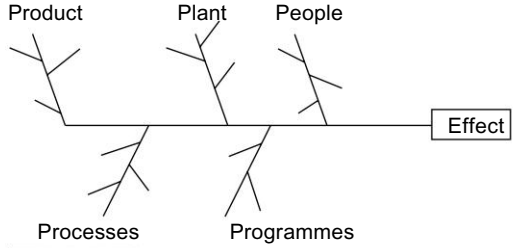


Fig. 23.8 **Fishbone Diagram with Five Ps**

Record the Causes The causes will emanate from a brainstorming session as was mentioned earlier. To a great extent the success of such a session depends on the group leader's way of conducting it. The causes are written on the appropriate branch of the diagram. In any exercise of this type, causes are just listed and no criticism of any cause is allowed at this stage.

Incubate and Analyse the Diagram Finally, the last step involves the analysis of the diagram. Sufficient time should be given to the concerned members to offer their suggestions. For this purpose, it is necessary that the diagram must remain on display for at least a few days. A major advantage of such an 'incubation' period is that as a result of some lapse of time, members are unable to recall who made what point. As such they are free to critically analyse the diagram to arrive at the most genuine causes leading to that particular problem.

Pareto Diagram

Another device used for quality improvement is the *Pareto Diagram*. A detailed fishbone diagram may cause considerable disappointment to the people involved as they may think that it is almost impossible to take care of so many defects and errors. To overcome such a gloomy outlook, Joseph Juran made a major contribution when he advised that TQM companies should distinguish between the *vital few* and the *trivial many*.

A *Pareto Chart* is a bar diagram that displays groups of error causes arranged in order of their frequencies of occurrence. Juran observed that in extremely complex systems, 80 per cent of defects in errors could be attributed to only 20 per cent of the causes. Figure 23.9 gives a hypothetical example of Pareto chart.

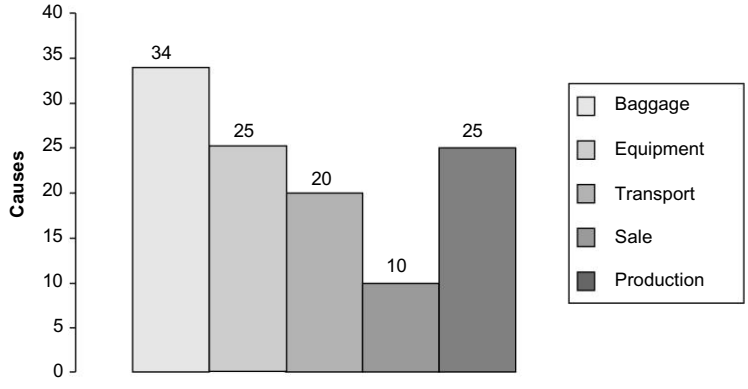


Fig. 23.9 **A Hypothetical Example of Pareto Chart**

Once the causes of defects and errors have been identified, the management can address itself to improve the quality of the systems' goods and services. Many a times this may necessitate the reallocation of the sources within the system. In case such an exercise is done systematically, TQM may be completely successful. This may result into the leading cause of defects or errors to drop to zero on the Pareto's chart. However, in such a situation, management will now be concerned with another part of the system to which they had not focused their attention earlier.

When a company adopts the TQM technique, it implies that a goal that inputs in each stage of an operation should not have any defect. This is because the operations at the preceding stage are sub-

jected to quality control. As the manufacturers have to obtain raw materials and components from suppliers, it is in their interest to ensure that the inputs are defect-free. It, therefore, becomes necessary for them to test inputs frequently. Since complete inspection of an entire batch is not possible on account of constraints on time and cost, a sample of the batch is inspected on the basis of which decision to either accept or reject the entire batch is made. This is an important area in quality control and is known as *acceptance sampling*. This is discussed in some detail in the section that follows.

23.11 ACCEPTANCE SAMPLING

Acceptance sampling involves sampling inspection by a purchaser who has to decide whether to accept a shipment of product. Thus, the objective of acceptance sampling is either to accept or to reject the product. It does not attempt to control the quality during the manufacturing process. As will be evident, this is altogether a different approach from what has been followed in control charts discussed earlier.

A major *advantage* of acceptance sampling is that it can motivate suppliers to improve the quality of their items. Suppose a company receives a batch of components from its supplier and finds that 10 per cent of the supply is defective. Although 90 per cent of it is free from defects, but the company may decide to reject the entire lot to ensure its qualitative output. This decision of the company would result into heavy loss to the supplier. He has to suffer even though a small proportion of his equipment was defective. In order to avoid such an eventuality, the supplier would be very particular from the very beginning to ensure his supplies are free from defects. In contrast, if the company rejects only 10 per cent of defective equipment, it amounts to imposing a high cost on itself and a low cost on the supplier.

Single-sampling Plan

When the decision on whether to accept or reject a lot is based on only one sample, the acceptance plan is said to be on a single-sampling plan. There are three things that need to be specified in a single-sample plan. These are: **(a)** number of items n in the lot from which the sample is chosen, **(b)** number of articles n drawn by random sample from the given lot, and **(c)** the acceptance number C, which specifies the maximum number of defective articles allowable in the sample. In case the number of defective articles crosses this limit in the sample drawn, the entire lot is to be rejected.

Double-sampling Plan

Double-sampling plan is obviously more complicated than the single-sampling plan. In this case, a lot is immediately accepted or rejected depending on the condition of the first sample. At times, the management finds that the first sample is neither good enough nor bad enough so as to take a decision one way or the other. In such a situation, it defers its decision until a second sample is drawn. On the basis of the evidence from both the first and the second samples, a decision is finally taken whether to accept or reject the lot.

A double-sampling plan depends on five specified numbers (besides N): n_1, c_1, n_2, $n_1 + n_2$ and c_2 ($>c_1$), which are used as follows:

First, a sample of size n_1 is taken. Let b_1 denote the number of defective pieces in the first sample and c_1 denote the number of defective pieces acceptable in the lot, then

(a) *accept the lot if $b_1 \leq c_1$*

(b) *reject the lot if $b_1 > c_1$*

(c) *an additional n_2 units are sampled if $c_1 < b_1 \leq c_2$*

Let b_2 be the total number of defective pieces in the combined sample of $n_1 + n_2$ units:
(d) *accept the lot if $b_2 \leq c_2$*
(e) *reject the lot if $b_2 > c_2$*

As mentioned earlier, double-sampling plans are more complicated than the single-sampling plan. But, as they are more powerful, they are more frequently used in quality control problems.

Multiple or Sequential Sampling Plan

We have seen earlier that when a single-sampling plan is unable to give us a clear decision, we take recourse to the double-sampling plan. It may just be possible that even a double-sampling plan may not give us a clear decision. In such a case, we may go on to have another sample before we reach a definite decision. Thus, three or more sampling plans can be used. This is known as multiple or sequential sampling. Since such plans are extremely complicated, they are seldom used in practice.

Selection of a Sampling Plan All practical sampling plans have an operating characteristic curve—OC curve for short. The following points are relevant in regard to the OC curve:

1. There is some probability that good lots will be rejected.
2. There is some probability that bad lots will be accepted.
3. Based on the theory of probability, it is possible to calculate these risks, which will depend on a number of factors such as the number of samples inspected, the acceptance number and the per cent defective in the lots submitted for sample inspection. Having decided the amount of risk, which can be tolerated, a sampling plan meeting these requirements can be devised.
4. The larger the sample size used in sample inspection, the nearer the OC curve will approach the ideal. But there is a limit to this as the cost of inspecting a larger number of parts will far exceed the benefits to be derived.

In order to formulate a sampling plan for a lot of N items, one has to decide on the sample size (n) to be used and the acceptance number (c). The values of n and c are determined keeping in mind both types of risk—the consumer's risk, which means accepting a lot of defective quality and the producer's risk, which means rejecting a lot of satisfactory quality. In hypothesis testing, the producer's risk corresponds to Type I error while the consumer's risk corresponds to Type II error.

Construction of an OC Curve

From the preceding discussion, it should be evident that the relationships between sample size (n), acceptance number (c), and the two types of risk—the producer's risk and the consumer's risk are very complicated. In this connection, sampling inspection tables are available, which considerably facilitate quality engineers to select an appropriate acceptance sample plan.

The OC curve of an acceptance sampling plan shows how far a sampling plan is capable to differentiate between good and bad lots. For any given fraction of defective p in a lot under consideration, the OC curve brings out the probability p_a, that is, the lot will be accepted as per the given sampling plan.

In drawing the OC curve, the two terms AQL(Acceptance Quality Level) and LTPD (Lot Tolerance Percentage Defective) are important. It may be noted that both the producer and the consumer know that there is always some risk that lots would pass outside the range of AQL and LTPD. The chance that lots would be rejected while actually having lower percentage of defectives than the AQL is known as the producer's risk and is represented by α. As against this, the chance of a lot being accepted with

a higher percentage of defectives than indicated by LTPD is known as the consumer's risk and is represented by β. The actual levels of these two types of risk are decided by the consumer and the producer by negotiation. In other words, such a negotiation would mean that there is an explicit agreement between the consumer and the producer in respect of the following:

(a) the acceptable quality lot, AQL
(b) the lot tolerance per cent defective, LTPD
(c) the producer's risk, α
(d) the consumer's risk, β

This is illustrated in Fig. 23.10, which shows an OC curve along with producer's risk (α) and consumer's risk (β). The figure also shows AQL and LTPD. The points on the horizontal scale represent possible lot of process quality levels while the vertical scale shows the probability that a lot of specified quality level will be accepted. At times an additional point, known as split risk point, is shown as given in Fig. 23.10. This point shows the indifference quality level, which corresponds to 0.5 probability of acceptance. There is an implicit assumption that the remaining points of the curve would be satisfactory.

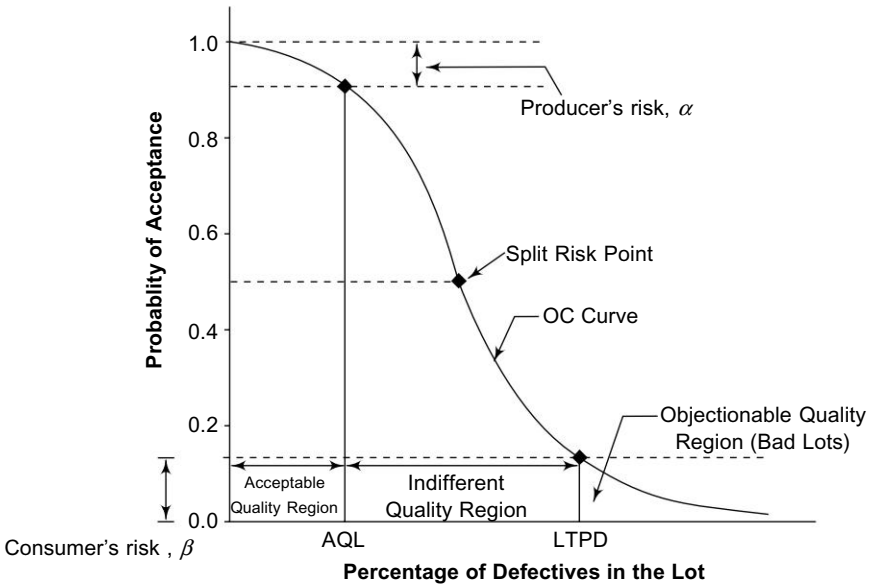


Fig. 23.10 OC Curve Along with Producer's and Consumer's Risk

Source Vohra, N.D.: *Quantitative Techniques in Management*, Tata McGraw-Hill Publishing Company Limited, New Delhi, 1990, p. 715. Reprinted by permission of the publishers.

The steepness of the OC curve will vary on account of variation in sample size. The larger the sample size, the steeper will be the OC curve and vice versa. Also, as the steepness of the OC curve increases, the zone between the acceptable qualities and the rejectable qualities will shrink. Another point worth noting relates to the location of the OC curve. Its location depends on the acceptance number, that is, the maximum number of defective items allowable for acceptance. If the acceptance number is reduced, the curve is shifted to the left. On the other hand, if the acceptable number is increased, the curve is shifted to the right.

Conclusion

This chapter has discussed various statistical techniques that are applied in quality control. In conclusion, we may reiterate that organisations must focus on the continuous improvement in quality of product and processes so that they may enhance their competitiveness both in the home and the overseas markets. For this, there must be real commitment on quality control from senior management.

GLOSSARY

Acceptance number	The maximum number of defective pieces in a lot, that can be accepted.
Acceptance sampling	The procedure used to decide whether to accept or reject a batch of input materials on the basis of their quality.
Action zones	The zones outside the action limits/lines on a control chart where the result is a clear indication of the need for action.
Attributes	Qualitative variables with only two categories.
C-Chart	A control chart used for attributes when the sample is constant and only the number of non-conformances is known.
Cause-and-effect diagram	A graphic display that illustrates the relationship between an effect and its contributory causes. It is also known as Fishbone Diagram.
Centre Line (CL)	A line on a control chart at the value of the process mean.
Control chart	A graphical method of recording results in order to readily distinguish between random and assignable causes of variation.
Control limits (lines)	Limits or lines set on control charts, which separate the zones of stability (no action required), warning (possible problems and the need to seek additional information) and action.
Ishikawa Diagram	Another name for cause-and-effect (or fishbone) diagram.
LCL	Lower Control Limit or line on a quality control chart.
Operating Characteristic curve (OC curve)	A graph showing how far a sampling plan is capable to differentiate between good and bad lots.
Outliers	Observations that do not come within the control limits on a control chart.
Pareto Chart	A graphical device used to show groups of error causes along with their frequencies of occurrence.
p-Chart	A control chart used for attributes showing the proportion of non-conforming items in a sample. *p* is the proportion of non-conforming items.
Process control	The management of a process by observation, analysis, interpretation and action designed to limit variation.
Producer's risk	The chance of rejecting a 'good' lot by the firm.

Qualitative variables	Those variables having categorical values rather than numerical values.
Quality	The totality of features and characteristics of a product or service so as to satisfy the customer's needs.
Quantitative variables	Variables having numerical values that are obtained from measuring or counting.
R-charts	Charts used in quality control for monitoring process variability. These are based on range.
Shewhart Charts	The control charts for attributes and variables, first proposed by Shewhart. These include mean and range, *np*, *p*, *c* and *u* charts.
SQC	Statistical Quality Control—similar to SPC but with an emphasis on product quality and less emphasis on process control.
Stable zone	The central zone between the warning limits on a control chart and within which most of the results are expected to fall.
Statsitical Process Control (SPC)	The use of statistically based techniques for the control of a process for transforming inputs into outputs.
Total Quality Management (TQM)	A set of approaches that involves consideration of processes in all the major areas so that the management can match its products or services to customer's expectations.
UCL	Upper Control Limit or line in a quality control chart.
Warning zone	The zone on a control chart between the warning and action limits and within which a result suggests the possibility of a change to the process.
$\bar{x}$ *-Chart*	A chart used in quality control for monitoring process means.

