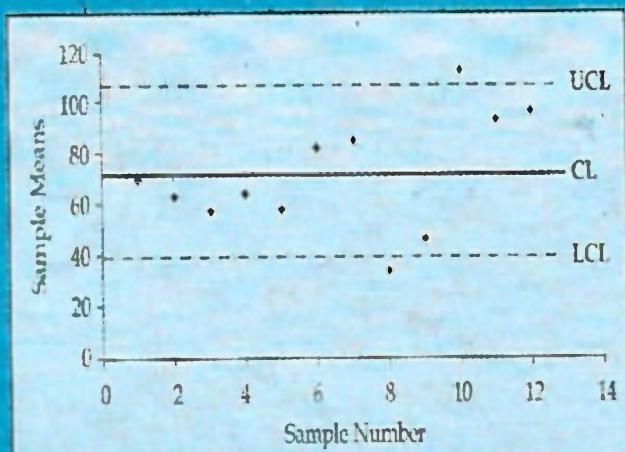
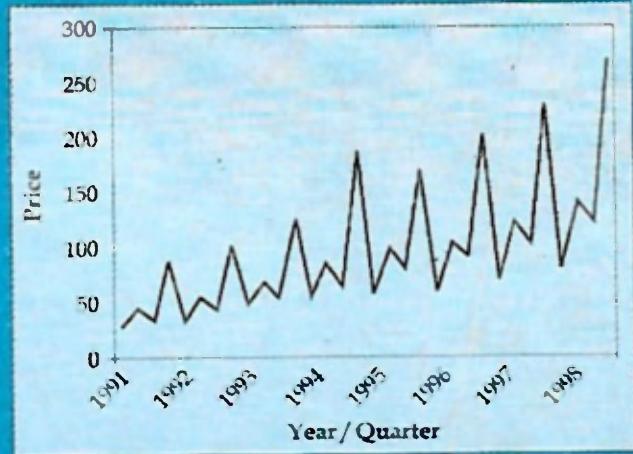


Business Statistics



Manindra Kumar Roy
&
Jiban Chandra Paul

Business Statistics

For BBA and MBA Students

Manindra Kumar Roy
B.Sc (Hons), M.Sc (Dhaka)
Ph.D (Kiev)
Professor of Statistics
University of Chittagong

and

Jiban Chandra Paul
B.Sc (Hons), M.Sc (Chittagong)
Ph.D (India)
Professor of Statistics
University of Chittagong

First Edition : November, 2012

Published by

Olga Roy

Chittagong University, South campus

Krishna Paul

Lalkhan Bazar, Chittagong

All rights reserved by the authors. No part of this
book may be reproduced in any form without the
written permission of the authors.

Distributors

Prime Book Depot

Jame-E-Mosjid Complex

Anderkilla, Chittagong

Phone: 031-2869258

Khan Book Centre

122, Islamia Market

Nilkhel, Dhaka

Phone: 02- 8612364

Cover Designed by

J.C. Paul

Computer Compose by

Mostafa Foysal Pasha

Printed by

JAHANGIR PRESS

Price : TK. 250 (News Print)

Price : TK. 300 (White Print)

**DEDICATED
TO
OUR PARENTS**

It is needless to say that now-a-days empirical research in any real life field can not be undertaken without application of statistical tools. Knowledge of Statistics is particularly very helpful for taking decision in the area of Business and Management under uncertainty. One has to perceive and conceive the proper interpretations of different concepts, and tools of Statistics in order to apply Statistics in the right manner. However, while teaching the Statistics related courses, such as 'Business Statistics', 'Statistics for Decision Making', 'Elementary Statistics for Business', etc. at different Private and Public universities or institutes, we have been observing that most of the students undertaking BBA or MBA programs, particularly, the students with Arts or Social Sciences background, cannot cope with the presentation and analytical approaches provided in the available text books on Business Statistics. Considering this fact, we intended to present the book 'Business Statistics' to our beloved students of BBA, MBA, CMA programs at private or public institutes.

The book is composed of eighteen chapters covering whole curriculum at BBA and MBA levels. It is designed to provide strong introductory understanding of applied statistical procedures so that the learners be able to handle a concrete statistical analysis in many business and economic problems. Keeping the matter of lack of higher mathematical backgrounds of our students in mind, the annoying tedious mathematical issues of statistics have been kept away in the book. On the other hand, emphasis has been given to the applications of statistical tools in real life problems. All the concepts of Statistics have been presented in a very simple way with illustrations, where applicable. The concepts have also been demonstrated with a number of examples. A substantial number of theoretical and applied exercises have also been provided at the end of each chapter with the belief that these would lead the learners to practical understanding of problems related to business and economics.

This book is the output of our combined efforts for about four years. We had to take help from different books, journals and websites. We are indebted to the authors of these books and journals. We are also thankful to our colleagues Professor Md. Emdadul Haque and Mr. Kanak Chowdhury, who co-operated us with inspirations and suggestions. The authors are also thankful to Mst. Pervin Akter, Assistant Registrar, Department of Statistics, for her co-operation during printing of the draft of this book. Above all, we had to deprive our beloved family members in many ways while preparing the manuscript of this book. We are also thankful to them for sacrificing their time and providing with active support in this effort without which it would have been difficult for us to complete this task.

Our efforts will be a success if the students and teachers of Business Statistics find the book useful in their learning and teaching process. We have no hesitation to mention that there might have some lacking in the book that we expect to improve in the next edition. So, any kind of suggestion and criticism for the enrichment and perfection of the book will be highly appreciated.

Chittagong
November, 2012

M. K. Roy
J. C. Paul

CONTENTS

Chapter 1. Introduction	
1.1 Origins and Historical Development of Statistics	1
1.2 Meaning of Statistics	1
1.3 Definition of Statistics	2
1.4 Business Statistics	4
1.5 Uses of Statistics	5
1.6 Limitation of Statistics	5
Questions	6
	7
Chapter 2. Some Important Concepts	8
2.1. Introduction	8
2.2. Types of Variables	11
2.3. Scales of Measurement	14
2.4. Types of Data	18
2.5. Descriptive and Inferential Statistics	22
Questions	24
Applications	25
Chapter 3. Methods of Data Collection	27
3.1. Introduction	27
3.2. Methods of Collecting Primary Data	28
3.3. Sources of Collecting Secondary Data	34
3.4. Selection of Appropriate Method	35
3.5. Processing of Data	35
Sample questionnaire	36
Questions	38
Chapter 4. Presentation of Data	39
4.1. Introduction	39
4.2. Condensing and Summarizing Data	39
4.3. Diagrams for Categorical Data	45
4.4. Pie Diagram or Chart	50
4.5. The Pareto Diagram	53
4.6. Pictogram	56
4.7. Condensing and Summarizing Quantitative Data	57
4.8. Construction of a Frequency Distribution for Continuous Data	59

4.9. Stem and Leaf Display	67
4.10. Graphical Representation of Quantitative Data	71
4.11. Histogram and Stem-leaf Display	77
4.12. Frequency Polygon	77
4.13. Frequency Curve	81
4.14. Cumulative Frequency polygon, cumulative frequency curve or Ogive	84
4.15. Line Graph	88
4.16. Some Additional Examples	89
4.17. Cross Tabulation of Statistical Data	95
4.18. Scatter Diagram	96
• Exercise	97
Application	98
Chapter 5. Describing Data with Numerical Measures	101
5.1. Introduction	101
5.2. Measures of Location or Central Tendency	101
5.3. Arithmetic Mean for Ungrouped Data	102
5.4. Median	116
5.5. Mode	129
5.6. Some Other Positional Measures	135
5.7. Geometric Mean	148
5.8. Harmonic Mean	152
5.9. Some more Measures of Central Tendency	155
Exercise	158
Application	159
Chapter 6. Measures of Dispersion	162
6.1 Introduction	162
6.2 Purposes of Dispersion	162
6.3. Properties of a Good Measures of Dispersion	163
6.4 Measures of Dispersion	163
6.5 Range	164
6.6 Inter - Quartile Range and Quartile Deviation	165
6.7 Mean Deviation	168
6.8 Variance and Standard Deviation	173
6.9 Coefficient of Variation	182
6.10 Measure of Relative Standing	190
6.11 Some Elementary Theorem and Examples	192
Questions	194

Exercise	194
Application	195
Chapter 7. Moments of a Distribution	200
7.1. Introduction	200
7.2. Shape Characteristics of a Distribution	204
7.3. Exploratory Data Analysis	212
Questions	217
Exercise	217
Applications	218
Chapter 8. Introduction to Probability	220
8.1. Introduction	220
8.2. Set Theory	220
8.3. Addition Laws of Sets	224
8.4. Venn Diagram	225
8.5. Tree Diagram	225
8.6. Concepts Related to Probability	228
8.7. Definition of Probability	234
8.8. Joint Probability and Marginal Probability	240
8.9. Conditional Probability	241
8.10. Independent Events	242
8.11. Laws of Probability	247
8.12. Bayes Theorem	254
8.13. Some Solved Problems on Probability	257
Questions	270
Exercise	270
Applications	271
Chapter 9. Random Variable	274
9.1. Introduction	274
9.2. Random variable	274
9.3. Discrete Random Variable	274
9.4. Continuous Random Variable	279
9.5. Mean and Variance of a Random Variable	281
Questions	287
Exercise	287
Applications	288

Chapter 10. Probability Distributions	290
10.1. Introduction	290
10.2. Bernoulli trial	291
10.3. Binomial distribution	280
10.4. Poisson Distribution	297
10.5. Normal Distribution	301
Questions	308
Exercise	309
Application	309
Chapter 11. Simple Correlation	312
11.1. Introduction	312
11.2. Correlation analysis	312
11.3. Simple Correlation	314
11.4. Assumption Underlying Karl Pearson's Correlation Coefficient or Simple Correlation Coefficient	316
11.5. Some Important Properties of Correlation Coefficient	316
11.6. Scatter Diagram	319
11.7. Probable Error of Correlation Coefficient	325
11.8. Rank Correlation	332
Questions	344
Exercises	345
Applications	345
Chapter 12. Simple Regression Analysis	348
12.1. Introduction	348
12.2. Population Regression Line and Model	350
12.3. Sample Regression Equation and Model	351
12.4. Relationship Between Correlation Coefficient and Regression Coefficients	354
12.5. Some Important Properties of Regression Coefficient	355
12.6. Difference between Simple Correlation and Simple Regression	356
12.7. The Coefficient of Determination r^2	357
12.8. Some Examples	357
Questions	369
Exercise	369
Application	370

Chapter 13. Index Numbers	373
13.1. Introduction	373
13.2. Some Characteristics of Index Numbers	374
13.3. Uses Index Number	375
13.4. Problems in the Construction of Index Numbers	375
13.5. Classification of Index Numbers	378
13.6. Features of Index Numbers	379
13.7. Methods of Constructing Index Number	380
13.8. Test of Accuracy of Index Number	399
13.9. The Chain Index Numbers – Change in Base Period	404
13.10. Base Shifting, Splicing and Deflating the Index Numbers	409
13.11. Consumer Price Index Number (Cost of Living Index Number)	415
Questions	420
Exercises	421
Applications	426
Chapter 14. Time Series Analysis and Forecasting	431
14.1. Introduction	431
14.2. Objectives of Time Series Analysis	432
14.3. Importance of Time Series Analysis in Business Decision-Making	432
14.4. Components of a Time Series	432
14.5. Mathematical Models for Time - Series Analysis	432
14.6. Description of Time Series Components	434
14.7. Measurement of Trend Component	439
14.8. Non-Linear Trends	454
14.9. Measurement of Seasonal Variation	467
14.10. Measurement of Cyclical Component	487
14.11. Forecasting	487
14.12. Smoothing Techniques	491
14.13. Autocorrelation Co-efficient	498
Question	500
Applications	501
Chapter 15. Sampling, Sampling distributions and Estimation	508
15.1. Introduction	508
15.2. Complete Enumeration or Census Method	509
15.3. Sampling	511

15.4. Methods of Sampling	514
15.5. Sampling and Non-sampling Errors	524
15.6. Sampling Distribution	526
15.7. Concept of Standard Error	527
15.8. Central Limit Theorem	528
15.9. Some Important Sampling Distributions	529
15.10. Concept of Estimation	549
15.11. Determination of Sample Size	558
Questions	562
Exercise	564
Applications	566
Chapter 16. Tests of Hypothesis	569
16.1. Introduction	569
16.2. Concepts of Hypothesis Testing	569
16.3. Survey of Important Test Statistics	579
16.4. Steps in Hypothesis Testing	581
16.5. Applications of Test Statistics	583
16.6. Hypothesis Testing for Single Population Mean	584
16.7. Test of Hypothesis Concerning Two Population Means	603
16.8. Test of Hypothesis Concerning Attributes	616
16.9. Test of Hypothesis about Correlation Co-efficient	625
16.10. Test of Hypothesis about Regression Co-efficient	627
16.11. Test of Significance of Single Variance (χ^2 Parametric Test)	638
16.12. Test of Hypothesis about Independence of Two Attributes (χ^2 Non-parametric Test)	640
16.13. Power of a Test	645
Questions	648
Exercises	650
Applications	651
Chapter 17. Statistical Quality Control	662
17.1. Introduction	662
17.2. Objectives of Quality Control	663
17.3. Causes of Variations	663
17.4. Uses of Statistical Quality Control (SQC)	664
17.5. Techniques of Statistical Quality Control	665
17.6. Process and Product Control	667

17.7. Control Chart for Variables	670
17.8. Control Chart for Attributes	683
17.9. Product Control	692
17.10. Types of Sampling Plan	695
17.11. Factors Related to An Acceptance Sampling Plan	696
Questions	702
Exercises	704
Applications	705
Chapter 18. Interpolation and Extrapolation	710
18.1. Introduction	710
18.2. Methods of Interpolation	711
18.3. Methods Based on Calculus of Finite Difference	714
Questions	725
Applications	725
Appendix. Statistical Tables	727
1. Random Numbers	727
2. Cumulative Binomial Probabilities	728
3. Cumulative Poisson Probabilities	733
4. Areas under the standard Normal Distribution	735
5. Critical Values of t	737
6. Critical Values of Chi-Square	738
7. Percentage Points of the F Distribution	740
8. Factors Useful in the Construction of Control Charts	747
References	748

CHAPTER - 1

INTRODUCTION

1.1. Origins and Historical Development of Statistics

Statistics as subject is barely a century old. In ancient time the word statistics was used as the collection of data for the purpose of taxation and military conscription. In this sense, statistics is as old as the human society itself.

The word 'statistics' seems to have been derived from the Latin word 'status' or the Italian word 'statista (meaning statesman)' or the German word 'statistik' or the French word 'statistique', each of which means a 'political state'. History tells us that in ancient times, the Pharaohs, the Hebrews, the Chinese, the Roman and the Greek rulers used to collect information about population, land, wealth, total number of employees, soldiers etc. to have the idea of the manpower of the country for formulation of administrative setup, fiscal, new taxes, levis and military policies of the government.

In Indian subcontinent, an efficient system of collecting official and administrative statistics existed even more than 2000 years ago. From Kautilya's Arthashastra it is known that even before 300 B.C., during the reign of Chandra Gupta Maurya (324-300 B.C.), a very good system of collecting 'vital statistics' and registration of births and death was in vogue. In 'Ain-I-Akbari' written by Abul Fazl (1596-97), we find that during Akbar's reign (1556-1605 A.D.) Raja Todarmal, the land and revenue minister maintained a very good record of land and agriculture statistics.

To formulate the commercial and industrial policy of state, Colbert of France introduced a Politico-Economic policy known as Mercantilism in the fifteenth century. Through this policy, the economic activities of the individuals are brought under the control of the state. This necessitated the elaborate and systematic collection of statistical data on trade, commerce and industry.

Birth, death, divorce, crime etc. vital statistics were systematically collected in England and studied by some mathematician, notably John Graunt (1620-1674) , Edmund Halley and William Petty (1623-1687) in the middle of the seventeenth century. John Graunt intensively worked with the vital statistics and came to a conclusion that the population of a country could be estimated from the birth and death statistics. He is known as the father of vital statistics for his fundamental works. Edmund Halley first constructed

a complete life table and made remarkable development in the field of life insurance. Their works were published in the name of Political arithmetic.

Information regarding population, industrial and agriculture statistics were collected in Germany at the end of the 18th century, in order to have an idea of the relative strength of different German state. In Germany statistical works were done in the name of 'Universal statistics'. A German professor Gottfried Ackenwall (1719-1772) in 1749 gave the first definition of statistics as a subject. He defined statistics as 'the political science of several countries'. Another definition of statistics was found in 1770 by Baron in the famous book "Elements of Universal Erudition". Here statistics was defined as "the science that teaches us what is the political arrangement of all the modern states of the known world". At that time statistics served as an overall picture of a country. That is why it was also known as the' Science of Kings'.

Today statistics as subject, originated from two quite dissimilar field viz. Political arithmetic or Science of kings and the theory of probability.

The game of chance is associated with the theory of probability developed mainly by the famous gamblers and mathematicians of Italy, France and Germany like Pascal (1623 - 1662), Fermat (1601 - 1665), Chaveliar-de-Mere, J. Bernoulli, De-Moivre (1667 - 17540), Laplace (1759 - 1827), Gauss (1777 - 1855) from the mid of the seventeenth century. But the notable development of statistics was started from the early of nineteenth century by the English Statistician Galton (1822 - 1921), K. Pearson (1857 - 1936), W.S. Gosset (1876 - 1937), R.A. Fisher (1890 - 1962) and others. Statistics got recognition as an independent discipline from the first quarter of the twentieth century. In 1911, Karl Pearson Founded the First University Statistics Department at University College, London. He is the founder of the first Statistical Journal "Biometrika" in 1901. Although the first stone of statistics was laid down by K. Pearson but R.A. Fisher, known as the father of the modern statistics, placed statistics on a very sound footing by applying it to various diversified field such as genetics biometry, education, agriculture etc. Today, statistics appears as an important branch of scientific knowledge.

1.2. Meaning of Statistics

The word Statistics is used in three different senses:

- 1) Statistics as a singular,
- 2) Statistics as a plural,
- 3) Statistics as a plural of statistic.

1.2.1. Statistics as a singular. Like physics, chemistry, mathematics, statistics as singular is considered as a separate scientific discipline.

Example 1.2.1.

- (i) Dr. M. Rahman is a professor of statistics,
- (ii) Mr. Zaman is a student of statistics,
- (iii) An introduction to Statistics is a book of statistics.

In these three examples, the word statistics is used as a singular. As a subject statistics has been defined as:

Definition 1.1. Statistics is that branch of knowledge that deals with the collection, organization, classification, presentation, summarization, analysis, and interpretation of statistical data in any field of inquiry.

1.2.2. Statistics as a plural. By statistics, we mean a set of numerical data relating to any field of inquiry. In early stage, the word statistics was used in this sense.

Example 1.2.2.

- (i) Statistics of daily production of a factory;
- (ii) Import-export statistics of Bangladesh for the year 2005-2006;
- (iii) Statistics of daily sales of a store etc. are some examples of statistics as plural.

An important definition of statistics as plural was given by Sechrist as :

Definition 1.2. By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a pre-determined purpose, and places in relation to each other.

Here any set of data or information may be referred to as statistics.

Example 1.2.3. The daily rainfall statistics for the last one-week is given in the following table:

Day	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Rainfall in mm	50	48	52	49	42	74	60

1.2.3. Statistics as plural of statistic. Any numerical value describing a characteristic of a sample is called a statistic.

Example 1.2.4.

- (i) Sample mean \bar{x} ,
- (ii) Sample variance s^2 etc.

If only one such measure is obtained, it is called statistic in the singular form.

Example 1.2.5. The daily sales of a store for last five days are Tk. 5000.00, 6000.00, 5500.00, 6500.00, 7000.00. The average sale of the store for last five day is Tk.6000.00. This average is calculated from a sample and it is called statistic.

Any function of a random sample is also called a statistic. Its value is used as an estimate of population parameter.

1.3. Definition of Statistics

It is very difficult to define a fast growing subject like statistics. In 1935 W.F. Willcox listed over a hundred definitions of statistics and the list was even then far from being exhaustive. Some definitions are old and narrow while others are modern and more comprehensive. Here we shall cite some old and modern definition of statistics.

As mentioned before, first definition of statistics was given a German professor Gottfried Ackenwall in 1749 and a second definition was given by Biron in 1770. These two definitions are old and too narrow. Here statistics is restricted to political state only. Today the subject statistics originated from two quite dissimilar field viz. political state and probability theory.

According to Professor R.A. Fisher, "The science of statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data."

Here statistics is considered as a branch of applied mathematics.

Croxton and Cowden gave a more comprehensive definition of statistics.

According to them Statistics may be defined as "the science, which deals with the collection, presentation, analysis and interpretation of numerical data.

This definition clearly points out four stages in a statistical investigation but excludes the organization of data and the inferential statistics. Moreover, statistics is considered here as a branch of science only.

In recent times, modern statistical theories and techniques have been developed in decision-making in the face of uncertainty in any field of enquiry. So a modern and more comprehensive definition of statistics may be given by "Statistics is a branch of scientific knowledge refers to the body of techniques and methodology developed for the collection, classification; organization, presentation and analysis of statistical data

and for the use of such data in decision-making in the face of uncertainty in any field of enquiry".

Now a day's statistics is not only a branch of science; its different methods and techniques have been successfully applied in almost all branches of knowledge. As a result, a number of new disciplines such as Econometrics, Tachometry, Biometrics, Biostatistics, Psychometrics, Sociometry etc. have been developed from statistics. Today Statistics collects, classifies, organizes, analyses data and takes decisions in the faces of uncertainty in any field of enquiries. So the definition of statistics cited above can be taken as an acceptable and exhaustive one. Actually, different statistical methods and techniques are the ways of knowing new facts in any field of knowledge.

1.4. Business Statistics

This is an elementary book on business statistics. The book is designed to serve as a textbook for the basic course in statistics taught in different departments of business school. Now we shall define business statistics. Business statistics may be defined as

"Statistical methods and techniques used to analyze data collected from various fields of business and management is called business statistics".

Business and management related data are also called business statistics in the plural sense.

The main purpose of this book is to familiar the business students with the most commonly used statistical methods of analysis and enable them to use business data effectively in making decision.

1.5. Uses of Statistics

The uses of statistics are unlimited. It is much harder to name a field in which statistics is not being used. Today, statistical tools are used in every spheres of life such as trade, industry and commerce, economics, biology, botany, astronomy, physics, chemistry, education, medicine, sociology, public administration, psychology, meteorology, agriculture etc. Here we shall discuss uses of statistics only in business.

1.5.1. Statistics in Business. Statistics is very important in business and commerce. The data collected from the field of business and commerce are called business statistics. History tells us that in order to control the business and economic activities of individual the west, European countries used to collect the business statistics during the period of mercantilism in the fifteenth century by state. Today, in the era of free market economy, business enterprises are big and competitions are high. The domain of

business is enshrouded by risks and uncertainties. Management has become a specialized job. A business manager has to plan, organize, supervise and control the operation of the enterprise. A successful manager must use the past and present statistical records and capable of making business decision, which can reduce the risk and uncertainty to minimum level. Different statistical methods are very important tools for proper organization in business and commerce. The important statistical techniques which are used in business and commerce are descriptive statistics such as different types of diagrams and graphs, different measures of central tendency and dispersion, correlation and regression analysis, time series analysis and business forecasting , index numbers and statistical inference etc. Statistical tools like control chart, sampling inspection plan etc. are being enormously used in controlling the quality of the product in industry by the modern management. Statistics is so important in business that many business organizations maintain their own statistical section. In the words of Ya-Lun-Chou, "In business, statistics has already made radical changes in maintaining and improving output in selecting and promoting personal, in efficient use of materials in projecting long-term capital requirements and forecasting sales, in estimating consumer's preferences, and in various other phases of business research and management. It is not an exaggeration to say that today nearly every decision in business is made with the aid of statistical data and statistical methods".

1.5.2. Function of statistics. The following are the important functions of statistics:

- a) It presents facts in a definite form;
- b) It implies mass of data;
- c) It facilitates comparison;
- d) It helps in formulation and test of hypothesis;
- e) It helps in prediction; and
- f) It helps in the formation of policies.

1.6. Limitation of Statistics

Although statistics has wide applications in every sphere of human knowledge, it is not free from limitations. The following are some of its important limitations:

1. Statistics does not deal with individuals. It only deals with aggregative values. For example, the production of a factory for a particular day, sale of a store for a particular day, national income of a country for a particular year are meaningless unless, they are compared with other similar figures.

2. Statistics cannot deal with qualitative characteristics. Statistics are numerical statements of facts. Thus qualitative characteristics such as honesty, poverty, intelligence, efficiency, blindness, deafness, culture, etc., cannot be studied directly. However, it may be possible to analyze such problems statistically by expressing them numerically. For example, the score obtained in a test can be used to study the efficiency of an employee.

3. Statistical results are true on an average. Statistical conclusions are not universally true; they are true only under certain conditions. This is because statistics as a science is less exact as compared to natural sciences like physics and chemistry.

4. Statistics can be misused. The greatest limitation of statistics is that it is liable to be misused. As the saying goes "Statistical methods are the most dangerous tools in the hands of the inexpert". The use of statistical tools by inexperienced and untrained persons might lead to fallacious conclusions. As King says, "Statistics are like clay of which one can make a God or Devil as one pleases". So it requires statistical skill and experience to draw sensible conclusion from the data; otherwise there is every chance of making wrong interpretation.

In conclusion we want to cite two important remarks made by two Statisticians. According Bowley "Statistics only furnishes a tool, necessary though imperfect, which is dangerous in the hands of those who do not know its use and its deficiencies". It is not the subject of statistics that is to be blamed but those people who twist the numerical data and misuse them either due to ignorance or deliberately for personal interest. As King points out. "Science of Statistics is the most useful servant but only of great value to those who understand its proper use".