LIST OF FORMULAE

1. Computation of grand mean:

$$\bar{\bar{x}} = \frac{\Sigma x}{n \times k} = \frac{\Sigma \bar{x}}{k}$$

where $\bar{\bar{x}}$ = grand mean
Σx = sum of all observations
n = number of observations in each sample
k = number of samples of the same size

2. Computation of upper and lower control limits for $\bar{x}$ -chart

(a) $\text{UCL} = \bar{\bar{x}} + \dfrac{3\bar{R}}{d_2\sqrt{n}}$

(b) $\text{LCL} = \bar{\bar{x}} - \dfrac{3\bar{R}}{d_2\sqrt{n}}$

where $\bar{\bar{x}}$ = grand mean
$\bar{R}$ = average of the sample ranges (i.e. $\Sigma R/k$)
d_2 = control chart factor from Appendix Table 11.
n = number of observations
UCL = upper control limit
LCL = lower control limit

3. $\sigma_R = d_3\sigma$

where σ_R = standard deviation of the sampling distribution of R
σ = population standard deviation
d_3 = a control chart factor given in Appendix Table 11.

4. Computation of upper and lower control limits for an R-chart

(a) $\text{UCL} = \bar{R} + \dfrac{3d_3\bar{R}}{d_2} = \bar{R}\left(1 + \dfrac{3d_3}{d_2}\right)$

(b) $\text{LCL} = \bar{R} - \dfrac{3d_3\bar{R}}{d_2} = \bar{R}\left(1 - \dfrac{3d_3}{d_2}\right)$

5. Alternative formulae for upper and lower control limits for an R-chart. Values of D_3 and D_4 are also given in Appendix Table 11. As the range cannot be negative, D_3 and LCL are to be taken as zero when $n \leq 6$.

(a) UCL = $\bar{R}\,D_4$, where $D_4 = 1 + 3d_3/d_2$

(b) LCL = $\bar{R}\,D_3$, where $D_3 = 1 - 3d_3/d_2$

6. (a) UCL for C-chart: $\bar{\text{C}} + 3\sqrt{\bar{\text{C}}}$

(b) LCL for C-chart: $\bar{\text{C}} - 3\sqrt{\bar{\text{C}}}$

A known or targeted value of p, if any, should be used in this and 7(a) and 7(b) formulae. In the absence of such value, the overall sample fraction given in formula 8 should be used.

7. (a) UCL for p-chart = $\bar{p} + 3\sqrt{\dfrac{(\bar{p}(1-\bar{p})}{n}} = \bar{p} + 3\sqrt{\dfrac{\bar{p}\,\bar{q}}{n}}$

(b) LCL for p-chart = $\bar{p} - 3\sqrt{\dfrac{(\bar{p}(1-\bar{p})}{n}} = \bar{p} - 3\sqrt{\dfrac{\bar{p}\,\bar{q}}{n}}$

8. Overall sample fraction:

$$\bar{\bar{p}} = \frac{\Sigma \bar{p}_i}{k}$$

where $\bar{\bar{p}}$ = overall sample fraction
$\bar{p}_i$ = sample fraction in the i^{th} sample
k = total number of samples

QUESTIONS

23.1 Given below are twenty statements. Indicate in each case whether it is true or false:

(a) The concept of quality indicates consistency but not reliability.

(b) SPC can be applied to any type of problem.

(c) The grand mean, $\overline{\overline{x}}$, does not capture any additional information other than what the individual sample means show.

(d) With the help of control charts, one can identify inherent variation.

(e) The term 'CL' denotes either of the two control limits in a control chart.

(f) The 3σ limits were first proposed and developed by Dr. Walter A. Shewhart for control charts.

(g) $\overline{x}$-charts and *R*-charts are control charts for qualitative variables that take on numerical values.

(h) TQM involves consideration of processes in all the major areas with which management is concerned.

(i) Occasionally, outliers can be a result of inherent variation.

(j) Acceptance sampling attempts to control the quality during the manufacturing process.

(k) In extremely complex systems, about 80 per cent of defects and errors can be attributed to only 20 per cent of the causes.

(l) Ishikawa diagrams and fishbone diagrams are two different diagrams used in quality controls.

(m) Control charts show step-by-step approach to statistical process control.

(n) A device to monitor the level of process output is in the form of an *R*-chart.

(o) TQM can be applied only to manufacturing industries and not to service industries.

(p) An OC curve can be used to determine either consumer's risk or producer's risk.

(q) A *p*-chart is used to monitor an attribute.

(r) A control chart can show us only random variation.

(s) Pareto diagrams are not used in quality control.

(t) The term AOQ stands for Average Operating Quality.

Multiple Choice Questions (23.2 to 23.14)

23.2 Which of the following cannot be considered quality?

(a) Consistency (b) Conformance to previously set standards

(c) Fitness for use (d) Luxuriousness

23.3 An OC curve is used to determine

(a) Producer's risk (b) Consumer's risk

(c) (a) and (b) (d) None of these

23.4 TQM uses the technique of

(a) R-chart (b) $\overline{X}$ chart (c) Fishbone diagram

(d) Pareto diagram (e) (b) and (c) (f) (c) and (d)

23.5 Which of the following diagrams is *not* used in quality control?

(a) R-chart (b) $\overline{X}$ chart

(c) Pareto diagram (d) OC curve

23.6 The UCL for an R chart is
(a) $\bar{R}\,D_3$ (b) $\bar{\bar{X}} + A_2\,\bar{R}$
(c) $\bar{R}\,[1 - (3d_3/d_2)]$ (d) $\bar{R}\,D_4$

23.7 In respect of attributes, which of the following charts is used?
(a) R chart (b) p-chart
(c) $\bar{X}$-chart (d) control chart for C
(e) none of these

23.8 Suppose the number of defects observed in 8 woollen carpets passed as satisfactory are:

No. of carpet	1	2	3	4	5	6	7	8
No. of defects	3	4	5	4	3	2	6	5

Which is the value of UCL?
(a) 8.5 (b) 10 (c) 12 (d) None of these

23.9 With reference to the OC curve, which of the following statements is *not* true when it increases?
(a) The steepness of the curve decreases.
(b) The zone between the acceptable qualities and the rejectable qualities widens.
(c) It becomes more steep.
(d) (a) and (b)
(e) (b) and (c)

23.10 The abbreviation CQI stands for
(a) Complete quality implementation (b) Continuous quality increases
(c) Constant quality increments (d) Continuous quality improvement

23.11 The process is said to be out of control in a $\bar{X}$-chart when the points are
(a) above the LCL (b) below the UCL
(c) between UCL and LCL (d) above the UCL
(e) below the LCL (f) (d) and (e)

23.12 R-charts and $\bar{X}$-charts are
(a) Charts for number of defects (b) Charts for attributes
(c) Charts for variables (d) None of these

23.13 A batch is rejected in double sampling, when
(a) $d_2 > c_2$ (b) $d_1 > c_2$
(c) Either (a) or (b) (d) Neither (a) nor (b)

23.14 TQM is used in situations when
(a) Some processes are extremely complicated.
(b) Companywide commitment is needed.
(c) Control charts are not useful.
(d) All of these
(e) (a) and (b) but not (c)

23.15 What is statistical quality control? How is it useful to industry?

23.16 Why should we use the term 'statistical quality control' instead of a shorter term 'quality control'?

23.17 What is a control chart? Describe how it is constructed and used.

23.18 What is meant by the process control in industrial statistics?

23.19 Write a brief note on the method of constructing control charts for $\bar{x}$ and R, giving the formulae for the upper and the lower control limits in both the cases.

23.20 Differentiate between control limits and tolerance limits.

23.21 Describe briefly the working of the p-chart.

23.22 Discuss the utility of Statistical Quality Control (SQC) from the producer's as well as consumer's point of view.

23.23 Write a note on "The Importance of an Efficient System of Statistical Quality Control in Modern Industry".

23.24 What do you understand by the acceptance sampling plan? Describe briefly the procedure for single and double-sampling plans.

23.25 Write a note on the operating characteristic curve, giving an example based on hypothetical data.

23.26 Explain the following terms in the context of SQC:

(a) Specification limits

(b) Tolerance limits

(c) Control limits

23.27 How does statistical quality control help the industry? Describe the procedure for drawing a control chart during production and indicate how you would detect lack of control in the production process.

23.28 State the advantages of a control chart for attributes over that for variables.

23.29 Explain the theoretical background of a control chart.

23.30 "The control charts make it possible to distinguish between those variations that are due to chance causes and those due to assignable causes." Explain the terms 'chance cause' and 'assignable cause', and elucidate the statement.

23.31 Illustrate situations where sampling inspection of manufactured products is used instead of complete inspection.

23.32 Distinguish between process control and product control. State the different types of acceptance sampling plans explaining their merits and demerits.

23.33 It is said that quality must be built into a product and that a control chart cannot cause a product to have high quality. How, then, does it make a contribution towards better quality?

23.34 In acceptance sampling, how producer's risk and consumer's risk correspond to Type **I** and Type **II** error in hypothesis testing? Explain.

23.35 What is an OC curve? Which OC curve would be called ideal?

23.36 What indications about lack of control are provided by a control chart? Work out control limits on an $\bar{x}$-chart for samples of size 4 if the process has to meet a lower and an upper specification limit of 79 cm and 104 cm, respectively.

23.37 The fraction defective $\bar{p}$ based on a sample of size $n = 400$ is 0.45. Compute the upper and lower control limits for (a) np-chart and (b) p-chart.

23.38 The mean number for matches in a box packed by machines has been found, from experience, to be 40, with a standard deviation of 2 matches. Results from recent tests on batches of random samples of 10 boxes show mean contents of 36, 35, 35, 34, 39, 41, 38, 39, 37 and 42. Ascertain whether the process is under control.

23.39 In order to determine whether or not a process producing bronze castings is in control, 20 subgroups of size 6 are taken. The quality characteristic of interest is the weight of the castings and it is found that $\bar{x}$ is 3.126 gm, and $R = 0.009$ gm.

(a) Estimate the standard deviation of the weight of castings.

(b) Assuming that the process is in control, find upper and lower control limits for the subgroup means.

(c) Assuming that the process is in control, find upper and lower control limits for the subgroup ranges.

23.40 A company is engaged in the manufacture of ball-bearings with the specification that they should have a mean diameter of 0.55 cm and a standard deviation of 0.01 cm. In order to determine whether the production is according to specifications, a sample of 6 ball-bearings is taken every half-hour and the mean diameter of the six is computed.

(a) Design a decision rule whereby one can ensure that the ball-bearings constantly meet the requirements.

(b) Show how to represent the decision graphically.

(c) How could even better control of the process be maintained?

23.41 The number of defects per unit inspected in rational subgroups of size 5 and for a sample of size 10 are as follows:

Sample no.	*Number of defects*					C_I	μ_i
1	2	1	0	1	2	6	6/5
2	1	1	0	0	1	3	3/5
3	2	2	2	1	0	7	7/5
4	1	0	0	0	1	2	2/5
5	3	1	2	1	1	8	8/5
6	3	1	2	0	2	8	8/5
7	1	1	0	0	1	3	3/5
8	2	2	1	1	2	8	8/5
9	2	1	1	0	1	5	1
10	3	2	3	1	0	9	9/5

Set up an appropriate control chart and give your comments.

23.42 A shampoo manufacturer wants that the contents of a bottle should measure 60 ±1 ml net. A SQC operation is established and the following data are obtained:

Sample no.	*Quantity (ml)*			
1	60.6	62.0	60.4	61.0
2	61.0	59.5	59.8	60.5
3	60.3	59.0	59.5	60.0
4	60.3	60.5	61.0	59.4
5	61.2	64.0	62.0	60.0

(a) Construct $\bar{x}$ and R-charts based on these five samples.

(b) What points, if any, have gone out of control?

23.43 A company manufactures a product that is packed in 1 kg tins. It utilizes an automatic filling equipment. It takes a sample of 5 cans every two hours and measures the filling in each of the

5 cans. The following table gives the measurements of filling (in grams) in the last 5 samples. Set up a control chart and state whether the process is under control. Assume $A_3 = 0.58$, $D_3 = 0$ and $D_4 = 2.115$.

Sample no.	*Individual measurements*				
1	1,001	1,002	1,000	998	999
2	996	998	1,001	998	999
3	995	1,000	1,003	1,001	1,002
4	1,000	1,001	999	998	1,002
5	994	996	996	1,000	999

23.44 A company has installed a machine to manufacture ball-bearings having a mean diameter of 5.74 mm and a standard deviation of 0.08 mm. In order to know whether the machine is working properly, the company takes a sample of 6 ball-bearings every 2 hours and the mean diameter is calculated from the sample.

(a) Design a decision rule to ensure that the ball-bearings produced by the company conform to required standards.

(b) Draw a control chart to show the decision rule in (a).

23.45 The following table shows data covering 24 consecutive production days on the number of defectives found in daily samples of 200. From the preliminary figures we found p, σ_p and the 3σ upper and lower control limits (UCL and LCL). Three points fell outside these limits, and investigation showed that on the first two of these three days, there were unusually large defectives because three new men had joined. On the third day, the die had been worn and actually fractured at the end.

Find the preliminary values for p, σ_p, UCL and LCL, and also revised values after eliminating the data for the three days for which assignable causes have been established.

Production day no.	*Number defective*	*Production day no.*	*Number defective*
1	10	13	12
2	5	14	15
3	10	15	8
4	12	16	14
5	11	17	4
6	9	18	10
7	19	19	11
8	4	20	11
9	12	21	26
10	27	22	13
11	25	23	10
12	9	24	11

23.46 You are given values of sample means ($\bar{X}$) and range (R) for 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
Sample means	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, $n = 5$, $A_2 = 0.58$, $D_2 = 0$ and $D_3 = 2.115$.

23.47 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	1	3	0	0	0	1	2
4	5	0	9	4	10	3	6														

Compute the values for an appropriate control chart, and give your comments.

23.48 A machine is designed to produce ball-bearings with mean diameter of 0.574 cm, and standard deviation of 0.008 cm. 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 the sample.

Design a rule whereby one can be fairly certain that the quality of products is conforming to the required standards. Give a sketch of the control chart.

23.49 Twenty tape-recorders were examined for quality control test. The number of defects for each tape-recorder are given below:

2, 4, 3, 1, 1, 2, 5, 3, 6, 7, 3, 1, 4, 2, 3, 1, 6, 4, 1, 1

Prepare a C-chart. What conclusions do your draw from it?

23.50 The following table gives the number of defects observed in 10 woollen carpets passing as satisfactory. Construct the control chart for the number of defects.

Carpet Number	1	2	3	4	5	6	7	8	9	10
Number of Defects	3	4	5	6	3	3	5	3	6	2

STATISTICAL PACKAGE FOR SOCIAL SCIENCES (SPSS)

As the name signifies, the SPSS is a software package especially designed to facilitate statisticians in the computation and analysis of varying types of data.

As is commonly known, handling of mass data without the aid of computers becomes very monotonous and tedious. Now that computers are in common use, it is a great relief to statisticians as they need not perform lengthy and difficult calculations manually. In addition, their use has cut down enormously the time required in calculations.

In this section, we will explain the procedure for using SPSS. Since a variety of problems can be handled through SPSS, the section covers some examples (mostly from the book) relating to different topics.

HOW TO DO FREQUENCY ANALYSIS USING SPSS

The **frequencies**' procedure provides statistics and graphical displays that are useful for describing several types of variables. For a first look at your data, the **frequencies**' procedure is a good place to start.

In order to explain how to do frequency analysis using SPSS, we have taken one example from Chapter 3. (Question No. 3.20).

As the data in the book are raw, before doing the frequency analysis, we should convert the raw data into categories. To convert the continuous data into categorical data, SPSS has a feature called 'Recode' in its Transform menu.

Click the TRANSFORM menu …
Click the RECODE..
Select the DIFFERENT VARIABLES option.

Figure SPSS-1 will pop up.

Select the variable ACCEPT and click the arrow button to push the variable to the right side, and give a new name RANGE for the new variable to be created (since we are going to convert the continuous variable into range variable.)

*Contributed by:
K Ashok Kumar, M Sc (Statistics) Statistician, SPSS South Asia Private Ltd., Bangalore
and
Girish P, BE, MBA (Marketing) Marketing Executive, SPSS South Asia Private Ltd., Bangalore.

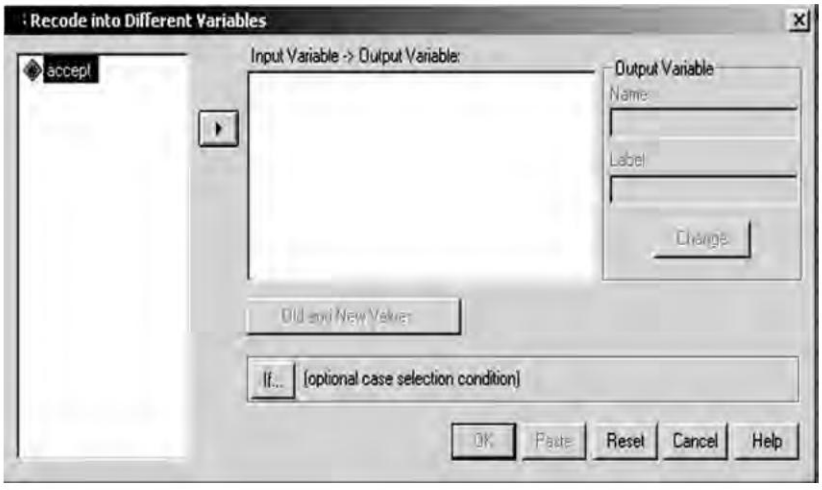


Fig. SPSS-1

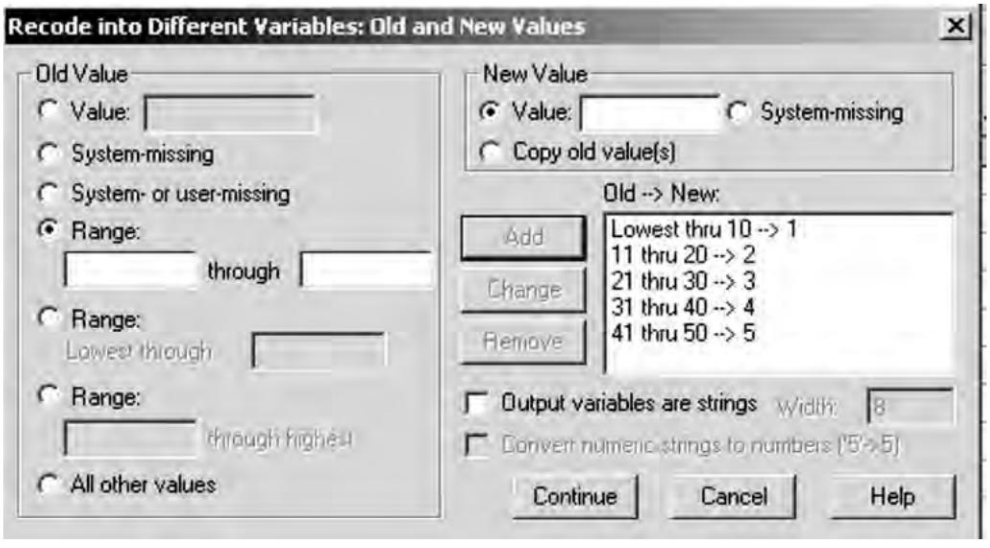


Fig. SPSS-2

Click the OLD and NEW VALUES option button.

Choose the option LOWEST THROUGH and enter the value 10 in the text box, and on the right hand side of new value enter value 1. Click ADD button so that the range LOWEST THROUGH 10 is updated and from now it will be called category 1.

Choose the fourth option and enter the value range 11 through 20, and on the right-hand side of new value, enter the value 2 and click ADD button so that values 11 to 20 will be categorised as 2nd category.

Similarly, enter the value 21 through 30, 31 through to 40, 41 through 50 and the corresponding value on the right-hand side as 3, 4, 5 respectively and update the values everytime by clicking the ADD button. By adopting this procedure, the continuous value that we had, has now been categorised.

Click CONTINUE and click OK.

Now, in data editor, a new variable called range is created.

To run a frequency analysis for the range variable that we have created now—

from the ANALYSE menu choose DESCRIPTIVE.

Choose FREQUENCIES.

This dialog box pops on the screen.

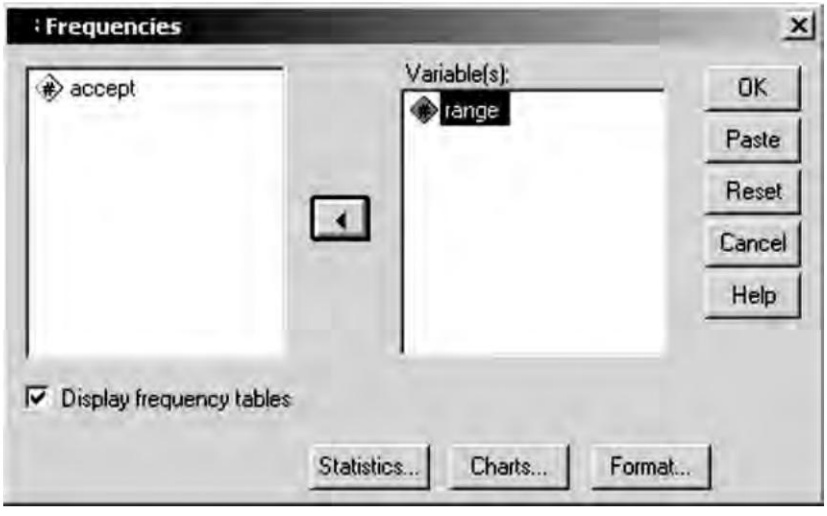


Fig. SPSS-3

Push the variable range from the left-hand side to the right side and click OK.

A new window opens up which is the output viewer (where all the outputs are displayed).

Table SPSS-1 Range for Accept

		Frequency	*Per cent*	*Valid Per cent*	*Cumulative Per cent*
Valid	1.00	2	4.0	4.0	4.0
	2.00	3	6.0	6.0	10.0
	3.00	12	24.0	24.0	34.0
	4.00	18	36.0	36.0	70.0
	5.00	15	30.0	30.0	100.0
	Total	50	100.0	100.0	

The output gives the frequency, per cent, valid per cent, and cumulative per cent for different categories in variable range.

To do a bar graph for the variable range—

from the graph menu choose BAR.

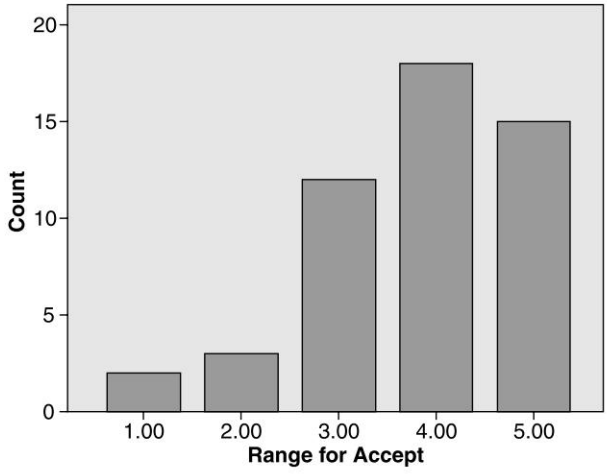


Fig. SPSS-4 **Range for Accept**

Choose simple and click DEFINE.
Move the variable range into the category axis and click OK.

MEASURES OF CENTRAL TENDENCY, DISPERSION AND SKEWNESS

To find mean, median, mode, standard deviation, variance and skewness*
Choose DESCRIPTIVE from the Analysis menu.
Choose FREQUENCIES option and click STATISTICS option.
Check the option mean, median, mode, standard deviation, variance and skewness

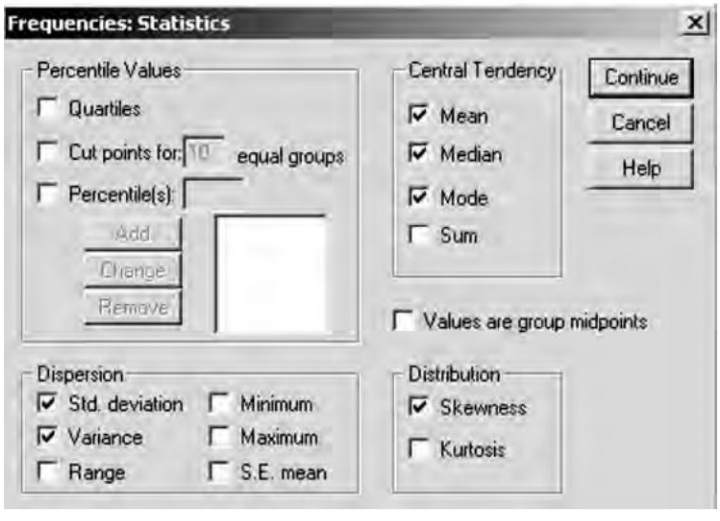


Fig. SPSS-5

*Same data as in Table SPSS-3.

Output for Measure of Central Tendency and Measure of Dispersion

Table SPSS-2

N	Valid	6400
	Missing	0
Mean		42.06
Median		41.00
Mode		39
Std. Deviation		12.290
Variance		151.032
Skewness		.299
Std. Error of Skewness		.031

The report gives the measures of central tendency and dispersion values.

The mean, median and the modal for the given data are 42.06, 41 and 39 respectively. The standard deviation is 12.29, and the variance is 151.032. The data are slightly skewed to the right as skewness is positive, being 0.299.

CHI-SQUARE ANALYSIS

To perform chi-square analysis*, choose DESCRIPTIVE from the analysis menu.

Choose CROSSTABS.

The crosstabulation window pops up.

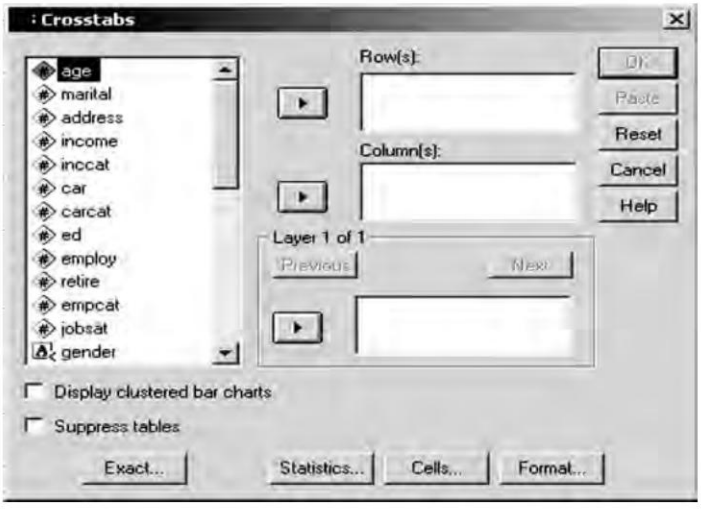


Fig. SPSS-6

*The data for this analysis have been taken from the 'demodata' included in the Students' Centre of the website.

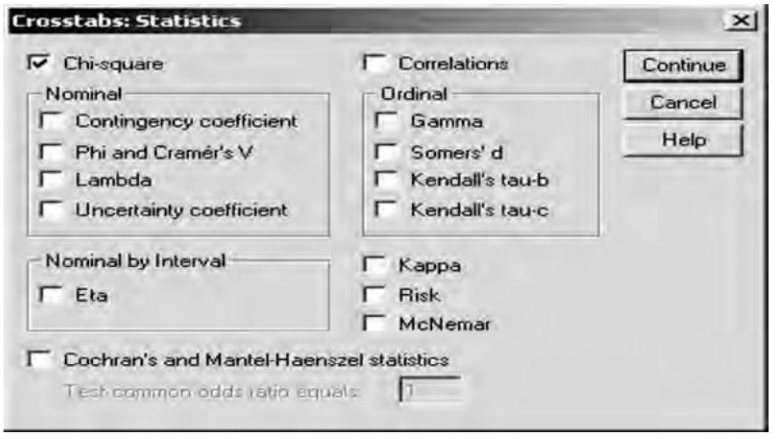


Fig. SPSS-7

Move one variable to the rows field and another variable to the column field.
Click the statistics option and check the checkbox chi-square.
Click CONTINUE and click OK.

Table SPSS-3 **Level of Education and Income Category in Thousands: Crosstabulation**

Count

		Income category in thousands				
		Under $25	*$25 - $49*	*$50 - $74*	*$75+*	*Total*
Level of education	Did not complete high school	322	537	224	307	1390
	High school degree	378	730	326	502	1936
	Some college	241	511	248	360	1360
	College degree	196	490	259	410	1355
	Post-undergraduate degree	37	120	63	139	359
Total		1174	2388	1120	1718	6400

The degrees of freedom (df) are $(r - 1)(c - 1) = (5 - 1)(4 - 1) = 12$.

Table SPSS-4 **Chi-Square Tests**

	Value	*df*	*Asymp. Sig. (2-sided)*
Pearson Chi-Square	87.404(a)	12	.000
Likelihood Ratio	87.564	12	.000
Linear-by-Linear Association	77.012	1	.000
N of Valid Cases	6400		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 62.83.

ANALYSIS OF VARIANCE

One-way Anova

To perform **ANOVA***, set up the two hypotheses.

Null Hypothesis (H_0) The means are same for the three groups.

Alternative Hypothesis (H_1) The means are different for the three groups.

From the **Analyse** menu.

Choose compare means, then choose **One-Way ANOVA.**

Move the *grade* into the dependent list and *group* into the factor option and click OK.