Questions

1. What is the meaning of the word Statistics? Discuss with suitable examples.
2. Briefly discuss origin and historical development of Statistics.
3. Define Statistics as a subject and cite some of its limitation.
4. What is Statistics? Bring out clearly the importance of statistics in business.
5. What do you mean by business statistics?
6. Define Statistics. Discuss the main function and limitations of statistics.

CHAPTER - 2

SOME IMPORTANT CONCEPTS

2.1. Introduction

In this chapter we shall first define some important basic concepts, which are needed to study and understand the subject statistics.

2.1.1. Population. Statistical methods are particularly useful for studying, analyzing and learning about population. Literally, population means total inhabitants of a country. But in Statistics it has a wide meaning.

Definition 2.1. Population is the totality or collection of all objects or individuals on which observations are taken on the basis of some characteristic of the objects in any field of enquiry.

Actually, it is the aggregate of individuals possessing some characteristic in common.

Example 2.1.1. The population may be

- i) All workers of a factory;
- ii) All employees of a firm;
- iii) All students of Chittagong University; etc.

Definition 2.2. Each individual of a population is called an experimental unit. Observations are collected on experimental units.

Example 2.1.2. Workers, employees and students are the experimental units of the above populations.

Example 2.1.3. An experimental unit may be

- i) An employee of a firm,
- ii) A student of a class,
- iii) A cow of a firm
- iv) A patient of a clinic
- v) A plot of an agriculture land etc.

The bold words of the above examples denote the experimental units.

Population may be finite and infinite.

Definition 2.3. Finite Population. A population is called finite if it contains finite number of experimental units. All the three examples cited in example 2.1.1 are finite populations.

Definition 2.4. Infinite population. A population is called infinite if it contains infinite number of experimental units.

Example 2.1.4. (i) In a coin tossing experiment, number of tosses required to get a head, (ii) The length of life of a bulb etc. are examples of infinite population.

2.1.2. Variable. It is a very important concept in statistics.

Definition 2.5. A Variable is a changeable characteristic of the experimental units under consideration. Actually, it is the characteristic of experimental units which varies from unit to unit.

It is customary to represent variables by the last capital letters of the English alphabets. That means, the variables are generally denoted by X, Y, Z, U, V, W etc.

Example 2.1.5.

- i) Age of a worker,
- ii) Religion of a student,
- iii) Wage of a worker,
- iv) Gender of a garment worker,
- v) Height of a student, etc. are some examples of variables.

The bold words of the above examples denote the variables. Actually population is named according to the characteristic of the experimental unit. For example, the ages of all workers of a factory, incomes of all workers of a factory are the population of age and the population of income respectively.

2.1.3. Observation, measurement or datum. An observation or measurement is obtained when a characteristic is measured on an experimental unit. Or when a variable is measured on experimental unit we get an observation or measurement. A single observation is called datum (Data in plural).

Definition 2.6. Data. A set of observations obtained from a particular enquiry is called data or a data set.

Usually data are the numerical results of scientific measurements. For example, it could be the

- i) Income of workers,
- ii) IQ-scores of students,
- iii) Examination marks of students in a class, etc.

Actually, data are the raw and disorganized facts and figures in any field of enquiry.

Definition 2.7. Sample. A sample is a part of a population that is taken and considered for study.

Usually, sample is a small but representative part of a population which contains a finite number of observations.

Some examples of sample are:

- i) Some workers of a factory,
- ii) Some employees of a firm,
- iii) Some students in a collage,
- iv) Some cows of a diary firm,
- v) Some trees of a forest,
- vi) Sale of stores for some days of a month etc.

Sample should represent the population characteristics under study. So selection of a representative sample is very important to take decision on whole population.

2.1.4. Sampling. Sampling is the process of selecting a sample from the population. Most of the times it is not feasible technically and economically to take entire population for analysis, so we must take a representative part of the population as a sample for the purpose of such analysis. Simple random sampling technique is an important procedure for selecting a random sample.

Definition 2.8. Simple random sample. A sample is called simple random sample if every elements of the population has an equal chance of being included in the sample.

Definition 2.9. Parameter. Any numerical value describing a characteristic of a population is called a parameter.

It is customary to represent parameters by Greek letters. By tradition the arithmetic mean of a population is denoted by the Greek letter μ (mu). Similarly, population variance (σ^2), correlation coefficient (ρ), regression coefficient (β), proportion (π) etc. are the examples of parameter. Note that a parameter is a constant value describing the population characteristic.

Definition 2.10. Statistic. Any numerical value describing a characteristic of a sample is called a statistic.

A statistic is usually represented by a small letter of the English alphabet. If the statistic is the sample arithmetic mean, it is denoted by \bar{x} . The sample variance (s^2), correlation coefficient (r), regression coefficient (b), sample proportion (p), etc. are the examples of some statistics. This means any summary value calculated from the sample is called statistic. Usually these values are used to estimate the corresponding population parameters.

The concepts of all the terms discussed so far are illustrated below with examples.

Example 2.1.6. Suppose we want to find the average sales of a certain commodity sold in 60 shops in Chittagong Metropolitan area. Then these 60 shops will be our population of interest. It is a finite population. Each shop of this area will be our experimental unit. The characteristic of interest is the amount of sales.

Example 2.1.7. Suppose there are 80 students in your class. We want to find the average height of these 80 students. Then these 80 students will be our population of interest. It is a finite population. Each student of this class will be our experimental unit. The characteristic of interest is height. Here height is the variable. If you collect numerical information on the height of all the students, then the collection of heights of 80 students will be the population data or the population of height. The average heights of these 80 students say 5.7 feet. Then $\mu = 5.7$ feet is our parameter, since it is a characteristic of the population.

Suppose it is not possible to get the population data. In that case, we shall take a random sample of 10 students to estimate the average height of the class. Then the 10 students will constitute the sample and the collection of the heights of 10 students will be the sample data or sample. Suppose the average height of these 10 students is 5.6 feet. Then the sample arithmetic mean $\bar{x} = 5.6$ feet is a statistic and value is used as an estimate of the population mean μ .

Size of a population. The size of a population is the number of observations or experimental units in it. It is usually denoted by N . Here the size of the population is $N = 80$ which is the total number of students in the class.

Size of a sample. The size of the sample is the number of observations or experimental units in it. It is denoted by n . Here $n = 10$ which is the number of students in the sample.

2.2. Types of Variables

According to whether a variable takes numerical or non-numerical values, it can be classified into two categories, viz.

- i) Qualitative variable and
- ii) Quantitative variable.

Definition 2.11. Qualitative variable. A variable is called qualitative when it measures qualitative characteristic on each experimental unit.

Qualitative variable cannot be measured on a natural numerical scale.

Characteristic of a qualitative variable are known as attribute. Qualitative variables produce qualitative data that can be classified according to different categories; hence they are often called categorical data. Here are some examples of qualitative variable:

- i) Religion of a student
- ii) Gender of a patient
- iii) Economic status of a person
- iv) Teaching performance of a professor
- v) Efficiency of a worker
- vi) Colour of a car entering the parking lot
- vii) Hair colour of a student
- viii) Quality of a finished product
- ix) Size of an industry

The bold words of the above examples are the qualitative variables..

Definition 2.12. Quantitative variable. A variable is called quantitative when it measures a numerical quantity or amount on each experimental unit.

Quantitative variables are usually denoted by the last capital English alphabets such as X, Y, Z, U, V, W etc. Some examples of quantitative variables are cited now .

1. X: Number of children per family
2. X: Piece of shirts produced daily by a garment factory
3. X: Production of rod in tons produced daily by steel mill.
4. X: Production of Sugar in kg produced daily in a sugar mill
5. X: Daily rainfall in inches in Chittagong city during the rainy season
6. X: Daily wage of workers of a factory,
7. X: Number of children per family of factory workers
8. X: Systolic blood pressure of a patient etc.

Note that there are differences in the types of numerical values that the quantitative variables assume. The number of children per family, for example can take on only the values $X = 0, 1, 2, 3\dots$ where as the daily rainfall in inches can take on any value greater than zero or less than a finite quantity, that means $0 < X < b$, where b is a positive quantity. It is to be noted here, that the number of children is a countable quantity, while the daily rainfall is a measurable quantity.

Hence, on the basis of whether a variable is countable or measurable, it is again classified as

- i) Discrete variable and
- ii) Continuous variable.

Definition 2.13. Discrete variable. A variable, which can take, only isolated or countable finite or infinite number of values is called a discrete variable.

Usually discrete variable takes natural number. Sometimes, it can also take countable number of fractional values. The following are some examples of discrete variables:

1. Number of defective items in a lot of 10 items,
2. Number of children per family,
3. Number of accidents per day in a busy corner of a road,
4. Number of printing mistakes per page of a book, etc.

Here, number of defective items, number of children per family, number of accidents per day and number of printing mistakes can take only integer values. Now we cite some examples where discrete variable can take fractional values. The discrete variables are underlined.

1. Size of shoes sold in a shop may be 3, 3.5, 4, 4.5 etc.
2. Size of nails in inches available in a shop may be 0.25, 0.50, 1.00, 1.25.
3. Coins in taka of a cash box may be 0.01, 0.05, 0.10, 0.25, 0.50, 1.00, 2.00, 5.00 etc.

Here the size of shoes, size of nails and coins in cash box take fractional but isolated values. So they are discrete variables. The most important characteristics of discrete variables is that they are countable.

Definition 2.14. Continuous variable. A variable, which can take infinitely many values in a certain range is called a continuous variable.

The values of a continuous variable cannot be counted, as they cannot take any isolated value. That is continuous variable can be measured only.

Some examples of continuous variables are:

- i) Age of a worker,
- ii) Systolic blood pressure of a patient,
- iii) Weight of an employee,
- iv) Weight of a package ready to be shipped,
- v) Height of a salesman,
- vi) Monthly salary of a worker etc. Here the bold words are the quantitative variables.

Examples 2.2.1. Identify each of the following underlined variables as qualitative or quantitative:

1. The number of unregistered taxicabs in a city.
2. The number of consumers who refuse to answer a telephone survey.
3. The winning time for a horse running in a race.
4. Gender of an employee of a garment factory.

5. Ethnic origin of a candidate for a public office.
6. Brands of soft drinks sold in a café

Solution. Variables in examples 1, 2, 3 are quantitative and discrete. In example 1, the number of unregistered taxicabs is a discrete variable that can take on any of the value $X = 0, 1, 2, \dots$ with a maximum value depending on the number of unregistered taxicabs. Similarly, in example 2, the number of consumers is a discrete variable that can take on any of the values $X = 0, 1, 2, \dots$ with a maximum value depending on the number of consumers called. In example 3, the winning time is the only quantitative and continuous variable in the test. Variables in examples 4, 5 and 6 are qualitative because qualitative response would be obtained for these variables.

Example 2.2.2. A medical research wants to estimate the survival time in years of a patient after the onset of a particular type of cancer and after a particular regime of radio therapy. A sample of 50 patients having cancer and radiotherapy who are not alive have been selected randomly from a cancer hospital.

- i) What is the population?
- ii) What is the sample?
- iii) What is the experimental unit?
- iv) What is the variable to be measured?
- v) Is the variable qualitative, or discrete or continuous?

Solution.

- i) The population of interest is the set of all patients listed in the registrar of cancer hospital having that particular type of cancer who died after undergoing the particular type of radiotherapy.
- ii) The 50 patients selected at random from the cancer hospital is the sample.
- iii) Every cancer patient who died having undergone the particular type of radiotherapy is an experimental unit.
- iv) Survival times in years is the variable to be measured.
- v) The variable is quantitative and continuous.

2.3. Scales of Measurement

World civilization is enriched by the idea of number and measurement. It was first felt in physical sciences but now a days it is spread nearly all branches of knowledge. At the end of the eighteenth century Lord Kelvin said, "when you can measure what you are speaking about and express it in numbers you know something about it; but when you can not express it in numbers, your knowledge is of a meager and unsatisfactory kind".

It is the belief of some researchers that if it is researchable then it must be measurable. If the measurement procedure in a statistical investigation is poor, the usefulness of the findings of the investigation will be severely affected. For every researcher, it becomes necessary to explain the variables under study as well as the level of measurements of the selected variables during the planning phase of the study. In the field of measurement it is also essential to know what the numbers actually imply. Now we want to explain the meaning of measurement.

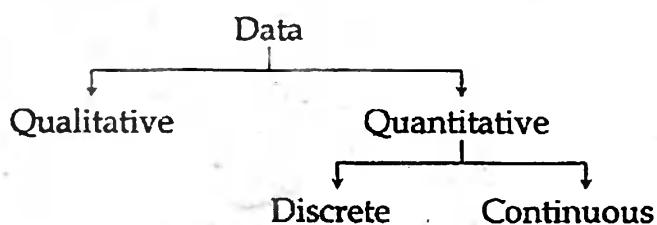
Definition 2.15. Measurement. It is a process of assigning numbers to some characteristics or variables or events according to scientific rules.

In statistics, measurement has a special meaning. A measurement is obtained when a variable is measured on experimental unit. All variables under statistical study can be measured. Since variables in study may be of different nature, they may have different level or scale of measurement. According to Stevens (1968), there are four levels or scales of measurements. They are

- i) Nominal scale;
- ii) Ordinal scale,
- iii) Interval scale and
- iv) Ratio scale.

Each scale is a way of assigning numbers to variables, and each has its limitation and degree of precision. According to different scale of measurement, variables are also called nominal, ordinal, interval and ratio.

The classification of the variables and measurement of scales is shown in following diagram



Qualitative variables are measured by nominal and ordinal scales where as quantitative variables are measured by interval and ratio scales. Now we shall define all the scales of measurement in turn with examples.

Definition 2.16. Nominal Scale. The scale of measurement by which we can classify and identify a qualitative variable according to different categories is called nominal scale.

The main purpose of this scale of measurement is to identify and then classify qualitative variables according to different categories. The variables

measured by nominal scale are also called nominal variables and the data obtained by this scale of measurement are called nominal data. For example, the defective status (defective or not defective) of each of 100 piece of cloths produced by a factory.

Here are some examples of nominal scale:

1. **Gender** of a garment worker (Male, female),
2. **Colour of eyes** of a worker (Black, green, brown),
3. **Religion** of a worker (Muslim, Hindu, Buddhist, Christian)
4. **Marital status** of a worker (Single, married, widowed, divorced or separate).

The variables of the above example are bolded and are measured by nominal scale.

In example 1, a student is measured according to gender. He may be male or female. That is worker is first identified and the classified according to two categories. Similarly, a worker can be identified and classified according to the colour of eyes or religion or marital status.

It is the weakest level of measurement. These measurements are not measured natural numerical scale. But sometimes, they are denoted by number or alphabet or both. But this number cannot be used for mathematical operation such as addition, subtraction, multiplication or division. For example, the model or registration number of a car, the holding number of a house, class identity number of a student etc.

Definition 2.17. Ordinal Scale. The scale of measurement by which we can classify, identify and rank a qualitative variable according to different categories is called ordinal scale.

Ordinal scale is the second level of measurement. It possesses all the characteristics of the nominal scale. Some examples are given below:

1. Grading of a student (A, B, C, D).
2. Size of a worker (Tall, medium, short).
3. Size of a factory (Big, medium and small).
4. Rating of an Executive (Excellent, good, fair, poor).
5. Economic status of a citizen (Higher class, meddle class, poor).
6. Health status of a worker (Excellent, good, poor).

In example 1, students are classified into four categories according grade A, B, C, D. Here a student having letter grade A is better than a student having letter grade B. That is letter grade A, B, C, D are orderly classified. In example 2, the size of workers are measured as tall, medium or short. Here the values of the six qualitative variables are meaningfully ranked or ordered and then classified according to different categories. Here

mathematical notation $>$ (greater than) or $<$ (less than) may be useful. But addition, subtraction, multiplication or division is not possible in ordinal scale. The variables measured by this scale are also known as ordinal variables and the data obtained by this scale of measurement are called ordinal data. For example, efficiency (Excellent, good, fair, poor) of each 75 workers of a factory is an example of ordinal scale.

Definition 2.18. Interval Scale. The scale of measurement by which we can measure a quantitative variable numerically on experimental unit with arbitrary zero as origin is called interval scale.

This is the third highest level of measurement. It preserves all the characteristics of nominal and ordinal scales. Some examples of interval scale are :

1. Body temperature of a patient.
2. Marks obtained by students in an examination.
3. Calendar time.

The variables considered in the above three examples are quantitative. They are measured in numerical scale. Body temperature is measured in Fahrenheit or Centigrade. These two measurements have no absolute zero as origin. So zero body temperature does not mean 'no temperature', 60° F is not twice as hotter as 30° F. But the temperature difference between 60° F and 30° F is the same as 100° F and 70° F. Similarly, the marks obtained by a student have no absolute zero or origin. Mathematical operation such as addition and subtraction are only possible by this scale of measurement. But multiplication or division is not meaningful in interval scale. The variables, which are measured by this scale, are known as interval variables and the data obtained by this scale of measurement are called interval data. For example, the scores of a sample of 50 students in a class test.

Definition 2.19. Ratio Scale. The level of measurement is called ratio scale when a quantitative variable is measured numerically on experimental unit with absolute zero as origin.

Ratio scale is the highest level of measurement. All kinds of mathematical operations are possible in this scale. It preserves all characteristics of the previous three scales of measurements. Some examples of this scale of measurements are given below:

1. Age of a worker.
2. Weight of a worker.
3. Height of a worker.
4. Number of printing mistake per page of a book.
5. Number of children per family.
6. Number of defects of a product.

All the variables in the above examples are quantitative. They are measured in natural numerical scale. Age, weight, height, number of printing mistakes, number of children and number of defects are all measured not only in natural numerical scale, they have also absolute zero as origin. Zero age of a worker means no age, zero weight or zero height means no weight or no height.

The variables measured by this scale of measurement are called ratio variable and the data obtained by this scale are called ratio data. For example, the age for each of 50 workers of a factory..

Example 2.3.1. Identify the scale of measurement on which the following underlined variables are measured:

1. Colour of a car entering in the parking lot.
2. Model of a car made by the fort motor company.
3. Holding number of a house in Dhaka city.
4. Smoking habit of a student.
5. Types of houses for the University teachers (A, B, C, D).
6. Socio-economic status of a citizen (upper, upper-middle, lower-middle, lower).
7. Size of an apple in a basket (big, medium, and small).
8. Quality of each student (Best, average, poor) in a class of 100 students.
9. Body temperature of a patient.
10. The grade point average (GPA) obtained by each student of a class of 100 students.
11. Room temperature of a house in winter.
12. Number of children per family of 100 families selected from an area.
13. Printing mistakes per page of a book.
14. Height of a cricket player.
15. Weight of a soldier in the Bengal Regiment.
16. Political affiliation of a citizen.

Solution. 1. Nominal Scale, 2. Nominal scale, 3. Nominal scale, 4. Nominal scale, 5. Ordinal Scale, 6. Ordinal Scale, 7. Ordinal Scale, 8. Ordinal Scale, 9 Interval scale, 10. Interval scale, 11. Interval scale, 12. Ratio scale, 13. Ratio scale, 14. Ratio scale, 15. Ratio scale, 16. Nominal scale.

2.4. Types of Data

Statistics is the science of data and that data are obtained by measuring the values of one or more variables on the experimental unit in the population or sample. Broadly speaking data may come from many ways such as:

- 1) According to origin,

- 2) According to variable,
- 3) According to scale of measurement,
- 4) According to time,
- 5) According to sources.
- 6) According subjects

2.4.1. According to origin. Data may be obtained from

- i) Population and
- ii) Sample.

Definition 2.20. Population Data. When data originate from each experimental unit of the population on the basis of some variable is called population data. It is also called census data.

Following are the examples of population data:

1. Heights of 500 students of a department may be population data if these 500 students are the experimental units of the population.
2. Salary of all the workers of a factory,
3. Marks obtained by all students of a class in a test, etc.

Definition 2.21. Sample Data. Data are called sample data when they are obtained from each experimental unit of a sample.

The following are examples of sample data:

1. Salary of 50 workers out of 500 is a sample data,
2. Ages of 40 workers randomly selected from 500 workers of a garment.
3. The health condition of 50 workers of a factory randomly selected from 1000 workers.

2.4.2. According to variable. Again all data can be classified as one of two general types according to variable as

- 1) Qualitative data and
- 2) Quantitative data.

Definition 2.22. Qualitative data. Qualitative data are obtained when a qualitative characteristic is measured on each experimental unit.

Qualitative data cannot be measured on a natural numerical scale. They can only be classified into one of a group of categories.

Examples of qualitative data include:

1. The defective status (defective or non-defective) of each of 100 bulbs produced by a Philip's company,
2. Working Efficiency of 60 workers of a factory (Excellent, good, satisfactory, poor),

3. Gender (male, female) of 5000 workers of a garment factory,
4. The eyesight (excellent, good, poor) of a sample of 100 male students of Chittagong University.

Sometimes qualitative data are called categorical data or count data. Often, we assign arbitrary numerical values to qualitative data for the ease of computer entry and analysis. But these assigned numerical values are simply codes. They cannot be meaningfully added, subtracted, multiplied or divided. For example, we might code Defective = 1; Non-defective = 2. Similarly, efficiency of a worker might rank from excellent = 1 to poor = 4. These are simply arbitrary selected codes for the categories and have no use beyond that.

Definition 2.23. Quantitative data. Quantitative data are obtained when a quantitative variable is measured on each experimental unit.

Quantitative data are measured on a naturally occurring numerical scale. The following are the examples of quantitative data:

1. The daily production of a factory for last one month.
2. The monthly export of tea for the first six months of the financial year 2007-2008.
3. The body temperature of 60 patients of clinic.
4. The number of family members of 150 workers of a factory.
5. The GPA scores of 200 S.S.C. students of Chittagong University School for the year 2008.

Again there are two types of quantitative data. They are

- 1) Discrete data and
- 2) Continuous data.

Definition 2.24. Discrete data. Data generated from a discrete variable is called discrete data.

Examples.

- (1) Number of babies born per day for last six months of the year 2007 in Chittagong Medical College,
- (2) Numbers of workers absent per day for last two months of a garment industry,
- (3) Number of children per family of 150 families of a village.

Definition 2.25. Continuous data. Data originated from a continuous variable is called continuous data.

Examples.

- 1) Weight in pound of 50 students of a class,
- 2) Body temperature of 70 patients of a hospital,

- 3) Heights in feet of 100 first year students of Chittagong University,
- 4) Weight in kg of 100 finished product of a factory.

2.4.3. Data according to scale of measurement. According to different level of measurements data are also known as

- i) Nominal data,
- ii) Ordinal data,
- iii) Interval data, and
- iv) Ratio data.

(i) **Nominal data.** A set of data is called nominal data when it is originated from a variable which is measured on nominal scale. Some examples of nominal data:

1. The defective status (defective or not defective) of each of 100 garments items manufactured by a factory,
2. Colour of each of 100 cars entering in a parking lot,
3. Types of ice creams (cup, choc-bar, lemon, and orange) available in a cooling corner.

(ii) **Ordinal data.** A set of data is called ordinal data when it is originated from a variable which is measured on ordinal scale. The following are some of its examples:

1. The size of a car (subcompact, compact, mid-size or full-size) rented by each of a sample of 40 business travelers,
2. The letter grade (A, B, C, D) obtained by 100 student of a school in H.S.C. examination for the year 2008,
3. The size (large, medium and small) of each apple of 150 apples in a basket.

(iii) **Interval data.** A set of data is called interval data when the variable is measured on interval scale. The following are the examples of interval data:

1. The grade point average (GPA) obtained by a sample of 50 students got admission in BBA course for the session 2007-2008,
2. The temperature (in degree Celsius) at which each in a sample of 40 pieces of heat-resistant plastic begins to melt.

(iv) **Ratio data.** A set of data is called ratio data when a variable is measured on ratio scale. Some examples of ratio scale are:

1. The number of printing mistakes per page of a book of 800 pages,
2. The weight of each of 150 cartons ready to be shipped,
3. The height of 50 students in a class.

2.4.4. Data according to time. According to time data may be classified as:

- i) Time series data;

- ii) Cross-section data; and
- iii) Panel data.

(i) Time series data. A time series data is a set of ordered observations on a quantitative characteristic of an individual or collective phenomenon taken at different points of time. Although it is not essential, it is common for these points to be equidistant in time.

1. Year-wise production of a firm for last 15 years,
2. Month-wise price of potatoes for last six years,
3. Population of a country for last fifteen years,
4. Productions of a factory for last ten years etc are some examples of time series data.

(ii) Cross-section data. Data are called cross-section when they are collected from the experimental units at a particular period of time.

1. Salary of the workers of a factory for the month of July, 2005.
2. Height of the students of a class,
3. Weight of patients of a clinic, etc. are the examples of cross-section data.

(iii) Panel data. This is the mixture of time series and cross-section data.

1. Year-wise prices of different food stuffs for last ten years,
2. Income and expenditure of 50 workers of a factory for last 10 years,
3. Month-wise productions of different items of a firm for last 15 years are some examples of panel data.

2.4.5. According to sources. According to source data are known as:

- i) Primary data; and
- ii) Secondary data.

The concepts and methods of collection of these types of data will be discussed in the next chapter.

2.4.6. According to subject. Data may be business data, agriculture data, medical data, meteorological data, economic data etc

2.5. Descriptive and Inferential Statistic

Descriptive and Inferential Statistics. Statistical methods are those procedures used in the collection, classification, presentation, analysis and interpretation of data obtained from any field of enquiry. These methods are categorized belonging to one of two following major methods:

- i) Descriptive statistics, and
- ii) Inferential statistics.

Most of the statistical information in newspapers, magazines, company

reports, and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical or numerical, referred to as descriptive statistics.