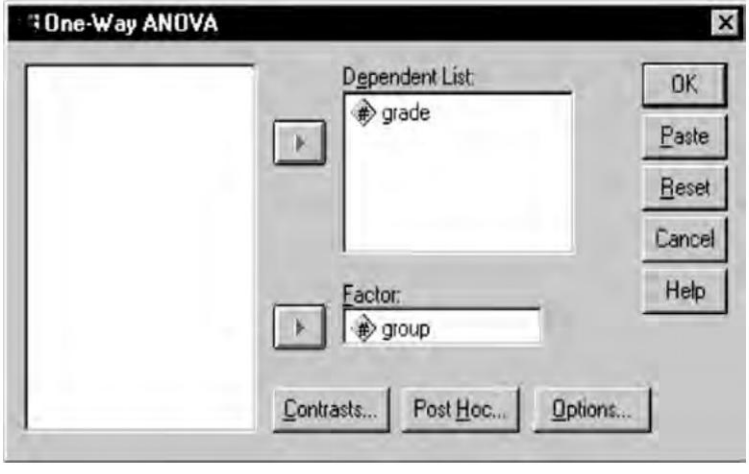


Fig. SPSS-8

Table SPSS-5 ANOVA

Grade	*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	10.000	2	5.000	2.308	.142
Within Groups	26.000	12	2.167		
Total	36.000	14			

Inference In one-way ANOVA, the total variation is partitioned into two components. *Between Groups* represents variation of the *group means* around the *overall mean*. *Within Groups* represents variation of the individual scores around their respective group means. Sig indicates the significance level of the F-test. Small significance values (<.05) indicate group differences. In this example, the significance level is not less than .05 (0.142). The null hypothesis is accepted so that the grade means for all the three groups are same.

*In order to describe this function, the data have been taken from Question No. 15.20 of Chapter 15.

Two-Way ANOVA

To perform a Two-Way ANOVA
From the ANALYSE menu
General Linear Model and then choose Univariate.
(To describe the procedure for two-way Anova, we have taken data from Example 14.3 on *Page 367*).

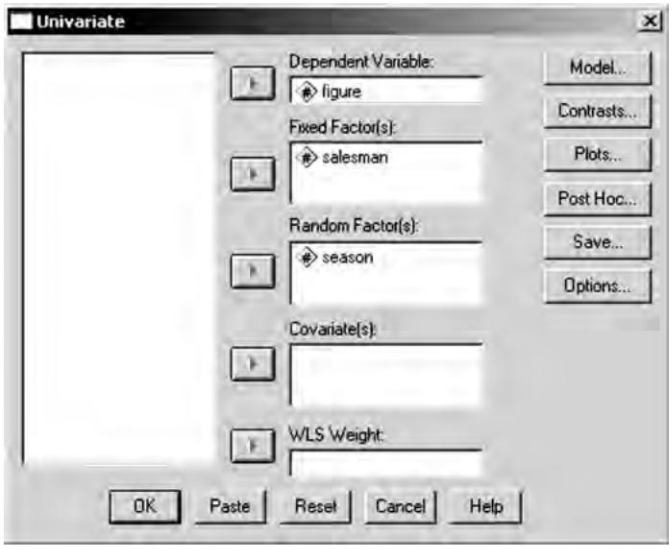


Fig. SPSS-9

Move the variable *figure* into the dependent variable option and move *salesman* into the fixed factor option and the *season* into the random factor and click OK.

Table SPSS-6 Tests of Between-Subjects Effects

Dependent Variable: figure

Source		*Type III Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
Intercept	Hypothesis	10800.000	1	10800.000	675.000	.001
	Error	32.000	2	16.000(a)		
salesman	Hypothesis	42.000	3	14.000	.618	.629
	Error	136.000	6	22.667(b)		
season	Hypothesis	32.000	2	16.000	.706	.531
	Error	136.000	6	22.667(c)		
salesman * season	Hypothesis	136.000	6	22.667	.	.
	Error	.000	0	.(c)		

a MS(season)
b MS(salesman * season)
c MS(Error)

Inference The Sig. value for salesman 0.629 (greater than 0.05) indicates that there is no significant difference among salesman. The Sig. value for season 0.531 (greater than 0.05) indicates that there is no significant difference among seasons.

TESTING OF HYPOTHESIS

How to perform t test

To explain t test using SPSS we have taken a numerical problem from Chapter 13 (*Q. No. 13.40.*).

Prob: The Null hypothesis test whether the mean score of exams taken before and after training is the same.

Alternative hypothesis treats that the means score of exams taken before and after training is not the same.

To perform *t* test analysis:

From the **analyse** menu

Select compare means option

Select independent t test option

The t test window pops up

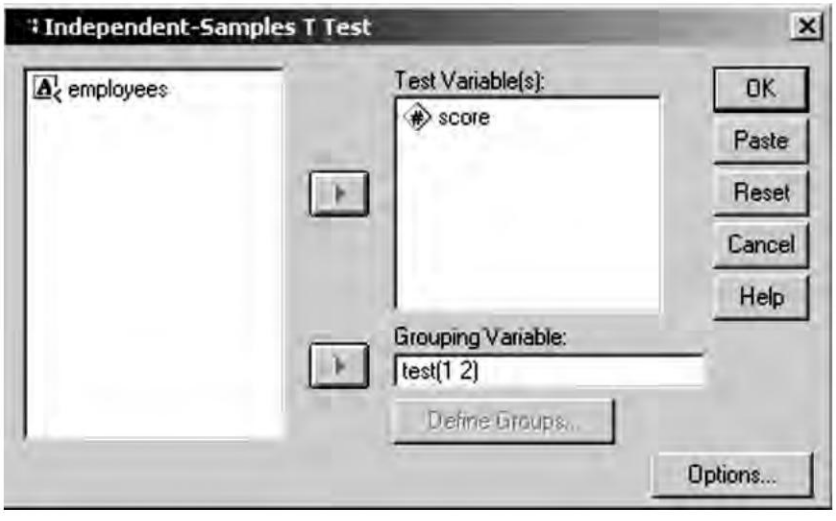


Fig. SPSS-10

Take the variable score in the test variable and the variable test (before training or after training) in the grouping variable and click DEFINE GROUPS and define the two values of the variable test (1 for before and 2 for after). Click CONTINUE and click OK.

In the output viewer the t test results will be displayed.

Table SPSS-7 Independent Samples Test

		t test for Equality of Means		
		t	*df*	*Sig. (2-tailed)*
Score	Equal variances assumed	–.679	18	.506

T test was performed to find whether the mean score between before and after training is same or not.

With the very high significant value, .506 (greater than 0.05) indicates that the null hypothesis should be accepted. In other words, there is no difference between pre-training and post-training scores.

REGRESSION ANALYSIS

Linear Regression estimates the coefficients of the linear equation, involving one or more independent variables, that best predict the value of the dependent variable.

To perform regression analysis, Question No. 16.34 from the book is taken.

From the ANALYSE menu

Choose REGRESSION

Choose LINEAR…

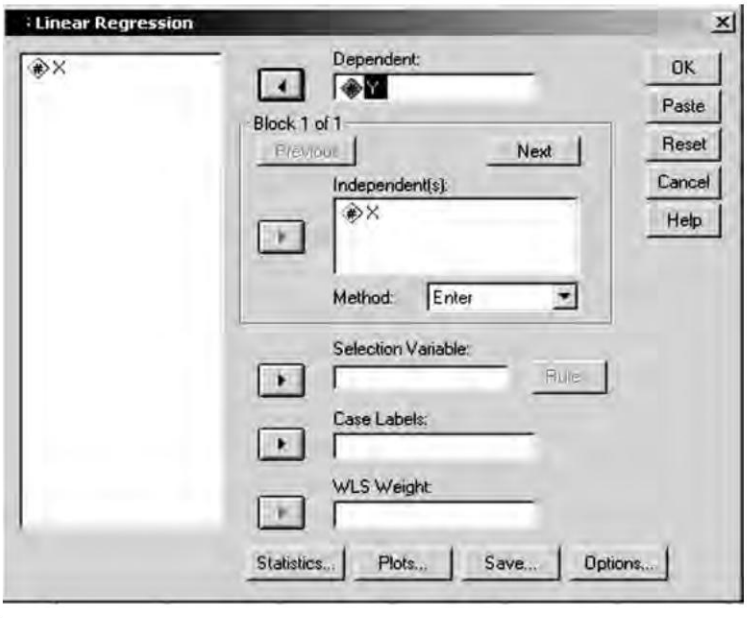


Fig. SPSS-11

Select the variable *Y* as the dependent variable and *X* as the independent variable and click OK.

Table SPSS-8 Model Summary

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
1	.694(a)	.482	.379	5.59212

a Predictors: (Constant), X

Table SPSS-9 Coefficients(a)

Model		*Unstandardised Coefficients*		*Standardised Coefficients*	*t*	*Stg.*
		B	*Std. Error*	*Beta*		
1	(Constant)	–1.968	10.397		–.189	.857
	X	.814	.377	.694	2.158	.083

a Dependent Variable:Y

Inference The correlation coefficient *r*, is the correlation between the observed and predicted values of the dependent variable. The values of *r* for models produced by the regression procedure range from 0 to 1. Larger values of *r* indicate stronger relationships.

The coefficient of determination, r^2, indicates the proportion of variation in the dependent variable explained by the regression model. The values of r^2 range from 0 to 1.

The value of r^2 0.482 explains that only 48% of the dependent variable is explained by independent variable.

The regression equation is Y = –1.968 + 0.814X

CORRELATION

To do correlation analysis correlation analysis is used to find the linear relationship between two variables. Question No. 17.33 from the book, that relates to two series—marks in English and marks in Mathematics of 5 students is taken here.

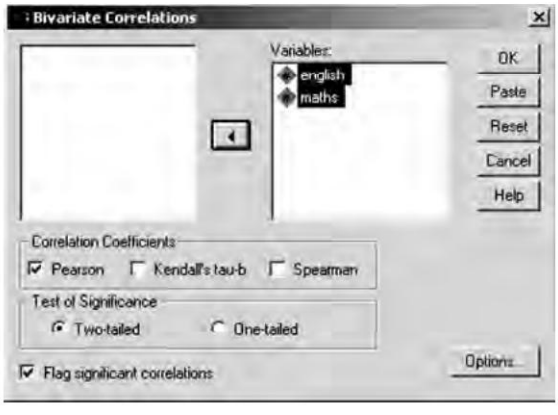


Fig. SPSS-12

Table SPSS-10 Correlations

		Marks in English	*Marks in Maths*
Marks in English	Pearson Correlation	1	.908(*)
	Sig. (2-tailed)		.033
	N	5	5
Marks in Maths	Pearson Correlation	.908(*)	1
	Sig. (2-tailed)	.033	
	N	5	5

Correlation coefficient *r* between marks in English and marks in Maths is .908 and is significant at the 0.05 level (2-tailed).

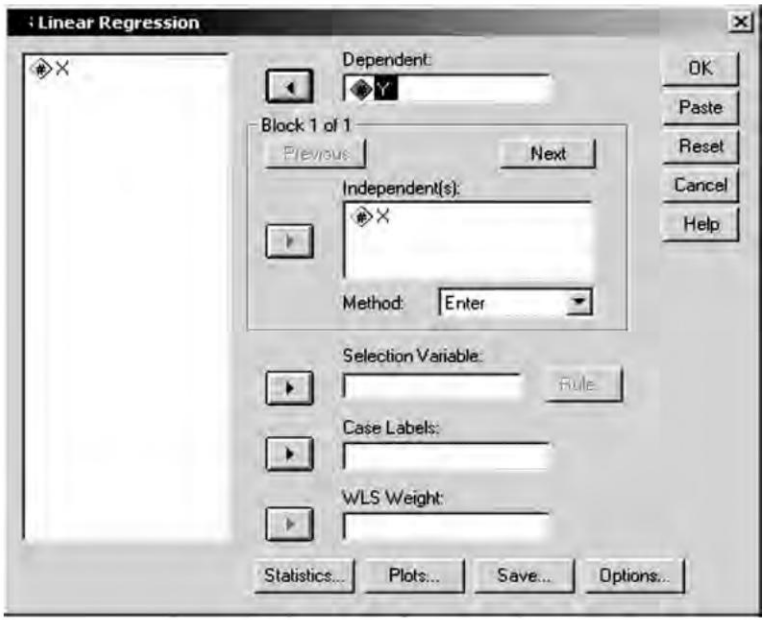


Fig. SPSS-13

Choose the variable Y in the dependent variable and X in the independent variable and click OK.

Table SPSS-11 Model Summary

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
1	.694(a)	.482	.379	5.59212

a Predictors: (Constant), X

Table SPSS-12 Coefficients(a)

Model		*Unstandardised Coefficients*		*Standardised Coefficients*	*t*	*Stg.*
		B	*Std. Error*	*Beta*		
1	(Constant)	–1.968	10.397		–.189	.857
	X	.814	.377	.694	2.158	.083

a Dependent Variable: Y

MULTIPLE REGRESSION

To explain how multiple regression works Question No. 18.20 has been taken.

To perform multiple regression analysis:

From the analyse menu, choose regression and move X1 to the dependent variable and X2 and X3 to the independent variables

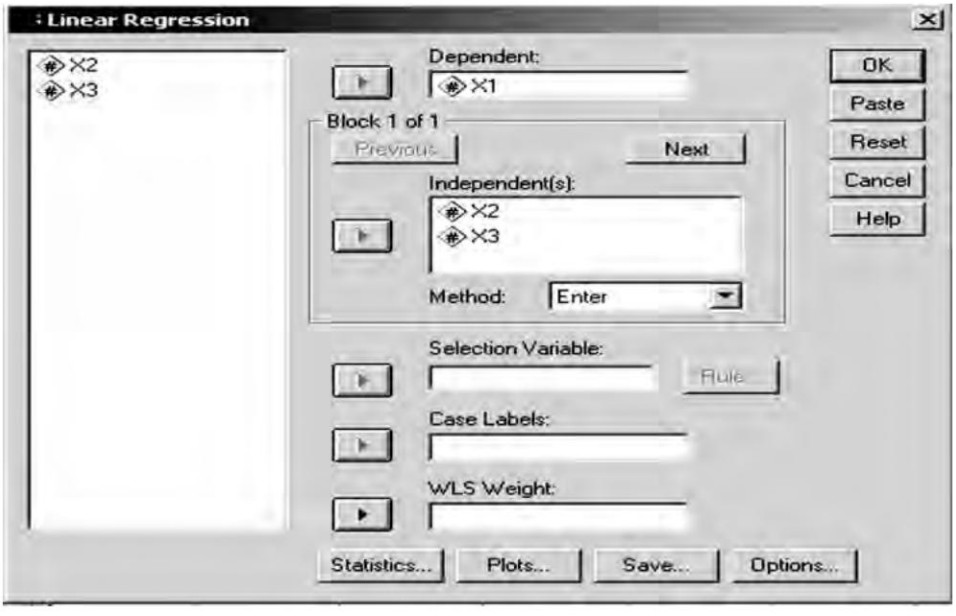


Fig. SPSS-14

Table SPSS-13 Model Summary

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
1	.986(a)	.972	.965	3.53612

a Predictors: (Constant), X3, X2

Table SPSS-14 **Coefficients(a)**

Model		*Unstandardised Coefficients*		*Standardised Coefficients*	*t*	*Stg.*
		B	*Std. Error*	*Beta*		
1	(Constant)	.471	2.672		.176	.865
	X2	1.894	1.495	.398	1.267	.246
	X3	3.054	1.615	.593	1.891	.101

a Dependent Variable: X1

The regression equation is:

$X_1 = 0.471 + 1.894\ X_2 + 3.054\ X_3$

Inference R, the multiple correlation coefficient, is the correlation between the observed and predicted values of the dependent variable. The values of R for models produced by the regression procedure range from 0 to 1. Larger values of R indicate stronger relationships.