Definition 2.26. Descriptive statistics comprises those methods concerned with collection and describing a set of data so as to yield meaningful information.

It can also be defined as: Descriptive statistics consists of procedure used to summarize and describe the important characteristics of a set of data.

If the data set is the entire population, we need only to draw conclusion based on the descriptive statistics. If it is not possible to get population data due to time, cost or other considerations, we have to take sample from the population. We have to infer about the whole population on the basis of the sample data. Thus, the branch of statistics that deals with generalization of results obtained from a sample to the whole population is called inferential statistics.

Definition 2.27. Inferential Statistics consists of procedures used to make inferences about population characteristics from information contained in a sample drawn from the population.

The objective of inferential statistics is to make inferences about the characteristics of a population from information contained in a sample.

Sometimes more than one variable can be generated from the experimental units of a population or sample. For example, we may get data on height, weight and age of a group of students. Here three variables (height, weight, age) are measured on a single experimental unit (student).

Definition 2.28. Univariate data result when a single variable is measured on a single experimental unit.

Definition 2.29. Bivariate data result when two variables are measured on a single experimental unit. Similarly, multivariate data result when more than two variables are measured.

Questions

1. Define population, sample, experimental unit and variables with suitable examples.
2. Identify the experimental units and variables of the following:
 - a. Gender of a garment worker,
 - b. Number of errors on a midterm exam,
 - c. Body temperature of a patient,
 - d. Age of a first year student,
 - e. Colour of a car entering the parking lot,
 - f. Number of typing errors per page of a manuscript,
 - g. Age of a cancer patient.
3. Distinguish between qualitative and quantitative variables. Identify each of the following variables as qualitative or quantitative:
 - a. The most frequent use of your microwave oven (reheating, defrosting, warming, other),
 - b. The number of consumers, who refuse to answer a telephone survey,
 - c. Number of unregistered taxicabs in a city,
 - d. Number of defective items per box each of which contains 10 items,
 - e. Size of a factory (large, medium and small),
 - f. Rating of performance of a newly elected politician (excellent, good, fair, poor),
 - g. Number of deaths in every year in the Chittagong University campus,
 - h. The size of a firm (big, medium and small).
4. Define discrete and continuous variables with examples. Identify following variable as continuous or discrete:
 - a. The number of automobile accidents per year in Dhaka city,
 - b. The length of time in minutes to complete a phone call,
 - c. The number of telephone calls made from a telephone booth,
 - d. The quantity of paddy produced per acre,
 - e. The quantity of milk produced in a year by a herd of cows,
 - f. Ages of workers of a factory,
 - g. Number of claims received by an insurance company in every year,
 - h. Systolic blood pressure of a patient.
5. What do you mean by a variable? Discuss different types of variables with examples.
6. Explain with examples how does a population differ from a sample.
7. Illustrate the difference between discrete and continuous variables with examples.
8. What do you mean by scale of measurement? Discuss different levels of

measurements with suitable examples.

9. Identify the experimental units, variables, types of variables and their scales of measurements on which the following variables are measured:
- Religion of an employee of a factory (Muslim, Hindu, Buddhist, Christen),
 - Official status of an employee (1st class, 2nd class, 3rd class, 4th class),
 - Monthly salary of an employee,
 - Monthly expenditure of an employee,
 - Number of family members of an employee,
 - Humidity at Chittagong city
 - Number of children per family of an employee,
 - Age of an employee,
 - Height of an employee,
 - Weight of an employee
 - IQ score of candidates applied for a post
 - Name of the students of BBA.

Applications

10. A data set consists of the age at the time of Chairmanship for each of past 20 chairmen in the department of statistics, University of Chittagong.
- Is the set of data a population or sample?
 - What is the experimental unit?
 - What is the variable measured?
 - Is the variable measured qualitative or quantitative?
 - Is the variable measured discrete or continuous?
 - What is the scale of measurement on which the variable is measured?
11. Suppose a medical researcher wants to find the average systolic blood pressure of the employees of a big firm. For this purpose, a sample of 50 employees has been selected randomly from that firm and their systolic blood pressure measurements have been obtained.
- What is the population?
 - What is the sample?
 - What is the experimental unit?
 - What is the variable being measured?
 - Is the variable qualitative or quantitative?
 - Is the variable discrete or continuous?
 - What is the scale of measurement on which the variable is measured?

12. University authority needs the following information about the students before making decision whether the student is eligible for any financial aid:
- GPA of H.S.C. examination,
 - GPA of S.S.C. examination,
 - Gender of applicant,
 - Parents' income,
 - Age of applicant.

Classify each of the above data as qualitative or quantitative. Also mention the corresponding scale of measurements.

13. Consider the set of all students enrolled in Business Statistics course this semester. Suppose you are interested to know the current grade point average (GPA's) of the students:
- Define the population and variable of interest.
 - Is the variable qualitative or quantitative?
 - Suppose you determine the GPA of every student of the class. Would this represent a census or sample survey?
 - Suppose you determine the GPA of randomly selected 10 students of the class. Would this represent a census or sample?
14. A researcher wants to survey on the houses or flats, which are available for the government officials in Dhaka city. For this he wants to know the following measurements
- Holding number of a house such as B₁ F₃, B₂ F₆, SE-7 etc.,
 - Area of a house or flat in square feet,
 - Room temperature in winter season,
 - Number of rooms per house,
 - Number of bedrooms per house,
 - Comfort: excellent, good and poor
 - Security system: (strong, medium, weak)

Identify the level of scales on which the variables are measured.

CHAPTER - 3

METHODS OF DATA COLLECTION

3.1. Introduction

Data constitute the foundation of statistical analysis and interpretation. They are the raw materials of any kind of statistical analysis. The success of any statistical investigation depends on the quality of data. Hence the first step in statistical work is to collect data through proper data collection technique. Data are collected from two important sources, namely

- i) Primary source and
- ii) Secondary source.

Depending on the source, data are called primary data and secondary data.

3.1.1. Primary data.

Definition 3.1.1. Primary data are those data which are collected by the investigator himself or by any research institution for the purpose of some specific inquiry or study.

Such data are collected afresh and for the first time and thus happen to be original in character and are generated by surveys conducted by individuals or research institutions.

For example, if a researcher is interested to know the salary structure of the female workers of a garment factory, he must take a survey and collect data on salary from the female workers of the relevant garment factory. The data so collected would be considered as a primary data.

Data collected by different government, public, private organizations, research bodies, research scholars, NGO's for their official records and research purpose from the field directly are primary data. Bangladesh Bureau of Statistics (BBS) is the main primary sources of government data.

3.1.2. Secondary data.

Definition 3.1.2. When an investigator uses data which have already been collected by others, such data are called secondary data.

Such data are primary data for the agency that collect them, and become secondary for some one else who uses these data for his own purposes. The secondary data can be obtained from Journals, reports, government publications, publications of professional and research organizations etc.

For example, if a businessman wants to study the movements of the share indices of different companies, he can get the required information or data from the share market related publications as secondary data. This means the primary data collected by an agency or organization; constitute the secondary data in the hands of other agencies.

All the data collecting methods are broadly divided into two categories viz.

- i) Methods of collecting Primary data, and
- ii) Methods of collecting Secondary data.

3.2. Methods of Collecting Primary Data

The collection of data may range from a simple observation to a large –scale survey on any defined population. The tools and techniques to be employed to collect data depend largely on the objectives of the study, the research design and the availability of the time, money and trained personnel. There are several primary data collecting methods each with its own advantages and disadvantages. The important primary data collecting methods are:

- 1. Through Interview,
- 2. Through Questionnaires,
- 3. Through Schedules,
- 4. Through Local agents,
- 5. Through observations,
- 6. Through experimentation.

A brief description on each of the above mentioned methods is provided below.

3.2.1. Through Interview. It is one of the important methods of collecting primary data. Interview methods can be categorized as:

- 1. Unstructured Interview,
- 2. Structured Interview,
- 3. Direct Personal or Face to face Interview,
- 4. Indirect Personal Interview,
- 5. Telephone Interview,
- 6. Computer-Assisted Interviewing.

(i) Unstructured Interview. By this method, the interviewer is free to ask any question to the respondents relating to the objective of the study. The main purpose of this method is to bring some preliminary issues to the surface so that the researcher can determine what variables need further in-depth investigation. Usually open ended questions are asked to the respondents. For example : Tell me something about your unit and department, and perhaps even the organization as a whole, in terms of work, employees, and whatever else you think is important.

(ii) **Structured Interview.** By this method, the interviewer has a list of predetermined questions to be asked to the respondents either personally, through the telephone, or through the medium of a PC. Here the same questions will be asked to everybody in the same manner. This is particularly important when a team of trained interviews conducts the survey.

(iii) **Direct personal or face to face Interview.** Under this method, the investigator presents himself personally before the respondents and ask questions in a face-to face contact and obtains a first hand information. This sort of interview is useful in the investigation of a specific issue or opinion survey.

Merits

1. The data collected by this method is very reliable and accurate,
2. The response rate is almost 100%.

Demerits

1. The method is not suitable for a large scale survey.
2. It is costly and time consuming.

Recommendation. The method is suitable when the field of enquiry is small and a greater degree of accuracy is required.

(iv) **Indirect Personal Interview.** Under this method, instead of directly approaching the respondents, the investigator interviews several third persons who are directly or indirectly concerned with the subject matter of the enquiry and who are in possession of the requisite information. For example, data regarding drug addicted, gamblers, smugglers, fighting between two groups of students etc. are successfully obtained by this method. This method is followed by the enquiry committees and commissions appointed by the government, universities and different organizations.

Merits

1. The method is less costly and less time consuming than the direct personal investigation.
2. Under this method, the enquiry can be formulated and conducted more effectively and efficiently

Limitation. The success of this method depends upon

1. The representative character of the witness.
2. The personal knowledge of the witness about the subject-matter of enquiry.
3. The ability of the investigator

The method is very popular in every day life.

(v) **Telephone Interview.** Under this method, the investigator contacts respondents on telephone and collects information from them. The method is best suited when information from a large number of respondents spread over a wide geographical area is to be obtained quickly, and the probable duration of each interview is, say, 10 minutes or less. Many market surveys are conducted through structured telephone interviews; particularly in developed region. The method is more convenient than personal interview.

Merits

1. It is less costly and less time-consuming and can be applied even to extensive fields of enquiries.
2. The method is more convenient than personal interview.

Limitation

1. The method excludes those who do not have telephone as also who have unlisted telephones.
2. The main limitation of this method is that the number of questions should be limited.

Most opinion surveys are done by this method.

(vi) **Computer-Assisted Interviewing (CAI).** Under this method, questions are flashed onto the computer screen and interviewer can enter the answers of the respondents directly into the computer. There are two types of computer-assisted interview programs: Computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI)

- **Computer-Assisted Telephone Interviewing.** Through this method the interviewer can collect data from people all over the world since PC is networked into the telephone system. The PC monitor promotes the questions with the help of software and the respondent provides the answers. The computer selects the telephone number, dials and places the responses in a file. The data are analyzed later.
- **Computer-Assisted personal Interviewing.** This method involves big investments in hardware and software. CAPI has an advantage in that it can be self-administered, that respondents can use their own computers to run the program by themselves once they receive the software and enter their responses, thereby reducing errors in recording.

3.2.2. Collection of data through questionnaires. This method of data collection is quite popular, particularly in case of big enquiries. Under this method, the investigator prepares a questionnaire containing a number of

questions relating to the objectives of the study. These questionnaires are sent by post to the respondents together with a polite covering letter explaining in details, the aims and objectives of collecting the information, and requesting the respondents to cooperate by furnishing the correct replies and returning the questionnaire duly filled in. In order to ensure quick response, the return postage expenses and some monetary incentives are given to the respondents by investigator. Now a day, questionnaires are also sent to the respondents through e-mail addresses. This method is usually adopted by the research workers, private individuals, private and public organizations and even by governments.

Merits

1. By this method, a large field of investigation may be covered at a very low cost.
2. This is the most economical method in terms of time, money and manpower.
3. Personal bias of the investigators or enumerators are eliminated by this method.

Limitations

1. The method can be applied to the educated respondents only.
2. The return rate of the filled questionnaires are not always satisfactory.

The method is suitable when the area of the investigation is large and the respondents are educated.

3.2.3 Collection of data through Schedule. Under this method, instead of sending the questionnaire through post, the investigator appoints local agents known as enumerators, who go to the respondents personally with the questionnaire called schedules, ask them the questions given there in, and record their replies. The enumerator should be given proper training to perform their job well. This method of data collection is very useful in extensive enquires and can lead to fairly reliable results.

Merits

1. The information collected through this method is more reliable.
2. Non-response rate is very low compared to questionnaire method.
3. The method can be applied to illiterate respondents.

Limitations

1. The method is more expensive and can be used by financially strong bodies or institutes only.
2. It is more time-consuming then mailed questionnaire method.
3. Skilled and well-trained enumerators needed to get reliable data.

Population census all over the world is conducted through this method.

Requirements of a good questionnaire. Questionnaire is a set of questions relating to the objective of the study. The success of the questionnaire method of collecting information depends largely on the proper designing of the questionnaire. Designing of questionnaire, requires a high degree of skill and experience on the part of the investigators. There is no hard and fast rules for designing or framing a questionnaire. While developing a questionnaire, the researcher has to be very clear on the following issues:

1. The number of questions should be related to the objective of the study.
2. The size of the questionnaire should not be too large.
3. Questions should be logically arranged.
4. Questions should be clear, brief, unambiguous, non-offending, and courteous in tone, corroborative in nature and to the point.
5. Questions of sensitive nature should be avoided.
6. The questionnaire should be simple and easy to understand.
7. The questions should be so designed that the respondents can easily comprehend and answer them.
8. Questions involving mathematical calculation should not be asked.
9. If a particular question needs clarification, it should be explained by way of a footnote.
10. The questionnaire should provide necessary instructions to the respondents.
11. The answers of most questions should be of multiple choices. That is closed questions should be more than the open questions.

Difference between questionnaire and schedule methods. Both questionnaire and schedule are popularly used methods of collecting data in research surveys. There is much resemblance in the nature of these two methods and this fact has made many people to remark that from practical point of view, the two methods can be taken to be the same. But from the technical point of view there is difference between the two. The important points of difference are:

1. The questionnaire is sent to the respondents in a covering letter whereas the schedule is filled out by the research workers or by the enumerators.
2. Mailed questionnaire method is cheaper and economical than the schedule method.
3. Non-response rate is higher in case of questionnaire method than schedule method.
4. Questionnaire method is slow compare to schedule method.

5. Questionnaire can be used in case of literate respondents, whereas schedule method can be even for illiterate respondents.
6. In case of questionnaire, it is not always clear who replies, but in case of schedule the identity of respondent is known.
7. The success of the questionnaire method depends on the quality of the questionnaire, but in case of schedules much depends upon the honesty and competence of enumerators.

3.2.4. Collection of data through local agent. Under this method, the information is not collected formally by the investigators, but by local agents, commonly known as correspondents, is appointed in different parts of the area under investigation. These agents collect information in their areas and transmit the same to the investigator. They apply their own judgments as to the best method of obtaining information. This method is usually employed by newspaper or periodical agencies who require information in different fields like economic trends, business, stock and share markets, sports, politics and so on.

Merits

1. The method is very cheap and economical for extensive investigations.
2. The required information can be obtained expeditiously since only rough estimates are required.

Limitations

Information collected through this method may be subject to bias due to personal choice of local agents.

3.2.5. Data collection through observation. Observation is one way to collect primary data. Observation is a purposeful, systematic and selective way of watching and listening to an interaction or phenomenon as it takes place. There are many situations in which observation is the most appropriate method of data collection; for example, when you want to study the animal behaviour, behaviour of a group of people etc. There are two types of observation:

- (i) Participant observation, and
 - (ii) Non-participant observation.
- (i) **Participant observation.** It is a method when the researcher participates in the activities of the group being observed in the same manner as its members, with or without their knowing that they are being observed. For example, one can study the economic, social, cultural behaviour of a class of people or race by this method.

(ii) Non-participant observation. When the researcher will not participate in the activities of the group but remains a passive observer, watching and listening to its activities and drawing conclusions from this. For example, a researcher wants to study the functions carried by a nurses in a hospital or a manager of a firm. As an observer, a researcher could watch, follow, and record the activities as they are performed. After making a number of observations, conclusions could be drawn about the functions of a nurse's carry out in the hospital or the manager in the firm.

In this method data are also collected by observation through mechanical devices such as the reading of X-ray films, E.C.G. etc. Meteorological data such as rainfall, temperature, humidity etc.; laboratory data of blood, urine, stool etc. are obtained by observations

3.2.6. Collection of data through experimentation. In natural sciences like physics, chemistry and in biological sciences like botany, zoology, biochemistry, microbiology and pharmacy, data are obtained when treatments are applied on experimental units in a controlled laboratory experiment. There are some important basic experimental designs such as (i) completely randomized design, (ii) randomized block design and (iii) Latin square design are used to obtain a set of good experimental data.

3.3. Sources of Collecting Secondary Data

There is no method of collecting secondary data like the method of collecting primary data. We can get secondary data from different sources. Before using secondary data the investigator should examine the following aspect :

1. Whether the data are suitable for the purpose of investigation.
2. Whether the data are adequate for the purpose of investigation.
3. Whether the data are reliable.

The chief sources of secondary data may be, broadly classified into the following two groups:

- (i) Published sources
- (ii) Unpublished sources

(i) Published Sources. There are a number of national organizations and also international agencies which collect and publish statistical data relating to business, trade, labor, price, consumption, production etc. These publications of the various organizations are used for sources of secondary data. Some of these published sources are as follows:

- Government, semi-government organizations, municipalities, local bodies, universities publish a lot official statistics in report form. Bangladesh Bureau of Statistics(BBS) is the main sources of government official statistics.

- International organization like UNO, FAO, UNPD, ILO, IMF, UNESCO, UNEP, WHO, USAID, etc have got their official, publications containing valuable statistical information.
 - Trade Journal, technical and financial journals as well as news papers contain a good deal of statistical data.
- (ii) **Unpublished sources.** The records maintained by private firms or business houses who may not like to release their data to any outside agency, the researches carried out by the research scholars in the Universities or Research Institutes may also provide useful statistical data.

3.4. Selection of Appropriate Method

Selection of appropriate method for data collection. One must always remember that each method of data collections has its uses and none is superior in all situations. For instance, telephone interview method may be considered appropriate if funds are restricted, time is also restricted and the data is to be collected in respect of few items with or without a certain degree of precision. This method is appropriate for conducting opinion survey. In case funds permit and more information is desired, personal interview method may be said to be relatively better. In case of time is ample, funds are limited and much information is to be collected with no precision, then questionnaire method can be regarded more reasonable. When funds are ample, time is also ample and much information with no precision is to be collected, then either direct personal interview or the questionnaire or the joint use of these two methods may be considered as an appropriate method of collecting data. When a vast geographic area is to be covered, the use of questionnaire supplemented by personal interviews will yield more reliable results.

With greater technological advancement and a reduction of hardware and software costs, computer-assisted interviews promise to become a primary method of data collection in the future.

The secondary data may be used in case the researcher finds them reliable, adequate and appropriate for his research.

Thus, the most desirable approach with respect to the selection of the method depends on the nature of the particular problem and on the time and resources available along with the desired degree of accuracy.

3.5. Processing of Data

The data, after collection has to be processed through editing and coding.

3.5.1. Editing. Once the set of data have been collected, it is necessary to process them for proper presentation. Editing of data is required as preparatory work before the tabulation and statistical analysis is carried out. This is quite a difficult job and requires a great deal of skill and experience. While editing primary data the following consideration need attention:

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

It should be noted that these days computer is extensively used to edit data.

3.5.2 Coding. When the data is to be processed by computer, it must be coded and converted into the computer language. For some qualitative characteristics, the code numbers can be assigned and identified. For example, to a question, "Are you married or single?" a code of 1 can be assigned to the qualitative answer "married" and a code of 0 can be assigned to the answer "single". This coding job should be accomplished while editing the data.

The next step is to code the responses. Nowadays Scanner sheets are available to collect data through questionnaire. Such sheets facilitate the entry of the responses directly into the computer without manual keying in of the data. Raw data can also be entered through any software program say SPSS, Excel etc. Statistical analysis can also be done with the help of these software.

Sample Questionnaire

Questionnaire

A study entitled 'Determinants of Credit Cards Users in Chittagong City' was prepared by a graduate student for the partial fulfillment of his M.S. degree. The main objectives of the study are as follows:

- To know the information of the cardholder's about their social and educational status.
- To know the information of the cardholder's about their income and expenditure behaviour.
- To know the idea about the age structure of the credit card users.
- To know the information about the income group who uses credit cards
- To know the information about the types of cards they use.
- To know the idea about the satisfaction level of the cardholders.
- To know the problems they frequently faced.

- To know the information about the monthly average transaction of money by the cardholders
- To take idea about the reason of using credit card.
- To acquire knowledge about credit card system and operational system of credit card.

Questionnaire

1. Name:

2. Age:

Year	Month	Day

3. Sex: Male Female

4. Educational qualification:

- Graduate and above HSC
 SSC Others

5. Marital status: Single Married Divorce

6. Is Spouse Employed? Yes No

7. Where do you reside? (For example: Agrabad)

8. Your resident status: Own Rented Company provided
 Government provided

9. Your job status: Government employee Businessman
 Retired Student Others

10. Car ownership: Yes No

11. Family economic status: Lower Middle Higher

12. Monthly Income: _____

13. Monthly Expenditure: _____

14. Spouse Income (if any): _____

15. Total Income (Both): _____

16. Do you use debit card: Yes No

17. Which Bank's Credit card do you use? _____

18. Type of your card:

- Local Classic/Silver Local Gold Dual Classic/Silver
 Dual/International Gold Others (Please specify)

19. Credit limit : _____

20. Average monthly transactions by credit card : _____

21. How many Credit card you use? One Two Three or more

22. You use your International credit card more in:

- Home country Abroad

23. For which purpose you use the credit card frequently?

- Business Purchases Smartness 24 hours service

24. You have taken card: Willingly By influence
25. Which problem you face usually?
 High interest rate Booth facilities Others (please specify)
26. What is your satisfaction status?
 Satisfied Highly satisfied Not satisfied
27. Do you want to continue in future? Yes No
28. Do you encourage your friends or relatives to take credit card?
 Yes No

Thank you for your co-operation

Questions

1. What are the main sources data? Define primary and secondary data.
Discuss the important methods of collecting primary data.
2. What is primary and secondary data? Distinguish between primary and secondary data. State the important sources of secondary data.
3. Discuss questionnaire and schedule methods for collecting primary data. Distinguish between questionnaire and schedule method.
4. Define secondary data. What are the sources of secondary data? State the precautions necessary for using secondary data.
5. What do you mean by editing the data? Explain clearly the procedure that you would adopt in editing primary data.
6. Discuss different types interview methods for collecting primary data.
7. Discuss how you can select an appropriate method for collecting data.

CHAPTER - 4

PRESENTATION OF DATA

4.1. Introduction

The data which has been just collected are in the raw or disorganized form. So, it is quite difficult to explore any information about the characteristics of interest from this raw data at a glance, particularly, in case of enormous data. Hence, it is required to arrange or present data using some statistical tools.

After collecting and editing data, the next important step towards processing the data is classification. Classification is the first statistical technique to condense the raw data. It is the process of arranging data in different groups or classes according to their affinities. Facts of one class differ from those of another class with respect to some characteristics called the basis of classification. It is like sorting of letters in a post office. The number of ways in which statistical data may be classified is almost limitless. We may get number of different types of variables in any statistical investigation through questionnaire or observation method. That is data may be either qualitative or quantitative in nature. Again quantitative data may be either discrete or continuous. Very often for further statistical analysis collected raw data may be classified according to variables under study such as:

- (i) Qualitative data,
- (ii) Discrete data, and
- (iii) Continuous data.

Once classification has been done, the classified data can be condensed and summarized in two basic forms:

- (i) Tabular form (known as Frequency distribution), and
- (ii) Diagrammatical and graphical form.

4.2. Condensing and Summarizing Data

From the statistical point of view, frequency distribution is the most important tabular form of condensing and summarizing a large mass raw statistical data. Some concepts related to frequency distribution are briefly described below. Illustration of the concepts is provided later with an example.

Frequency. Number of times an observation occurs in a set of data is called the frequency of that observation. Number of observations in a class is called the frequency of that class or category.

Frequency distribution. A frequency distribution is a tabular summary of data where observations are divided into different non overlapping classes or categories and frequency of each class or category is arranged accordingly.

According to nature of variable, there may be three types of frequency distribution, namely:

1. Frequency distribution for qualitative or categorical variable,
- 2 Frequency distribution for discrete variable and
3. Frequency distribution for continuous variable.

4.2.1. Some Numerical measure of Categorical or qualitative data. Although qualitative data are non numerical in nature, we can measure them by simple mathematical formulae. Now, we cite some of these measures.

Proportion. The proportion of a category is the ratio of the number of elements falling in the category to the total number of elements under study. For example,

$$\text{Proportion of smoker} = \frac{\text{number of smokers}}{\text{Total number of smokers + non smokers}}$$

Another example may be the proportion of defective items produced by a factory. It is a pure number and is always less than 1. Note that, in this expression, the numerator is always a part of denominator.

Percent. Percent of a category is obtained by multiplying the proportion of that category by 100. It is always expressed per 100.

Proportion and percentage are used in the same purpose. But percent is more useful for comparison than proportion.

Rates and ratios are other two mathematical formulae by which we can express qualitative data meaningfully. All these measures are used for comparison between two classes.

Ratio. A ratio is a fraction used to show the magnitude of a quantity relative to the magnitude of another.