R^2 is the proportion of variation in the dependent variable explained by the regression model. The values of R^2 range from 0 to 1.

The value of R^2 0.972 explains that as much as 97.2% of the total variation observed in the dependent variable X_1 is explained by the regression equation.

Partial Correlation Coefficient

To perform Partial Correlation Coefficient*

From the ANALYSE menu

Choose CORRELATE

PARTIAL…

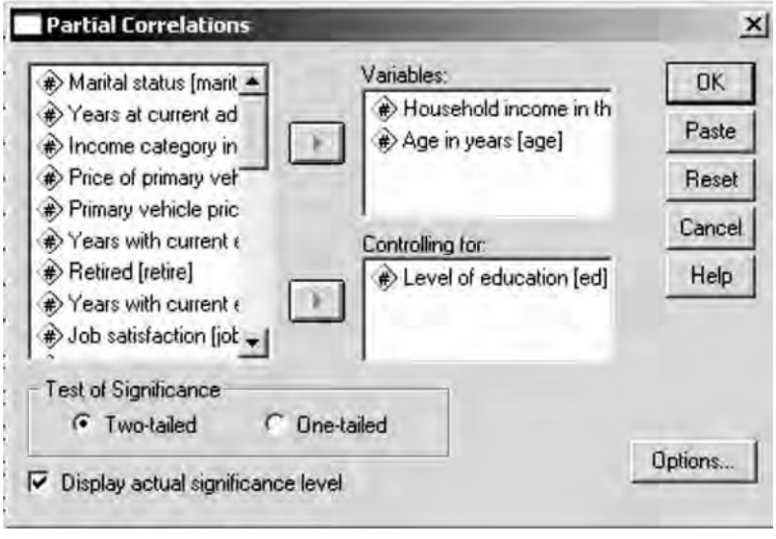


Fig. SPSS-15

*The data for this topic have been taken from the Demodata, included in the Students' Centre of the website.

Move the two variables for which you want to find the correlation and move the controlling variable in the controlling for option and click OK.

Table SPSS-15 Correlations

Control Variables			*Household income in thousands*	*Age in years*
Level of education	Household income in thousands	Correlation	1.000	.352
		Significance (2-tailed)	.	.000
		df	0	6397
	Age in years	Correlation	.352	1.000
		Significance (2-tailed)	.000	.
		df	6397	0

Inference The partial correlation coefficient between income and age controlled with education is 0.352.

TIME SERIES ANALYSIS

To perform centered moving average (data taken from Example 19.7)
From the TRANSFORM menu
Choose CREATE TIME SERIES

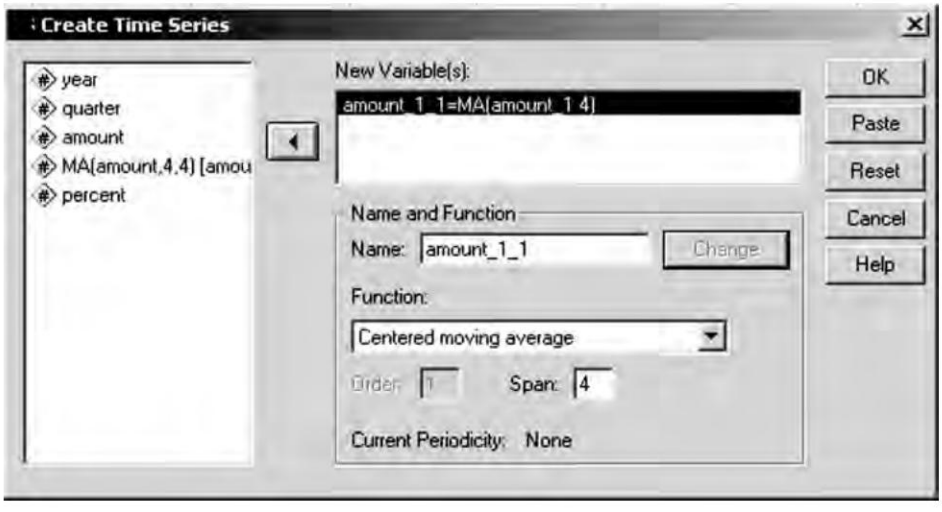


Fig. SPSS-16

Move the variable amount into the new variable and from the function choose CENTERED moving average and in the span option enter the value 4 and click CHANGE button and click OK.

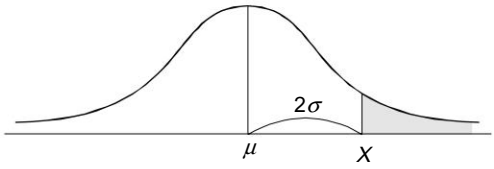
Year	*Quarter*	*Amount*	*Amount_1*
1.00	1.00	68.00	
1.00	2.00	62.00	
1.00	3.00	61.00	63.13
1.00	4.00	63.00	62.25
2.00	1.00	65.00	61.13
2.00	2.00	58.00	60.25
2.00	3.00	56.00	60.38
2.00	4.00	61.00	61.38
3.00	1.00	68.00	62.88
3.00	2.00	63.00	64.50
3.00	3.00	63.00	65.50
3.00	4.00	67.00	65.25
4.00	1.00	70.00	63.88
4.00	2.00	59.00	62.38
4.00	3.00	56.00	60.50
4.00	4.00	62.00	58.75
5.00	1.00	60.00	57.63
5.00	2.00	55.00	56.50
5.00	3.00	51.00	
5.00	4.00	58.00	

The 4-quarter moving average will be displayed in the data editor, a new variable will be created by the name amount_1. The moving averages are shown in the last column of the above table, the original data are in column (3) of the above table.

APPENDIX
Tables

- Appendix Table 1 — Areas Under the Normal Curve
- Appendix Table 2 — Percentile Values of the Student's t Distribution
- Appendix Table 3 — Binomial Probabilities
- Appendix Table 4 — (a) Values of $e^{-\lambda}$ for Computing Poisson Probabilities
  (b) Poisson Probabilities
- Appendix Table 5 — The Chi-square Distribution (Values of χ^2)
- Appendix Table 6 — Percentiles of the F-Distribution
- Appendix Table 7 — Values for Rank Correlation for Combined Areas in Both Tails
- Appendix Table 8 — Critical Values of T in the Wilcoxon Matched-Pairs Test
- Appendix Table 9 — Partial Table of Critical Values of U in the Mann-Whitney Test
- Appendix Table 10 — Critical Values of D in the Kolmogorov-Smirnov One-Sample Test
- Appendix Table 11 — Control Chart Factors
- Appendix Table 12 — Random Numbers

Appendix Table 1 Areas Under the Normal Curve



Example

$$z = \frac{X - \mu}{\sigma}$$

$P[z > 1] = .1587$

$P[z > 1.96] = .0250$

Normal Deviate z	*.00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

Appendix Table 2 Percentile Values of the Student's *t* Distribution

df \ $1-\alpha$	.75	.90	.95	.975	.99	.995	.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	1.330	1.734	2.101	2.552	2.878	3.992
19	.688	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	.674	1.282	1.645	1.960	2.326	2.576	3.291

Appendix Table 3 Binomial Probabilities

						p					
n	*k*	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*	*.50*
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000
	1	.0500	.1000	.1500	.2000	.2500	.3000	.3500	.4000	.4500	.5000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1562
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0516
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0009	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0312
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0004	.0046	.0815	.0459	.0865	.1361	.1875	.2322	.2627	.2734

(Contd.)

Appendix Table 3 *(Contd.)*

		p									
n	*k*	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208

(Contd.)

Appendix Table 3 *(Contd.)*

		p									
n	*k*	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*	*.50*
12	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095
	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571
	6	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095
	7	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095
	8	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571
	9	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873
	10	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009
	2	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056
	3	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222
	4	.0037	.0348	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611
	5	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222
	6	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833
	7	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095
	8	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833
	9	.0000	.0000	.0000	.0003	.0018	.0066	.0183	.0408	.0762	.1222
	10	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0093	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417

(Contd.)

Appendix Table 3 *(Contd.)*

		p									
n	*k*	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
15	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
16	0	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001	.0000
	1	.3706	.3294	.2097	.1126	.0535	.0228	.0087	.0030	.0009	.0002
	2	.1463	.2745	.2775	.2111	.1336	.0732	.0353	.0150	.0056	.0018
	3	.0359	.1423	.2285	.2463	.2079	.1465	.0888	.0468	.0215	.0085
	4	.0061	.0514	.1311	.2001	.2252	.2040	.1553	.1014	.0572	.0278
	5	.0008	.0137	.0555	.1201	.1802	.2099	.2008	.1623	.1123	.0667
	6	.0001	.0028	.0180	.0550	.1101	.1649	.1982	.1983	.1684	.1222
	7	.0000	.0004	.0045	.0197	.0524	.1010	.1524	.1889	.1969	.1746
	8	.0000	.0001	.0009	.0055	.0197	.0487	.0923	.1417	.1812	.1964
	9	.0000	.0000	.0001	.0012	.0058	.0185	.0442	.0840	.1318	.1746
	10	.0000	.0000	.0000	.0002	.0014	.0056	.0167	.0392	.0755	.1222
	11	.0000	.0000	.0000	.0000	.0002	.0013	.0049	.0142	.0337	.0667
	12	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0040	.0115	.0278
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0029	.0085
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
17	0	.4181	.1668	.0631	.0225	.0075	.0023	.0007	.0002	.0000	.0000
	1	.3741	.3150	.1893	.0957	.0426	.0169	.0060	.0019	.0005	.0001
	2	.1575	.2800	.2673	.1914	.1136	.0581	.0260	.0102	.0035	.0010
	3	.0415	.1556	.2359	.2393	.1893	.1245	.0701	.0341	.0144	.0052
	4	.0076	.0605	.1457	.2093	.2209	.1868	.1320	.0796	.0411	.0182
	5	.0010	.0175	.0668	.1361	.1914	.2081	.1849	.1379	.0875	.0472
	6	.0001	.0039	.0236	.0680	.1276	.1784	.1991	.1839	.1432	.0944
	7	.0000	.0007	.0065	.0267	.0668	.1201	.1685	.1927	.1841	.1484
	8	.0000	.0001	.0014	.0084	.0279	.0644	.1134	.1606	.1883	.1855
	9	.0000	.0000	.0003	.0021	.0093	.0276	.0611	.1070	.1540	.1855
	10	.0000	.0000	.0000	.0004	.0025	.0095	.0263	.0571	.1008	.1484
	11	.0000	.0000	.0000	.0001	.0005	.0026	.0090	.0242	.0525	.0944
	12	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0021	.0215	.0472
	13	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0021	.0068	.0182
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052

(Contd.)

Appendix Table 3 *(Contd.)*

		p									
n	*k*	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
17	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
18	0	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000	.0000
	1	.3763	.3002	.1704	.0811	.0338	.0126	.0042	.0012	.0003	.0001
	2	.1683	.2835	.2556	.1723	.0958	.0458	.0190	.0069	.0022	.0006
	3	.0473	.1680	.2406	.2297	.1704	.1046	.0547	.0246	.0095	.0031
	4	.0093	.0700	.1592	.2153	.2130	.1681	.1104	.0614	.0291	.0117
	5	.0014	.0218	.0787	.1507	.1988	.2017	.1664	.1146	.0666	.0327
	6	.0002	.0052	.0301	.0816	.1436	.1873	.1941	.1655	.1181	.0708
	7	.0000	.0010	.0091	.0350	.0820	.1376	.1792	.1892	.1657	.1214
	8	.0000	.0002	.0022	.0120	.0376	.0811	.1327	.1734	.1864	.1669
	9	.0000	.0000	.0004	.0033	.0139	.0386	.0794	.1284	.1694	.1855
	10	.0000	.0000	.0001	.0008	.0042	.0149	.0385	.0771	.1248	.1669
	11	.0000	.0000	.0000	.0001	.0010	.0046	.0151	.0374	.0742	.1214
	12	.0000	.0000	.0000	.0000	.0002	.0012	.0047	.0145	.0354	.0708
	13	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0044	.0134	.0327
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0039	.0117
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0009	.0031
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
19	0	.3774	.1351	.0456	.0144	.0042	.0011	.0003	.0001	.0000	.0000
	1	.3774	.2852	.1529	.0685	.0268	.0093	.0029	.0008	.0002	.0000
	2	.1787	.2852	.2428	.1540	.0803	.0358	.0138	.0046	.0013	.0003
	3	.0533	.1796	.2428	.2182	.1517	.0869	.0422	.0175	.0062	.0018
	4	.0112	.0798	.1714	.2182	.2023	.1491	.0909	.0467	.0203	.0074
	5	.0018	.0266	.0907	.1636	.2023	.1916	.1468	.0933	.0497	.0222
	6	.0002	.0069	.0374	.0955	.1574	.1916	.1844	.1451	.0949	.0518
	7	.0000	.0014	.0122	.0443	.0974	.1525	.1844	.1797	.1443	.0961
	8	.0000	.0002	.0032	.0166	.0487	.0981	.1489	.1797	.1771	.1442
	9	.0000	.0000	.0007	.0051	.0198	.0514	.0980	.1464	.1771	.1762
	10	.0000	.0000	.0001	.0013	.0066	.0220	.0528	.0976	.1449	.1762
	11	.0000	.0000	.0000	.0003	.0018	.0077	.0233	.0532	.0970	.1442
	12	.0000	.0000	.0000	.0000	.0004	.0022	.0083	.0237	.0529	.0961
	13	.0000	.0000	.0000	.0000	.0001	.0005	.0024	.0085	.0233	.0518
	14	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0082	.0222
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0022	.0074
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

(Contd.)