A ratio can consist of any two numbers. Number of students per teacher is called student-teacher ratio. It is computed by the formula,

$$\text{Student-teacher ratio} = \frac{\text{Total number of students}}{\text{Total number of teachers}}$$

On the other hand sex ratio is the total number males per one hundred females. That is,

$$\text{Sex ratio} = \frac{\text{Total number of males}}{\text{Total number of females}} \times 100$$

For example, sex ratio 105 of a population means there are 105 males per 100 females in the population.

Rate. The rate of category is the ratio of the number elements of the category to the total number of elements under study.

It is normally expressed per 100, per 1,000 or per 10,000, etc. For example,

$$\text{Literacy rate} = \frac{\text{Number of literate persons}}{\text{Total number of persons in the population}} \times 100$$

It is expressed per 100 persons. Birth rate and death rate are expressed per 1,000. But maternal mortality rates are usually expressed per 10,000.

Class. A class is one of the categories into which qualitative data can be classified

Class frequency. The class frequency is the number of observations in the data set falling in a particular class

Relative frequency. The class relative frequency is the class frequency divided by the total number of observations n in the data set. The formula for relative frequency is,

$$\text{Relative frequency} = \frac{\text{Frequency of a class}}{n}$$

Percent frequency. The percent frequency in each class is the class relative frequency $\times 100$.

$$\text{Percent frequency} = \text{Relative frequency} \times 100.$$

It can be easily verified that the sum of the frequencies is n , the sum of the relative frequencies is 1 and the sum of the percentages is 100.

4.2.2 Frequency distribution of categorical or qualitative data. Frequency distribution for categorical variable is a table where observations of a qualitative variable are divided into different mutually exclusive categories. Here each category is called class. Usually it contains three columns. Different categories of the qualitative variable as class are in the first column, observations fall in each class presented by tally marks are in the second column, and frequencies for each class are shown in the third column.

Sometimes, qualitative data are shown as relative frequency distribution or percent frequency distribution.

A relative frequency distribution is a tabular summary of data showing the relative frequency for each class.

A percent frequency distribution is a tabular summary of data showing the percent frequency for each class.

Let us consider some examples to demonstrate the construction and interpretation of frequency distribution for qualitative data.

Example 4.2.1. The owner of a departmental store wants to know the weekly requirements of different categories of shirts demanded by the customer. For this, he collected the information on the numbers of different categories of shirts sold by the store in last one week. The data obtained by the store are :

M, S, L, XL, M, M, M, S, S, S, M, M,
XL, XL, L, M, M, M, M, S, S, L, L, XL,
M, M, M, S, S, L, L, S, M, M, L, L,
M, M, L, XL, S, S, S, M, M, M, S, XL,
L, L, M, M, M, S, M, M, S, XL, L.

Here S, M, X, and XL denote small, medium, large and extra large size of shirt respectively. Construct a frequency distribution table for the above categorical data and comment.

Solution. Here the data are ordinal. First, we draw a table with three columns and six rows. The first column contains the sizes of the shirts as class, the second column contains tally marks to count number of shirts in each class and the third column contains the number of shirts in each class as frequency. To construct the frequency distribution, follow the following steps:

1. There are four sizes of shirts, denoted by S, M, L and XL which are placed in the first column as categories or classes.
2. Now we read the observations one by one and put a tally mark in the second column, against each observation we read. For example, the first observation is M. So, put a tally mark against the class M. We continue the process until we finish putting tally marks for all the observations in the data set.
3. Now, we count the number of tallies in each class. This count is the frequency of that class. As we know frequency is the number of times that an observation occurs in the data. For example, the number of tallies in category M is 26. So the frequency of the medium size shirt is 26.

It is seen that small size appears 15 times, medium size appears 26 times, large size appears 12 times and ex-large size appears only 7 times. These counts are summarized in a frequency distribution presented in Table 4.2.1.

Table 4.2.1. Frequency distribution of the shirts sold by a departmental store.

Size of the shirts	Tally marks	Frequency (# of shirts)
S		15
M		26
L		12
XL		7
Total		60

Comment : This frequency distribution provides summary information of the distribution of the different sizes of shirts sold by the store in a week. Viewing the frequency distribution, we can say that medium is the most popular size of the shirt sold by the store. The owner of the store can use the information to plan how many different sizes of shirts he should keep this store to satisfy the need of the customers.

To get more information about characteristics of the data, we need relative and percent frequency distribution of the frequency distribution obtained in example 4.2.1.

Example 4.2.2. Construct relative and percent frequency distribution table for the frequency distribution given below:

Class (Size of the shirt)	Frequency (number of shirts)
S	15
M	26
L	12
XL	7
Total	60

Find the percentage of medium size shirt sold by the store in the last week.

Solution. The relative frequencies and percent frequencies are calculated by the following formulae:

$$\text{Relative frequency} = \frac{\text{Frequency of a class}}{n}$$

Percent frequency = Relative frequency \times 100. For example, the relative frequency for L = 12/60 = 0.20. Percent frequency = 0.20 \times 100 = 20.

Other relative frequencies and percent frequencies of different classes are calculated in the same way and are placed in third and fourth columns of the table 4.2.2.

Table 4.2.2. Relative and percent frequency distributions of different sizes of shirts sold by a store.

Size of shirt (class)	Frequency	Relative frequency	Percent frequency
S	15	0.25	25
M	26	0.43	43
L	12	0.20	20
XL	7	0.12	12
Total	60	1.00	100.

Comment : The percentage of the medium size shirt sold by the store is 43%, which is the highest. On the other hand, the lowest percentage is the ex-large size, which is only 12%. This table gives us better information about the characteristics of the data than the frequency distribution table 4.2.1. Accordingly, the owner of the store can take better plan.

Example 4.2.3. The administration of a factory wants to know the health status of the workers of his factory. They conducted a survey on 55 workers. The health conditions of 55 workers of the factory were found as follows:

G, P, A, P, P, A, A, A, A, A, A, A, A, A,
 A, P, G, G, P, A, A, P, G, P, A, A, P, G, A,
 A, G, P, A, A, A, G, P, P, A, A, G, P, P,
 P, A, G, A, A, A, P, P, G.

Here G, A and P denote good health, average health and poor health respectively. Construct a frequency distribution table and comment.

Solution. Here the data are ordinal. A worker can be classified as having good health, average health or poor health. The above data can be displayed by a categorical frequency distribution as follows:

Table 4.2.3. Health Status of 55 workers of a factory.

Health Status	Tally marks	Frequency (# of workers)
Good		10
Average		29
Poor		16
Total		55

This frequency distribution provides a summary of the health condition of 55 workers of the factory. This summary provides more insight than the

original data. It is seen from the frequency distribution that the health condition of most of the workers are average.

Example 4.2.4. Construct a relative and percent relative frequency distribution of the following frequency distribution.

Class (Health status)	Frequency (number of workers)
Good	10
Average	29
Poor	16
Total	55

Comment on the health condition of the workers of the factory.

Solution. Relative frequency and percent relative frequency are calculated by the following formulae:

$$\text{Relative frequency} = \frac{\text{Frequency of a class}}{n}$$

$$\text{Percent} = \text{Relative frequency} \times 100.$$

The relative frequencies and the percent frequency of the distribution are shown in the third and fourth column respectively in table 4.2.4.

Table 4.2.4. Relative and percent frequency distribution of health condition of 55 workers

Health Status	Frequency	Relative Frequency	Percent frequency
Good	10	0.18	18
Average	29	0.53	53
Poor	16	0.29	29
Total	55	1.00	100

Comment : The percentage of workers having good health, average health and poor health are 18%, 53% and 29% respectively. It is seen that most of the workers have average health condition which is 53%. On the other hand only 18% have good health condition and 29% of the workers have poor health condition. Better medical care should be given to the workers to insure quality and quantity of goods.

4.3. Diagrams for Categorical Data

Once the qualitative data have been categorized and summarized in a frequency distribution table, the next step is to display the frequency distribution of the qualitative data by different diagrams. Important diagrams for presenting categorical data are:

- (i) Bar diagram,

- (ii) Pie-diagram,
- (iii) Pareto diagram,
- (iv) Pictogram.

4.3.1. Bar diagram. There are three types of bar diagram. They are :

- (i) Simple bar diagram,
- (ii) Multiple bar diagram and
- (iii) Component bar diagram.

Here we shall discuss the diagrams one by one

4.3.2. Simple bar diagram. Simple bar diagram is the most popular diagrammatical representation of qualitative data. By simple bar diagram, only one qualitative variable can be exhibited. First a frequency distribution table for qualitative data is constructed. Then a bar of fixed width above each category or class is drawn. Actually bars are looked like rectangles of equal width over classes. The height of the bar over each class is equal to the class frequency. The bars should be separated by uniform gap to emphasize the fact that each class (category) is separate. Sometimes, bar diagrams are drawn for relative frequency or percent frequency distributions too. In those cases, the heights of the bars are taken proportional to class relative or percent frequencies.

Bar diagrams are widely used because they are easy to construct and easy to understand. They are effective tools for visual interpretations or comparison of data. Television, newspapers, magazines, books and reports make substantial use of this diagram to visually communicate to the viewer.

The main limitation of these diagrams is that only one qualitative variable can be displayed through these. More than one qualitative variables can be exhibited by multiple or component bar diagrams.

Example 4.3.1. The frequency distribution of the health condition of 55 workers of a factory is as follows:

Category (Health condition)	Frequency (Number of workers)
Good	10
Average	29
Poor	16
Total	55

Construct a bar diagram with the frequency table.

Solution. The different category of the health condition of the workers is plotted on the x-axis and the number of workers as frequency is shown on the y-axis. The bar diagram is shown in Fig. 4.3.1.

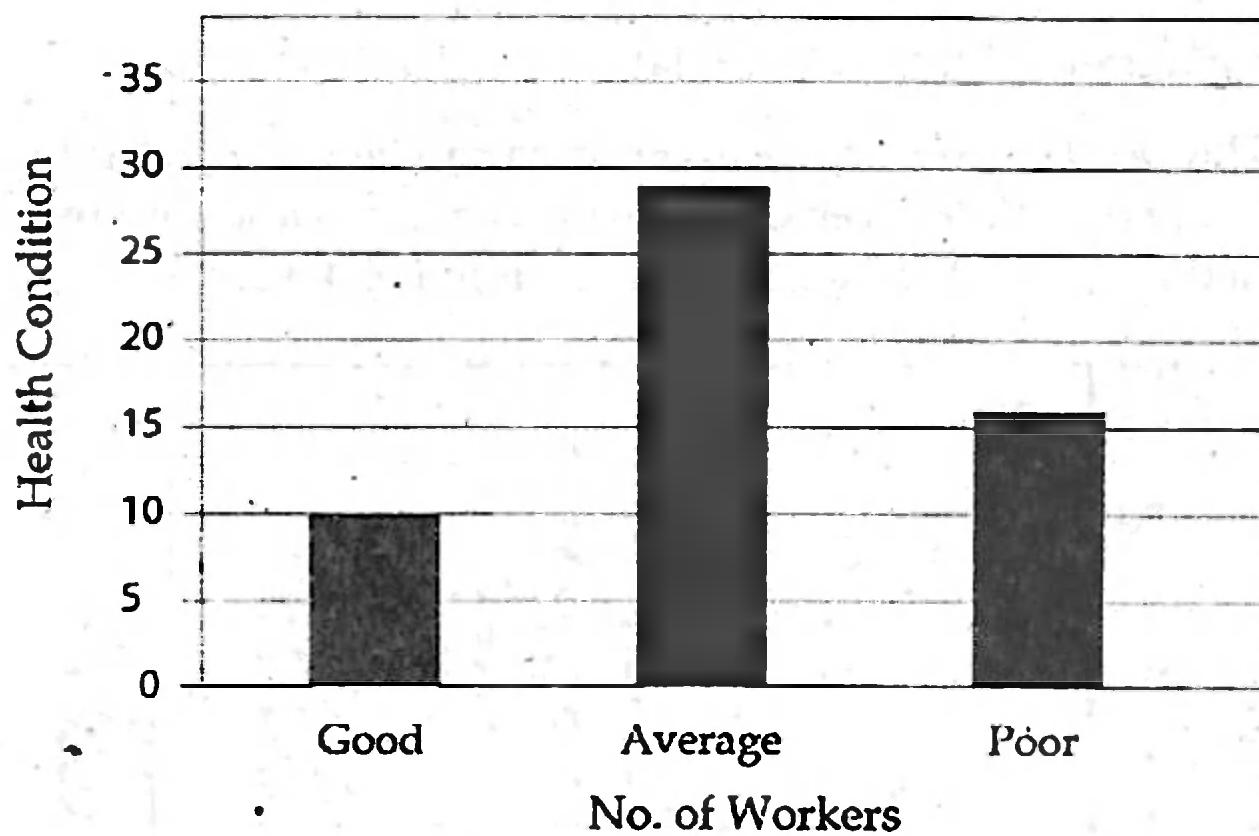


Fig. 4.3.1. Bar diagram of the health condition of 55 workers of a factory.