Appendix Table 3 *(Contd.)*

		p									
n	*k*	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*	*.50*
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000
	2	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002
	3	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011
	4	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046
	5	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148
	6	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370
	7	.0000	.0020	.0160	.0545	.1124	.1643	.1844	.1659	.1221	.0739
	8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201
	9	.0000	.0001	.0011	.0074	.0271	.0654	.1158	.1597	.1771	.1602
	10	.0000	.0000	.0002	.0020	.0099	.0308	.0686	.1171	.1593	.1762
	11	.0000	.0000	.0000	.0005	.0030	.0120	.0336	.0710	.1185	.1602
	12	.0000	.0000	.0000	.0001	.0008	.0039	.0136	.0355	.0727	.1201
	13	.0000	.0000	.0000	.0000	.0002	.0010	.0045	.0146	.0366	.0739
	14	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0049	.0150	.0370
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0049	.0148
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0046
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Appendix Table 4(a) Values of $e^{-\lambda}$ for Computing Poisson Probabilities

λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$
0.1	0.90484	2.6	0.07427	5.1	0.00610	7.6	0.00050
0.2	0.81873	2.7	0.06721	5.2	0.00552	7.7	0.00045
0.3	0.74082	2.8	0.06081	5.3	0.00499	7.8	0.00041
0.4	0.67032	2.9	0.05502	5.4	0.00452	7.9	0.00037
0.5	0.60653	3.0	0.04979	5.5	0.00409	8.0	0.00034
0.6	0.54881	3.1	0.04505	5.6	0.00370	8.1	0.00030
0.7	0.49659	3.2	0.04076	5.7	0.00335	8.2	0.00027
0.8	0.44933	3.3	0.03688	5.8	0.00303	8.3	0.00025
0.9	0.40657	3.4	0.03337	5.9	0.00274	8.4	0.00022
1.0	0.36788	3.5	0.03020	6.0	0.00248	8.5	0.00020
1.1	0.33287	3.6	0.02732	6.1	0.00224	8.6	0.00018
1.2	0.30119	3.7	0.02472	6.2	0.00203	8.7	0.00017
1.3	0.27253	3.8	0.02237	6.3	0.00184	8.8	0.00015
1.4	0.24660	3.9	0.02024	6.4	0.00166	8.9	0.00014
1.5	0.22313	4.0	0.01832	6.5	0.00150	9.0	0.00012
1.6	0.20190	4.1	0.01657	6.6	0.00136	9.1	0.00011
1.7	0.18268	4.2	0.01500	6.7	0.00123	9.2	0.00010
1.8	0.16530	4.3	0.01357	6.8	0.00111	9.3	0.00009
1.9	0.14957	4.4	0.01228	6.9	0.00101	9.4	0.00008
2.0	0.13534	4.5	0.01111	7.0	0.00091	9.5	0.00007
2.1	0.12246	4.6	0.01005	7.1	0.00083	9.3	0.00007
2.2	0.11080	4.7	0.00910	7.2	0.00075	9.7	0.00006
2.3	0.10026	4.8	0.00823	7.3	0.00068	9.8	0.00006
2.4	0.09072	4.9	0.00745	7.4	0.00061	9.9	0.00005
2.5	0.08208	5.0	0.00674	7.5	0.00055	10.0	0.00005

Appendix Table 4(b) Poisson Probabilities

					λ					
k	*.005*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
0	.9950	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
1	.0050	.0099	.0192	.0291	.0384	.0476	.0565	.0653	.0738	.0823
2	.0000	.0000	.0002	.0004	.0008	.0012	.0017	.0023	.0030	.0037
3	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

					λ					
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0002	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

					λ					
k	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

					λ					
k	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008

(Contd.)

Appendix Table 4(b) *(Contd.)*

11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

					λ					
k	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1734	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

					λ					
k	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0280	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

					λ					
k	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892

(Contd.)

Appendix Table 4(b) *(Contd.)*

4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0255
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001

					λ					
k	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0245	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0264
13	.0058	.0065	.0073	.0081	.0089	.0098	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

(Contd.)

Appendix Table 4(b) *(Contd.)*

	λ									
k	*7.1*	*7.2*	*7.3*	*7.4*	*7.5*	*7.6*	*7.7*	*7.8*	*7.9*	*8.0*
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	..0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0699	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1486	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1413
8	.1321	.1337	.1351	.1363	.1373	.1382	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0613	.0640	.0667	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

Appendix Table 5 The Chi-square Distribution (Values of χ^2)

Area in the right tail

df \ α	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	$.0^3157$	$.0^3628$	.00393	.0158	0.642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0401	.103	.211	.446	.713	2.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.414	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.139	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	36.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703
40	22.164	23.838	26.509	29.051	32.345	34.872	39.335	44.165	47.269	51.805	55.759	60.436	63.692	73.402
50	29.707	31.644	34.764	37.689	41.449	44.313	34.335	54.723	58.164	63.167	67.505	72.613	76.154	86.661
60	37.485	39.699	43.188	46.459	50.641	53.809	59.335	65.227	68.972	74.397	79.082	84.580	88.379	99.607

Appendix Table 6 Percentiles of the *F*-Distribution

$F_{.95(n1,\,n2)}$ $a = 0.05$

n_1 = degrees of freedom for numerator

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	2.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	21.5	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	3.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.43	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

n_2 = degrees of freedom for denominator

(Contd.)

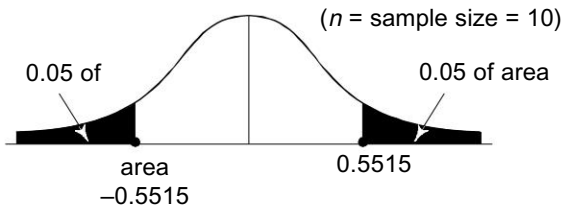
Appendix Table 6 *(Contd.)*

n_1 = degrees of freedom for numerator $F_{.99(n_1, n_2)}$ $\alpha = 0.01$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

n_2 = degrees of freedom for denominator

Appendix Table 7 Values for Rank Correlation for Combined Areas in Both Tails



n	*0.20*	*0.10*	*0.05*	*0.02*	*0.01*	*0.002*
4	0.8000	0.8000				
5	0.7000	0.8000	0.9000	0.9000		
6	0.6000	0.7714	0.8286	0.8857	0.9429	
7	0.5357	0.6786	0.7450	0.8571	0.8929	0.9643
8	0.5000	0.6190	0.7143	0.8095	0.8571	0.9286
9	0.4667	0.5833	0.6833	0.7667	0.8167	0.9000
10	0.4424	0.5515	0.6364	0.7333	0.7818	0.8667
11	0.4182	0.5273	0.6091	0.7000	0.7455	0.8364
12	0.3986	0.4965	0.5804	0.6713	0.7273	0.8182
13	0.3791	0.4780	0.5549	0.6429	0.6978	0.7912
14	0.3626	0.4593	0.5341	0.6220	0.6747	0.7670
15	0.3500	0.4429	0.5179	0.6000	0.6536	0.7464
16	0.3382	0.4265	0.5000	0.5824	0.6324	0.7265
17	0.3260	0.4118	0.4853	0.5637	0.6152	0.7083
18	0.3148	0.3994	0.4716	0.5480	0.5975	0.6904
19	0.3070	0.3895	0.4579	0.5333	0.5825	0.6737
20	0.2977	0.3789	0.4451	0.5203	0.5684	0.6586
21	0.2909	0.3688	0.4351	0.5078	0.5545	0.6455
22	0.2829	0.3597	0.4241	0.4963	0.5426	0.6318
23	0.2767	0.3518	0.4150	0.4852	0.5306	0.6186
24	0.2704	0.3435	0.4061	0.4748	0.5200	0.6070
25	0.2646	0.3362	0.3977	0.4654	0.5100	0.5962
26	0.2588	0.3299	0.3894	0.4564	0.5002	0.5856
27	0.2540	0.3236	0.3822	0.4481	0.4915	0.5757
28	0.2490	0.3175	0.3749	0.4401	0.4828	0.5660
29	0.2443	0.3113	0.3685	0.4320	0.4744	0.5567
30	0.2400	0.3059	0.3620	0.4251	0.4665	0.5479

Appendix Table 8 Critical Values of *T* in the Wilcoxon Matched-Pairs Test

N	*Level of Significance for One-Tail Test*		
	.025	.01	.005
	Level of Significance for Two-Tail Test		
	.05	.02	.01
6	0	—	—
7	2	0	—
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Appendix Table 9 Partial Table of Critical Values of U in the Mann-Whitney Test

Critical Values for One-Tail Test at $\alpha = .025$ or a Two-Tail Test at $\alpha = .05$

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	66	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

Critical Values for One-Tail Test at $\alpha = .05$ or a Two-Tail Test at $\alpha = .10$

n_2 \ n_1	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

Appendix Table 10 Critical Values of D in the Kolmogorov-Smirnov One-Sample Test

Sample Size	*Level of Significance for D = Maximum* $[F_0(X) - F_e(X)]$				
n	*.20*	*.15*	*.10*	*.05*	*.01*
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.21	.22	.24	.27	.32
30	.19	.20	.22	.24	.29
35	.18	.19	.21	.23	.27
Over 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Appendix Table 11 Control Chart Factors

Number of Observations in Sample	Chart for Averages: Factors for Control Limits			Chart for Standard Deviations: Factors for Central Line		Chart for Standard Deviations: Factors for Control Limits				Chart for Ranges: Factors for Central Line		Chart for Ranges: Factors for Control Limits				
n	A	A_1	A_2	c_2	$1/c_2$	B_1	B_2	B_3	B_4	d_2	$1/d_2$	d_3	D_1	D_1	D_3	D_4
2	2.121	3.760	1.880	.5642	1.7725	0	1.843	0	3.267	1.128	.8865	0.853	0	3.686	0	3.267
3	1.732	2.394	1.023	0.7236	1.3820	0	1.858	0	2.568	1.693	0.5907	0.888	0	4.358	0	2.575
4	1.500	1.880	0.729	0.7979	1.2533	0	1.808	0	2.266	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	1.596	0.577	0.8407	1.1894	0	1.756	0	2.089	2.326	0.4299	0.864	0	4.918	0	2.115
6	1.225	1.410	0.483	0.8636	1.1512	0.026	1.711	0.030	1.970	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	1.277	0.419	0.8882	1.1259	0.015	1.672	0.118	1.882	2.704	0.3698	0.833	0.205	5.203	0.076	1.924
8	1.061	1.175	0.373	0.9027	1.1078	0.167	1.638	0.185	1.815	2.847	0.3512	0.820	0.387	5.307	0.136	1.864
9	1.000	1.094	0.337	0.9139	1.0942	0.219	1.609	0.239	1.761	2.970	0.3367	0.808	0.546	5.394	0.184	1.816
10	0.949	1.028	0.308	0.9227	1.0837	0.262	1.584	0.284	1.716	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.973	0.285	0.9300	1.0753	0.299	1.561	0.321	1.679	3.173	0.3152	0.787	0.812	5.534	0.256	1.744
12	0.866	0.925	0.266	0.9359	1.0684	0.331	1.541	0.354	1.646	3.258	0.3069	0.778	0.924	5.592	0.284	1.716
13	0.832	0.884	0.249	0.9410	1.0627	0.359	1.523	0.382	1.618	3.336	0.2998	0.770	1.026	5.846	0.308	1.692
14	0.802	0.849	0.235	0.9453	1.0579	0.384	1.507	0.406	1.594	3.407	0.2935	0.762	1.121	5.693	0.329	1.671
15	0.775	0.816	0.223	0.9490	1.0537	0.406	1.492	0.428	1.572	3.472	0.2880	0.755	1.207	5.737	0.348	1.652
16	0.750	0.788	0.212	0.9523	1.0501	0.427	1.478	0.448	1.552	3.523	0.2831	0.749	1.285	5.779	0.364	1.636
17	0.728	0.762	0.203	0.9551	1.0470	0.445	1.465	0.466	1.534	3.588	0.2787	0.743	1.359	5.817	0.379	1.621
18	0.707	0.738	0.194	0.9576	1.0442	0.461	1.454	0.482	1.518	3.640	0.2747	0.738	1.426	5.854	0.392	1.608
19	0.688	0.717	0.187	0.9599	1.0418	0.477	1.443	0.497	1.503	3.689	0.2711	0.733	1.490	5.888	0.404	1.596
20	0.671	0.697	0.180	0.9619	1.0396	0.491	1.433	0.510	1.490	3.735	0.2677	0.729	1.548	5.992	0.414	1.586
21	0.655	0.679	0.173	0.9638	1.0376	0.504	1.424	0.523	1.477	3.778	0.2647	0.724	1.605	5.950	0.425	1.575
22	0.640	0.662	0.167	0.9655	1.0358	0.516	1.415	0.534	1.466	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.647	0.162	0.9670	1.0342	0.527	1.407	0.545	1.455	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.632	0.157	0.9684	1.0327	0.538	1.399	0.555	1.445	3.895	0.2567	0.712	1.759	0.031	0.452	1.548
25	0.600	0.619	0.153	0.9696	1.0313	0.548	1.392	0.565	1.435	3.931	0.2544	0.709	1.804	6.058	0.459	1.541

Appendix Table 12 Random Numbers

First Thousand

	1–4	*5–8*	*9–12*	*13–16*	*17–20*	*21–24*	*25–28*	*29–32*	*33–36*	*37–40*
1	23 15	75 48	59 01	83 72	59 93	76 24	97 08	86 95	23 03	67 44
2	05 54	55 50	43 10	53 74	35 08	90 61	18 37	44 10	96 22	13 43
3	14 87	16 03	50 32	40 43	62 23	50 05	10 03	22 11	54 38	08 34
4	38 97	67 49	51 94	05 17	58 53	78 80	59 01	94 32	42 87	16 95
5	97 31	26 17	18 99	75 53	08 70	94 25	12 58	41 54	88 21	05 13
6	11 74	26 93	81 44	33 93	08 72	32 79	73 31	18 22	64 70	68 50
7	43 36	12 88	59 11	01 64	56 23	93 00	90 04	99 43	64 07	40 36
8	93 80	62 04	78 38	26 80	44 91	55 75	11 89	32 58	47 55	25 71
9	49 54	01 31	81 08	42 98	41 87	69 53	82 96	61 77	73 80	95 27
10	36 76	87 26	33 37	94 82	15 69	41 95	96 86	70 45	27 48	38 80
11	07 09	25 23	92 24	62 71	26 07	06 55	84 53	44 67	33 84	53 20
12	43 31	00 10	81 44	86 38	03 07	52 55	51 61	48 89	74 29	46 47
13	61 57	00 63	60 06	17 36	37 75	63 14	89 51	23 35	01 74	69 93
14	31 35	28 37	99 10	77 91	89 41	31 57	97 64	48 62	58 48	69 19
15	57 04	88 65	26 27	79 59	36 82	90 52	95 65	46 35	06 53	22 54
16	09 24	34 42	00 68	72 10	71 37	30 72	97 57	56 09	29 82	76 50
17	97 95	53 50	18 40	89 48	83 29	52 23	08 25	21 22	53 26	15 87
18	93 73	25 95	70 43	78 19	88 85	56 67	16 68	26 95	99 64	45 69
19	72 62	11 12	25 00	92 26	82 64	35 66	65 94	34 71	68 75	18 67
20	61 02	07 44	18 45	37 12	07 94	95 91	73 78	66 99	53 61	93 78
21	97 83	98 54	74 33	05 59	17 18	45 47	35 41	44 22	03 42	30 00
22	89 16	09 71	92 22	23 29	06 37	35 05	54 54	89 88	43 81	63 61
23	25 96	68 82	20 62	87 17	92 65	02 82	35 28	62 48	91 95	48 83
24	81 44	33 17	19 05	04 95	48 06	74 69	00 75	67 65	01 71	65 45
25	11 32	25 49	31 42	36 23	43 86	08 62	49 76	67 42	24 52	32 45

Second Thousand

	1–4	*5–8*	*9–12*	*13–16*	*17–20*	*21–24*	*25–28*	*29–32*	*33–36*	*37–40*
1	64 75	58 38	85 84	12 22	59 20	17 69	61 56	55 95	04 59	59 47
2	10 30	25 22	89 77	43 63	44 30	38 11	24 90	67 07	34 82	33 28
3	71 01	79 84	95 51	30 85	03 74	66 59	10 28	87 53	76 56	91 49
4	60 01	25 56	05 88	41 03	48 79	79 65	59 01	69 78	80 00	36 66
5	37 33	09 46	56 49	16 14	28 02	48 27	45 47	55 44	55 36	50 90
6	47 86	98 70	01 31	59 11	22 73	60 62	61 28	22 34	69 16	12 12
7	38 04	04 27	37 64	16 78	95 78	39 32	34 93	24 88	43 43	87 06
8	73 50	83 09	08 83	05 48	00 78	36 66	93 02	95 56	46 04	53 36
9	32 62	34 64	74 84	06 10	43 24	20 62	83 73	19 32	35 64	39 69
10	97 59	19 95	49 36	63 03	51 06	62 06	99 29	75 95	32 05	77 34
11	74 01	23 19	55 59	79 09	69 82	66 22	42 40	15 96	74 90	75 89
12	56 75	42 64	57 13	35 10	50 14	90 96	63 36	74 69	09 63	34 88
13	49 80	04 99	08 54	83 12	19 98	08 52	82 63	72 92	92 36	50 26
14	43 58	48 96	47 24	87 85	66 70	00 22	15 01	93 99	59 16	23 77
15	16 65	37 96	64 60	32 57	13 01	35 74	28 36	36 73	05 88	72 29

(Contd.)

Appendix Table 12 *(Contd.)*

16	48 50	26 90	55 65	32 25	87 48	31 44	68 02	37 31	25 29	63 67
17	96 76	55 46	92 36	31 68	62 30	48 29	63 83	52 23	81 66	40 94
18	38 92	36 15	50 80	35 78	17 84	23 44	41 24	63 33	99 22	81 28
19	77 95	88 16	94 25	22 50	55 87	51 07	30 10	70 60	21 86	19 61
20	17 92	82 80	65 25	58 60	87 71	02 64	18 50	64 65	79 64	81 70
21	94 03	68 59	78 02	31 80	44 99	41 05	41 05	31 87	43 12	15 96
22	47 46	06 04	79 56	23 04	84 17	14 37	28 51	67 27	55 80	03 68
23	47 85	65 60	88 51	99 28	24 39	40 64	41 71	70 13	46 31	82 88
24	57 61	63 46	53 92	29 86	20 18	10 37	57 65	15 62	98 69	07 56
25	08 30	09 27	04 66	75 26	66 10	57 18	87 91	07 54	22 22	20 13

ANSWERS TO CHAPTER-END QUESTIONS

TRUE OR FALSE

Barring Chapters 3 and 4, each chapter has a number of "Concept" questions. These are invariably the first questions and are given in the form of statements. The students are asked to indicate whether the statement is true or false. Below are given their answers, where a 'T' indicates that the statement is true and 'F' indicates that the statement is false.

Chapter–1: Question 1

(a) F (b) F (c) T (d) F (e) F (f) T (g) F (h) T (i) T (j) F

Chapter–3: Question 1

(a) F (b) F (c) F (d) T (e) F (f) T (g) T

Chapter–6: Question 1

(a) F (b) F (c) F (d) T (e) T (f) F (g) T (h) F (i) T (j) F (k) F (l) T

Chapter–7: Question 1

(a) F (b) F (c) T (d) T (e) F (f) T (g) T (h) F (i) T (j) F

Chapter–8: Question 1

(a) T (b) F (c) F (d) F (e) T (f) F (g) T (h) F (i) T (j) F
(k) T (l) T

Chapter–9: Question 1

(a) T (b) T (c) T (d) T (e) F (f) F (g) F (h) T (i) F (j) T

Chapter–10: Question 1

(a) T (b) F (c) F (d) T (e) T (f) F (g) F (h) F (i) T (j) T

Chapter–11: Question 1

(a) F (b) T (c) F (d) F (e) T (f) F (g) T (h) F (i) T (j) T
(k) F (l) T

Chapter–12: Question 1

(a) T (b) T (c) F (d) F (e) F (f) T (g) T (h) F (i) T (j) T (k) F (l) T

Chapter–13: Question 1
(a) T (b) F (c) F (d) F (e) T (f) T (g) T (h) F (i) F (j) T (k) F (l) T

Chapter–14: Question 1
(a) F (b) F (c) T (d) F (e) F (f) T (g) F (h) T (i) T (j) T (k) T (l) T

Chapter–15: Question 1
(a) T (b) T (c) F (d) F (e) F (f) T (g) T (h) F (i) F (j) T

Chapter–16: Question 1
(a) F (b) F (c) T (d) F (e) T (f) T (g) T (h) F (i) T (j) T

Chapter–17: Question 1
(a) F (b) T (c) T (d) F (e) F (f) F (g) T (h) F (i) T (j) F (k) T (l) F
(m) T (n) T (o) F

Chapter–18: Question 1
(a) F (b) T (c) F (d) T (e) F (f) T (g) T (h) T (i) T (j) T (k) T (l) F

Chapter–19: Question 1
(a) F (b) T (c) F (d) F (e) T (f) T (g) F (h) T (i) F (j) T (k) F (l) T
(m) F (n) F

Chapter–20: Question 25
(a) T (b) F (c) T (d) T (e) F (f) T (g) T (h) T (i) F (j) F

Chapter–21: Question 1
(a) T (b) T (c) F (d) F (e) T (f) T (g) T (h) F (i) F (j) T (k) T (l) T

Chapter–22: Question 1
(a) T (b) F (c) T (d) F (e) F (f) T (g) T (h) T (i) T (j) F (k) F (l) F

Chapter–23: Question 1
(a) F (b) T (c) F (d) F (e) F (f) T (g) T (h) T (i) T (j) F (k) T (l) F
(m) T (n) T (o) F (p) F (q) T (r) F (s) F (t) F

MULTIPLE-CHOICE QUESTIONS

Chapter 1

1.2 (d) **1.3** (c) **1.4** (d) **1.5** (d)
1.6 (d)

Chapter 3

3.2 (c) **3.3** (c) **3.4** (c) **3.5** (d)
3.6 (d) **3.7** (c) **3.8** (c) **3.9** (a)
3.10 (d)

Chapter 6

6.2 (b) **6.3** (c) **6.4** (b) **6.5** (a)
6.6 (b) **6.7** (d) **6.8** (d) **6.9** (d)
6.10 (c) **6.11** (d) **6.12** (a) **6.13** (b)

Chapter 7

7.2 (a)	**7.3** (c)	**7.4** (c)	**7.5** (c)
7.6 (b)	**7.7** (d)	**7.8** (d)	**7.9** (c)
7.10 (a)	**7.11** (c)		

Chapter 8

8.2 (c)	**8.3** (c)	**8.4** (c)	**8.5** (c)
8.6 (c)	**8.7** (d)	**8.8** (c)	

Chapter 9

9.2 (d)	**9.3** (c)	**9.4** (a)	**9.5** (d)
9.6 (d)	**9.7** (d)	**9.8** (b)	**9.9** (c)
9.10 (d)	**9.11** (d)	**9.12** (c)	**9.13** (b)
9.14 (c)	**9.15** (d)	**9.16** (a)	**9.17** (c)

Chapter 10

10.2 (d)	**10.3** (d)	**10.4** (c)	**10.5** (c)
10.6 (e)	**10.7** (d)	**10.8** (d)	**10.9** (d)
10.10 (a)	**10.11** (d)	**10.12** (d)	**10.13** (c)
10.14 (c)	**10.15** (c)		

Chapter 11

11.2 (d)	**11.3** (d)	**11.4** (d)	**11.5** (c)
11.6 (d)	**11.7** (c)	**11.8** (c)	**11.9** (c)
11.10 (d)	**11.11** (c)	**11.12** (d)	**11.13** (d)
11.14 (c)	**11.15** (b)	**11.16** (c)	

Chapter 12

12.2 (e)	**12.3** (c)	**12.4** (c)	**12.5** (e)
12.6 (a)	**12.7** (d)	**12.8** (d)	**12.9** (a)
12.10 (b)	**12.11** (d)	**12.12** (c)	

Chapter 13

13.2 (c)	**13.3** (a)	**13.4** (b)	**13.5** (c)
13.6 (d)	**13.7** (c)	**13.8** (e)	**13.9** (e)
13.10 (e)	**13.11** (c)	**13.12** (c)	**13.13** (c)
13.14 (b)			

Chapter 14

14.2 (b)	**14.3** (d)	**14.4** (a)	**14.5** (e)
14.6 (b)	**14.7** (d)	**14.8** (d)	**14.9** (b)
14.10 (b)	**14.11** (d)		

Chapter 15

15.2 (a)	**15.3** (c)	**15.4** (d)	**15.5** (c)
15.6 (a)	**15.7** (c)	**15.8** (c)	**15.9** (c)
15.10 (e)			

Chapter 16

16.2 (d)	**16.3** (b)	**16.4** (d)	**16.5** (d)
16.6 (c)	**16.7** (c)	**16.8** (c)	**16.9** (d)
16.10 (c)	**16.11** (d)	**16.12** (d)	**16.13** (c)

Chapter 17

17.2 (d)	**17.3** (f)	**17.4** (d)	**17.5** (b)
17.6 (b)	**17.7** (c)	**17.8** (b)	**17.9** (c)
17.10 (b)	**17.11** (d)	**17.12** (b)	

Chapter 18

18.2 (c)	**18.3** (c)	**18.4** (c)	**18.5** (a)
18.6 (c)	**18.7** (a)	**18.8** (b)	**18.9** (c)

Chapter 19

19.2 (b)	**19.3** (b)	**19.4** (d)	**19.5** (d)
19.6 (c)	**19.7** (d)	**19.8** (e)	**19.9** (c)
19.10 (d)	**19.11** (c)	**19.12** (d)	**19.13** (c)
19.14 (e)	**19.15** (c)		

Chapter 20

20.2 (e)	**20.3** (a)	**20.4** (d)	**20.5** (b)
20.6 (b)	**20.7** (a)	**20.8** (d)	**20.9** (b)
20.10 (c)	**20.11** (c)		

Chapter 21

21.2 (b)	**21.3** (a)	**21.4** (d)	**21.5** (c)
21.6 (b)	**21.7** (c)	**21.8** (c)	**21.9** (c)
21.10 (d)	**21.11** (c)	**21.12** (c)	**21.13** (d)
21.14 (b)			

Chapter 22

22.2 (e)	**22.3** (e)	**22.4** (b)	**22.5** (a)
22.6 (c)	**22.7** (d)	**22.8** (d)	**22.9** (e)
22.10 (c)	**22.11** (e)	**22.12** (d)	

Chapter 23

23.2 (d)	**23.3** (c)	**23.4** (f)	**23.5** (d)
23.6 (d)	**23.7** (b)	**23.8** (b)	**23.9** (d)
23.10 (d)	**23.11** (f)	**23.12** (c)	**23.13** (c)
23.14 (d)			

SELECTED NUMERICAL PROBLEMS

Chapter 6

6.18 Male 20% Female 80%

6.21 Rs 180 per student per month.

6.23 (a) Mean = 5.1 Median = 5 Mode = 5 (b) No mode, Mean = 49.8, Median = 49.5

6.26 $Q_1 = 45.33$ Median 54.46
$D_7 = 61.89$ $P_{80} = 67.56$
6.28 $Q_1 = 67$, Q_2 (median) = 75, $Q_3 = 83$
6.29 (a) 2.50 (b) 0.417 (c) Arithmetic mean is a poor average used for ratios.
(d) Geometric mean is more suitable for averaging ratios.
6.30 (a) 8 (b) 15.66 **5.31** 7.2%
6.32 19.72% **5.34** 23.8
6.36 40 km/hr
6.40 A. mean = 443.4 Median = 444.5 Made = 445.1

Chapter 7

7.25 Quartile deviation 2.235
7.26 Set (a) Mean deviation 4.25
Set (b) Mean deviation 2.25
Set (a) shows greater dispersion.
7.27 Method 1: 12 kg
Method 2: 15 kg
7.30 11.55; 0.34
7.31 (a) Mean of 1st set = 8 Mean of 2nd set = 8
(b) Variance of 1st set = 18 Variance of 2nd set = 24
(c) Mean of combined sets = 8
(d) Variance of combined sets = 20.25
7.32 Mean of combined sets = 11
Variance of combined sets = 35.25
7.34 σ_2^2 will be smaller than σ_1^2.
7.36 Variance 2490.01 σ 49.9 coefficiant of Variation 31.58
7.39 Corrected $\bar{x}$ = 19.15 and corrected σ = 4.66
7.40 0–10, 10–20, 20–30, 30–40, 40–50, 50–60, 60–70
7.42 (i) Factory A (ii) Factory A (iii) Mean = 473.33, σ = 45.83
7.43 (i) Company B (ii) Company B
7.44 Mean = 6 σ = 3
7.47 $\sigma_2 = 6.64$
7.48 σ = 22.36

Chapter 8

8.23 0.7; As skewness is positive, the shape of the distribution is elongated to the right.
8.25 –0.27; The distribution is negatively skewed, i.e., skewed on the left side.
8.28 Mean = 7 Variance = 16
$\beta_1 = 1$ $\beta_2 = 0.63$
8.29 μ_4 should be 1875
8.31 Bowley's coefficient of skewness = 0.5
8.32 The four central moments about the mean are:
$\mu_2 = 177.39$ $\mu_3 = 47.982$ $\mu_4 = 95{,}009.364$
$\beta_1 = (47.982)^2/(177.39)^3$ $\beta_2 = 95{,}009.364/(177.39)^2$

Chapter 9

9.33 (a) 5/12 (b) 1/4
9.34 84%
9.35 (i) 1/35 (ii) 2/7 (iii) 24/35
9.37 15/28
9.39 Probability of losing 5/36 Probability of winning 1/6
Hence ratio 5:6 between P(L) and P(W).
9.40 0.143
9.41 0.244
9.45 (i) 0.5385 (ii) 0.1538 (iii) 0.6 (iv) 0.238
9.47 (i) 0.6667 (ii) 0.2500 (iii) 0.4545 (iv) 0.5385
9.56 (i) 0.66 (ii) 0.05
9.62 (a) 0.0068 (b) 0.0163

Chapter 10

10.18 0.712
10.19 $p = 1/5$
10.20 (a) 0.051 (b) 0.996 (c) 0.949
10.21 (i) 240 (ii) 352 (iii) 1696
10.22

Value of variable:	0	1	2	3	4	5
Expected frequency:	147	147	74	25	6	1

10.23 0.1114
10.28 Mean = 6 $\sigma = 2.449$
10.29 (i) 50 (ii) 353
10.35 (i) $P(0 \le Z \le 1) - (0 \le Z \le 0.5)$
(ii) 0.1525 (iii) 0.3830
10.37 $P(X > 30) = 0.0062$
10.41 (i) 0.2266 (ii) 0.4649 (iii) 0.1598
10.44 Mean = 30, Variance = 12

Chapter 11

11.24 (a) 70 (b) 495 (c) 91,390
11.25 (a) 15 (b) 45 (c) 300
11.26 (a) 1/3003 (b) 1/53130
11.29 (a) 2.8926 (b) 0.8264
11.32 (a) 34.13% (b) 47.72% (c) H_0 is rejected. Manufacturer's claim is not acceptable
11.33 (a) 0.9544 (b) 0.0456 (c) 0.0002 (d) 0.0030
11.35 (a) 0.9794 (b) 1–0.9988
11.41 (a) 0.1056 (b) 0.5714
11.43 (a) 10 Kg (b) 0.5 Kg

Chapter 12

12.19 17.16 to 32.84
12.20 (a) 42.59 (b) 43.41
12.21 65.016 to 66.584
12.22 0.58% to 10.54%
12.24 (a) 1.729 (b) 2.365 (c) 2.462 (d) 2.797 (e) 2.262 (f) 2.353
12.25 0.48 to 0.68
12.26 14.9% to 45.1%
12.27 (a) 12.496, 13.704 (b) 2.651, 2.809 (c) 28.280, 28.920
12.28 0.075; 0.125
12.31 (a) 0.0625 ± 0.0235 (b) 563
12.33 819

Chapter 13

13.24 $Z = -12$. Manager's claim cannot be justified.
13.25 $t = -3.57$; $df = 24$. H_0 is rejected. Average life of presses is different from 14,500 hours.
13.26 (i) $t = 1.6$, (ii) H_0 is accepted. (iii) $\bar{x}$ should be 220.65 to reject H_0.
13.27 $Z = -0.2$, both at 1% and 5% level of significance H_0 is accepted. As such the aircraft manufacturer should buy the aluminium sheets.
13.28 $Z = 0.022$. Yes. The company's assumption is acceptable.
13.29 $Z = -0.3$, Yes. The credit manager's belief is acceptable.
13.30 $Z = -2.17$, No. The stock broker's claim is rejected.
13.31 $Z = -0.268$, H_0 is accepted. The claim of insurance agent is not justifiable.
13.32 $Z = -2$. H_0 is accepted at 1% level of significance.
However, H_0 is rejected at 5% level of significance.
13.33 $t = -7.2$. H_0 that there is no difference in Type A and B batteries is rejected.
13.34 $t = -2.172$. H_0 is rejected. There is significant difference in the mean life of two brands.
13.38 $Z = -0.85$. No significant difference in the quality of two brands of bulbs.
13.39 $Z = -3.89$. H_0 is rejected. There is significant difference between average sales of A and B.
13.41 $t = 4.96$, $df = 13$. H_0 is rejected. There is significant difference in the mean life of Type I and Type II bulbs.
13.42 H_0 that there is no difference in the proportion of defective items in the two factories cannot be rejected at 5% level of significance.

Chapter 14

14.24 $\chi^2 = 11.25$, $df = 1$. H_0 is rejected both at 0.05 and 0.01 levels of significance, which means that the coin is not fair.
14.26 20.5, $df = 5$ Highly significant.
14.29 $\chi^2 = 2.111$, $df = 2$ Not significant.
14.30 2.074, $df = 4$ Not significant.
14.31 $\chi^2 = 8.434$, $df = 3$. H_0 is rejected. The given results are not commensurate with the general examination results.
14.32 $\chi^2 = 2.097$, $df = 3$. H_0 is accepted. The sampling techniques are not significantly different.
14.35 $\chi^2 = 4.72$, $df = 3$. H_0 is accepted. The experiment supports the theory.

14.38 $\chi^2 = 3.522$, $df = 2$. H_0 is rejected. Drug and sugar pills do not differ significantly in curing cold.
14.43 $\Sigma\chi^2 = 48.2$, $\Sigma df = 25$. Neither is significant at 5% level of significance. Combined χ^2 significant at 5%.

Chapter 15

15.20 $F_0 = 2.3$. There in no difference in the three teaching methods.
15.21 F (temperatures) = 0.6 and F (detergents) = 5.74. Both are not significant.
15.22 F (between machine types) = 18.4
F (between workers) = 6.58
Both the null hypotheses are rejected. Hence, (a) there is significant difference in the mean productivity of the four machines and (b) five men too differ with respect to mean productivity.
15.24 F (between machine types) = 18.4
F (between workers) = 6.58.
There is significant difference both in the mean productivity of machines as also of five men.
15.25 The second ANOVA table is identical with the one given. As such there is no change in the ratios or relative positions.
15.28 $F_{3,8} = 5.39$
15.29 $F_{3,8} = 7.78$
15.30 $F_{3,8} = 0.286$

Chapter 16

16.26 $\hat{y} = 11.7 + 1.525\,X$ Rs 36,100
16.27 $\hat{y} = -3 + 2\,X$ (a) 3,997 (b) 6,397
16.28 $\hat{y} = 0.8 + 0.27\,X$
16.30 $\hat{y} = 8.805 + 1.015\,X$ 34.18
16.31 $\hat{y} = 0.57 + 0.058\,X$ (a) 4920 units (b) 6370 units
16.32 $C = 48.8 + 0.42\,Y$ Rs 111.8
16.33 $\hat{y} = 2.6 + 2.1\,X$ 17.3 kg
16.35 $\hat{y} = 99.5 - 10\,X$ 24.5
16.36 $\hat{y} = 15.1 + 0.61\,X$ $\hat{X} = -5.2 + 1.36\,Y$
16.39 $\hat{y} = 40 + 4\,X$ $\hat{X} = 18.08 + 0.16\,Y$
(ii) Rs 140 crore (iii) Rs 42.08 crore
16.43 $B = -2 + 0.96$ A $B = 60.4$ (when A is 65)
16.44 $\hat{X} = 15.085 + 0.514\,Y$ $X = 30.505$ (when $Y = 30$)
$\hat{y} = -1.638 + 1.244\,X$ $Y = 60.562$ (when $X = 50$)
16.60 $y = -1.02 + 0.58$ b
when $y = 2.5$, $x = 6.07$
16.62 (i) $x = 0.24$ (ii) Rs 24 lakhs
$y = 32\,x + 36$

Chapter 17

17.32 $r = 0.89$ **17.33** $r = 0.91$

17.34 No change $r = 0.91$
17.36 $r = 0.6$
17.37 $r = -0.256$
17.39 $r = -0.705$
17.41 $r = 0.388$
17.44 $r = 0.665$
17.46 $r_s = 0.685$

Chapter 18

18.20 $X_1 = 0.45 + 1.9\,X_2 + 3.05\,X_3$
18.22 $X_1 = 16.479 + 0.389\,X_2 - 0.623\,X_3$
When $X_2 = 10$ and $X_3 = 22$, then X_1 would be 6.663
18.27 $r_{23.1} = 0.0966$ $R_{1.23} = 0.84$
$X_1 = 2.74\,X_2 - 0.157\,X_3$
When $X_2 = 74$ and $X_3 = 85$, then X_1 would be 189.42
18.28 $X_1 = -63.578 + 1.074\,X_2 + 1.015\,X_3$
$R_{1.23} = 0.42$
18.29 $X_1 = 8.3 + 0.41\,X_2 + 0.229\,X_3$

Chapter 19

19.35 $Y = 6.99 + 0.06\,X$ origin 1.1.1984; X-unit = Half year
19.36 $Y = 76 + 4.86\,X$ origin 1.7.1994; X-unit: 1 year
Y_{2001} = Rs 1,10,000 approx.
19.37

1974	'75	'76	'77	'78	'79	'80	'81	'82
		5.8	6.2	6.6	7.0	7.4		

(Unit Rs '00000)

19.40

QI	**QII**	**QIII**	**QIV**	
110	105	90	95	400

19.41 $\hat{Y} = 372 + 24x$
x-unit: 1 year
Origin: 1.7.1992
19.43 $Y = 90 + 2\,X$
Origin: 1.7.1992, X-unit = 1 year
19.44

QI	**QII**	**QIII**	**QIV**	
123.2	91.4	84.8	100.6	400

19.47 $Y = 12.314 - 0.6\,X - 0.857\,X^2$
Sales for 2001 = –3.798 and hence the model is not suitable.

Chapter 20

20.18 $T_{(-)} = 19$ $T_{(+)} = 101$
Reject H_0. The claim of the company is not justified.
20.19 $Z = 0.287$ The data do not show lack of randomness.
20.22 There is no difference in the substandard performance before and after the incentive scheme.
20.23 Observed value: $U = 8$. Critical value: $C = 3$
Do not reject H_0. The price of building lots in resort area A is less than that in resort area B.
20.26 H_0 is accepted. The arrangement of F's and A's was random.
20.27 H_0 is not rejected. The claim of the bus company is justified.

Chapter 21

21.36 108.07

21.37

1995	1996	1997	1998	1999	2000
100	120	125	167	108	130

21.40

1997	176
1998	141
1999	197

21.41 Chain index

1991	1992	1993	1994	1995
100	197.5	349.1	170.2	351.0

21.43 Overall increase in cost of living: 26.1%
Item-wise increase: Food 16%, Rent 20%, Clothing 25%, Fuel 25% and Miscellaneous 50%.

Chapter 22

22.27 Expected cost for testing Rs 600, for not testing Rs 360.
Hence it is more profitable to install the machine without testing it.

22.28 (a) Maximum EMV in case of two boats = Rs 4,500
(b) EVPI = Rs 5,350

22.31

Stocking Items	0	1	2	3	4
EMV in Rs	0	1.5	2.0	1.0	–1.5

The retailer should stock two items as it gives the maximum EMV.

22.34 EMV: 16 doses Rs 32.0
20 doses Rs 36.6
30 doses Rs 31.6
40 doses Rs 13.1
Since the act 'purchase 20 vaccine per week' gives the highest EMV, the optimal act for the physician would be to purchase 20 doses.

22.35 EMV: Accept proposal Rs 2.85 million
EMV: Reject proposal Rs 3.2 million
This shows that the proposal should be rejected.

22.36 EMV: Act X Rs 1,16,000 Act Y Rs 1,24,000
Act Z Rs 0. Hence Act Y should be chosen as it is the best.

22.46 (i) a_3 (ii) a_1 (iii) a_3 22.50 30000 units plant

Chapter 23

23.37 (a) 180 ± 29.85 (b) 0.45 ± 0.075

23.39 (a) 0.0035
(b) $CL_{\bar{x}} = 3.126$ $UCL_{\bar{x}} = 3.130$ $LCL_{\bar{x}} = 3.122$
(c) $CL_R = 0.009$ $UCL_R = 0.018$ $LCL_R = 0$ (zero)

23.42

	$\bar{x}$-chart	R-chart
(a)	LCL = 59.142	LCL = 0
	CL = 60.6	CL = 2
	UCL = 62.058	UCL = 4.564

(b) As all the samples are within the control limits, the process variation is also under control.

23.45

Preliminary:	$p = 0.06$	$\sigma_p = 0.0168$
	$UCL_p = 0.1104$	$LCL_p = 0.0096$
Revised:	$p = 0.05$	$\sigma_p = 0.0154$
	$UCL_p = 0.0962$	$LCL_p = 0.0038$

BIBLIOGRAPHY

Aczel, Amir D. and Jayavel Sounderpandian: *Complete Business Statistics*, New Delhi, Tata McGraw-Hill Publishing Company Limited, 2002 (5th edition).

Alan, J.B. Anderson: *Interpreting Data—A First Course in Statistics*, London, Chapman and Hall, 1989.

Berenson, Mark L. and David M. Levine: *Basic Business Statistics: Concepts and Applications*, Prentice-Hall International, Inc., 1996 (6^{th} edition).

Bryman, Alan and Duncan Cramer: *Quantitative Data Analysis for Social Scientists*, London, Routledge, 1992.

Chase, Warren and Fred Brown: *General Statistics*, John Wiley & Sons, Inc., 1997.

Ehrenberg, A.S.C.: *A Primer in Data Reduction*, Chichester, John Wiley & Sons, Inc., 1982.

Evans, Michael K: *Practical Business Forecasting*, Oxford, Blackwell Publishers, 2002.

Grant, E.L. and Leavenworth, R.W: *Statistical Quality Control*, New York, McGraw-Hill, 1996.

Levin, R.I., D.S. Rubin, J.P. Stinson and E.S. Gardner, Jr.: *Quantitative Approaches to Management,* New York, McGraw-Hill Book Company, 1992 (Eighth edition).

Levin, Richard I. and David S. Rubin: *Statistics for Management*, New Delhi, Prentice-Hall of India Private Limited, 1999 (Seventh edition).

Levine, David M., Patricia P. Ramsey and Mark L. Berenson: *Business Statistics for Quality and Productivity*, New Jersey, Prentice-Hall International, Inc., 1995.

Levine, David M., Timothy C. Krehbiel and Mark L. Berenson: *Business Statistics: A First Course*, Delhi Pearson Education, 1st Indian Reprint, 2004.

Mann, Prem S.: *Statistics for Business and Economics*, New York, John Wiley & Sons, Inc., 1995.

Morris, Clare: *Quantitative Approaches in Business Studies*, London, Pitman, 1996 (Fourth Edition).

Oakland, John S.: *Statistical Process Control*, Oxford, Butterworth-Heinemann, 1999 (Fourth Edition).

Oakland, John S.: *Total Quality Management*, Oxford, Butterworth-Heinemann, 1993 (Second Edition).

Parsons, Robert: *Statistics for Decision Makers*, London, Harper & Row, Publishers, 1974.

Reichman, W.J.: *Use and Abuse of Statistics*, Baltimore, Penguin Books, Inc., 1975.

Siegel, Sidney: *Nonparametric Statistics for the Behavioral Sciences*, New York, McGraw-Hill Book Company, Inc., 1956.
Spiegel, Murray R.: *Theory and Problems of Statistics*, London, McGraw-Hill Book Company, 1992.
Spoomer, Ann and Colin Lewis: *An Introduction to Statistics for Managers*, London, Prentice Hall, 1995.
Yamane, T.: *Statistics: An Introductory Analysis*, New York, Harper & Row, Publishers, 1973.

INDEX