Example 4.3.2. Construct a bar diagram with the following frequency table.

Size of the shirt	S	M	L	XL
Frequency (number of shirts)	15	26	12	7

Solution. The bar diagram of the above data is shown by plotting different sizes of shirts on the X-axis and the number of shirts sold on the Y-axis. The bar diagram is displayed in Fig. 4.3.2.

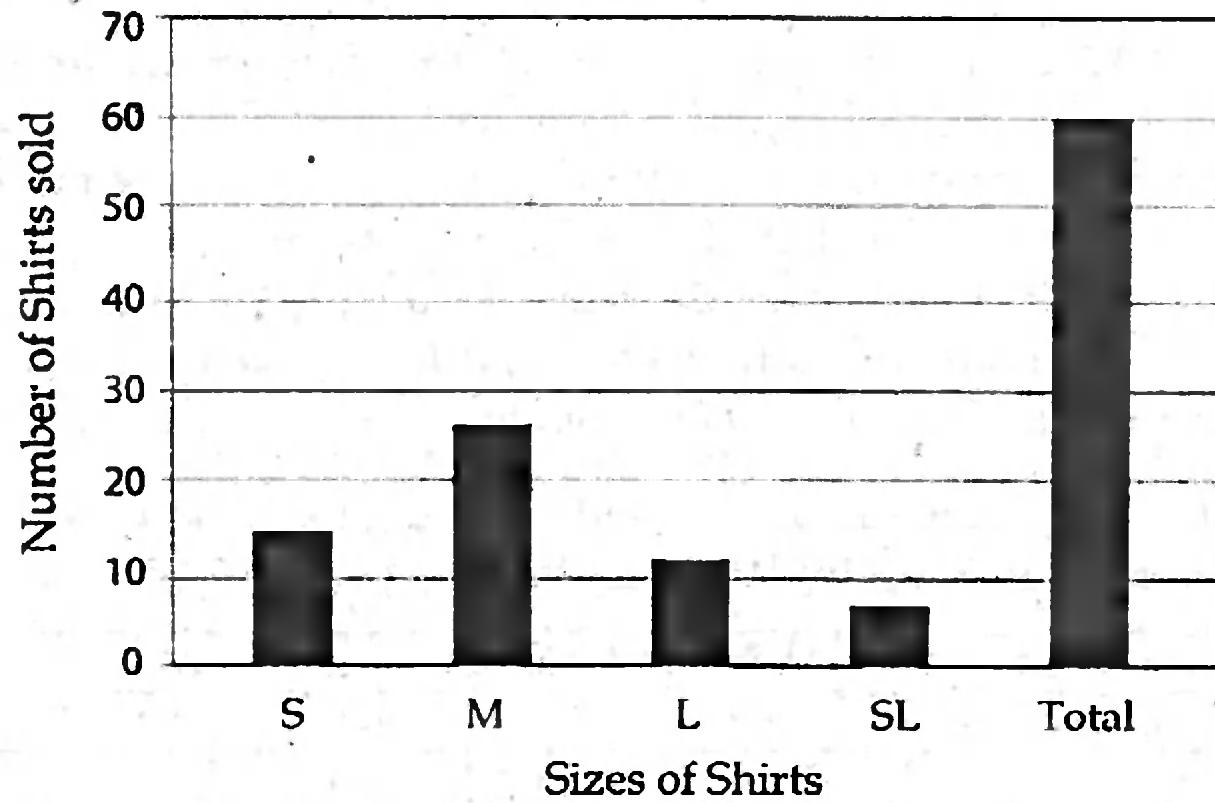


Fig. 4.3.2. Bar diagram of the shirts sold by a store.

Example 4.3.3. The following data give the expenditure budget in core taka of different sector of a country for the financial year 2005.

Sector	Transport	Education	Agriculture	Industry	Others	Total
Expenditure	25	40	80	70	55	270

Draw a bar diagram with the above data.

Solution. The bar diagram for the above data can be constructed by plotting the expenditure of different sectors in the vertical axis and different years in the horizontal axis. The bar diagram is shown in fig. 4.3.3.

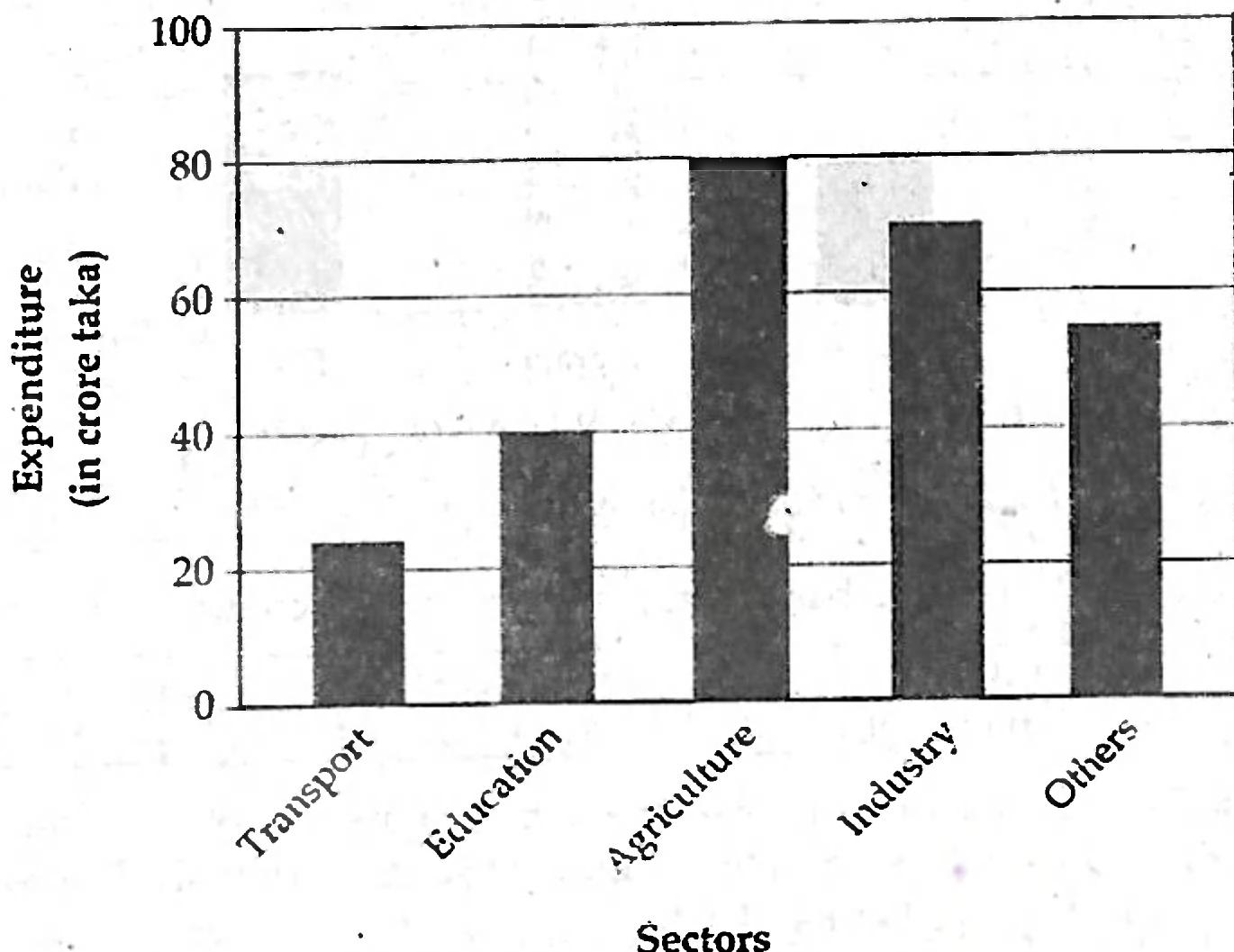


Fig. 4.3 3. Bar diagram of Expenditure Budget of different sectors.

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

Example 4.3.4. The following data give the pieces of different sizes of shirts sold by departmental stores for last five years:

Year /Size of the shirt	2005	2006	2007	2008	2009
Small	40	50	60	80	100
Medium	100	175	225	300	370
Large	60	90	140	180	220
Extra-Large	20	30	40	60	90

Construct a multiple bar diagram to compare the sales of different sizes of shirts for the last five years.

Solution. The multiple bar diagram for the above data can be constructed by plotting the different years in the horizontal axis and all types of sizes in vertical axis. The multiple bar diagram is shown in figure 4.3.4.

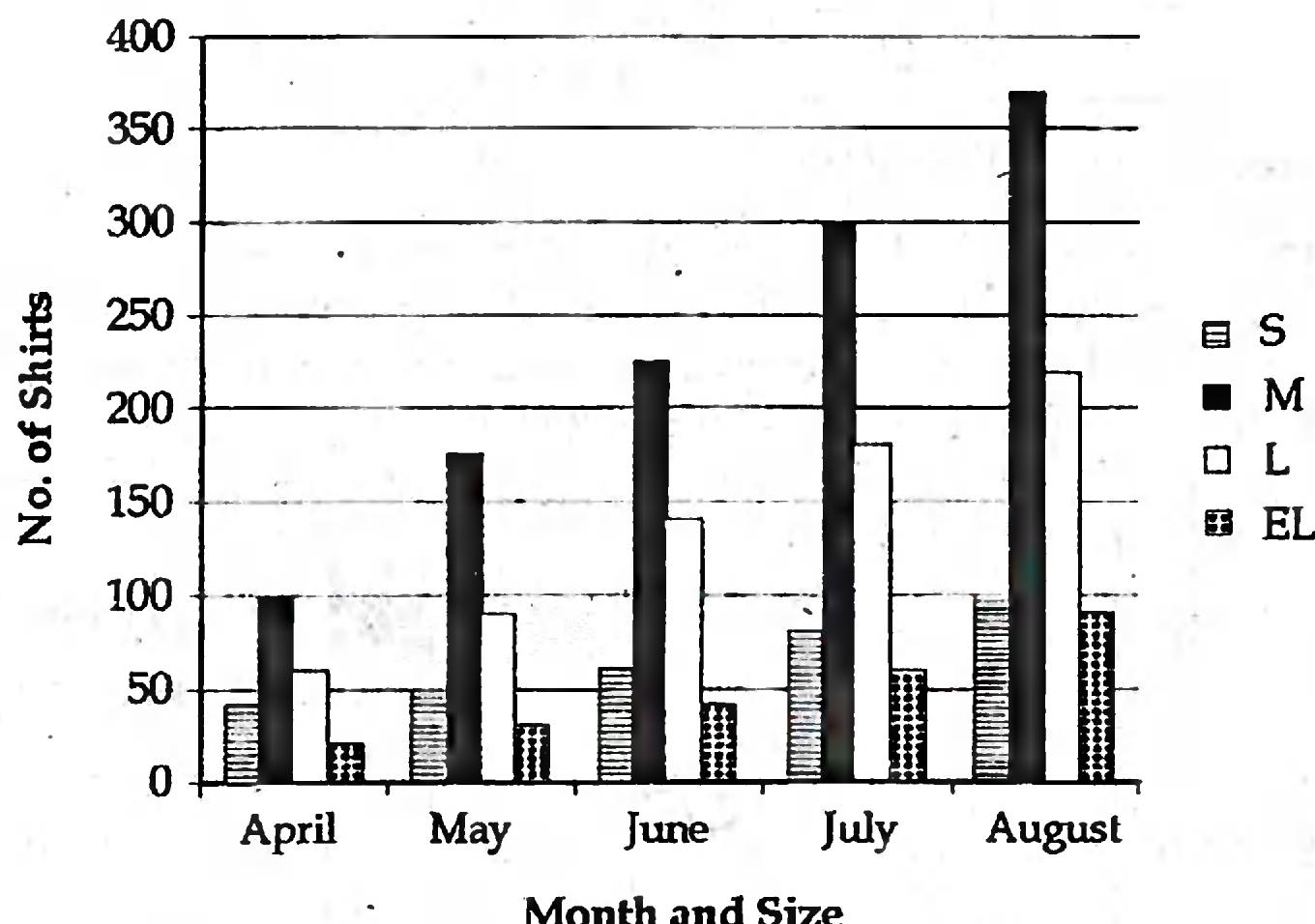


Fig. 4.3.4. Multiple bar diagram of different sizes of shirts for the year 2005-2009.

A matched problem to solve. Represent the following data by multiple bar diagram :

	Corporate sector profits in crores taka	
	2007-2008	2008-2009
Gross profits	3100	31500
Profits before tax	1600	1800
Profits after Tax	1350	1400
Retained profits	900	1050

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

Example 4.3.5. The following table gives the expenditure in taka of a family on different items for last three years:

	Year .	2005	2007	2009
Item	Food	5000	8500	12000
	Clothing	1500	2500	4000
	Education	2000	3000	5000
	Others	2500	4000	7000

Draw a component bar diagram with the above data.

Solution. The component bar diagram for the given data is constructed by plotting expenditure of different items in different years in vertical axis against items in horizontal axis and drawing rectangular vertical bars for expenditures in different years as shown below:

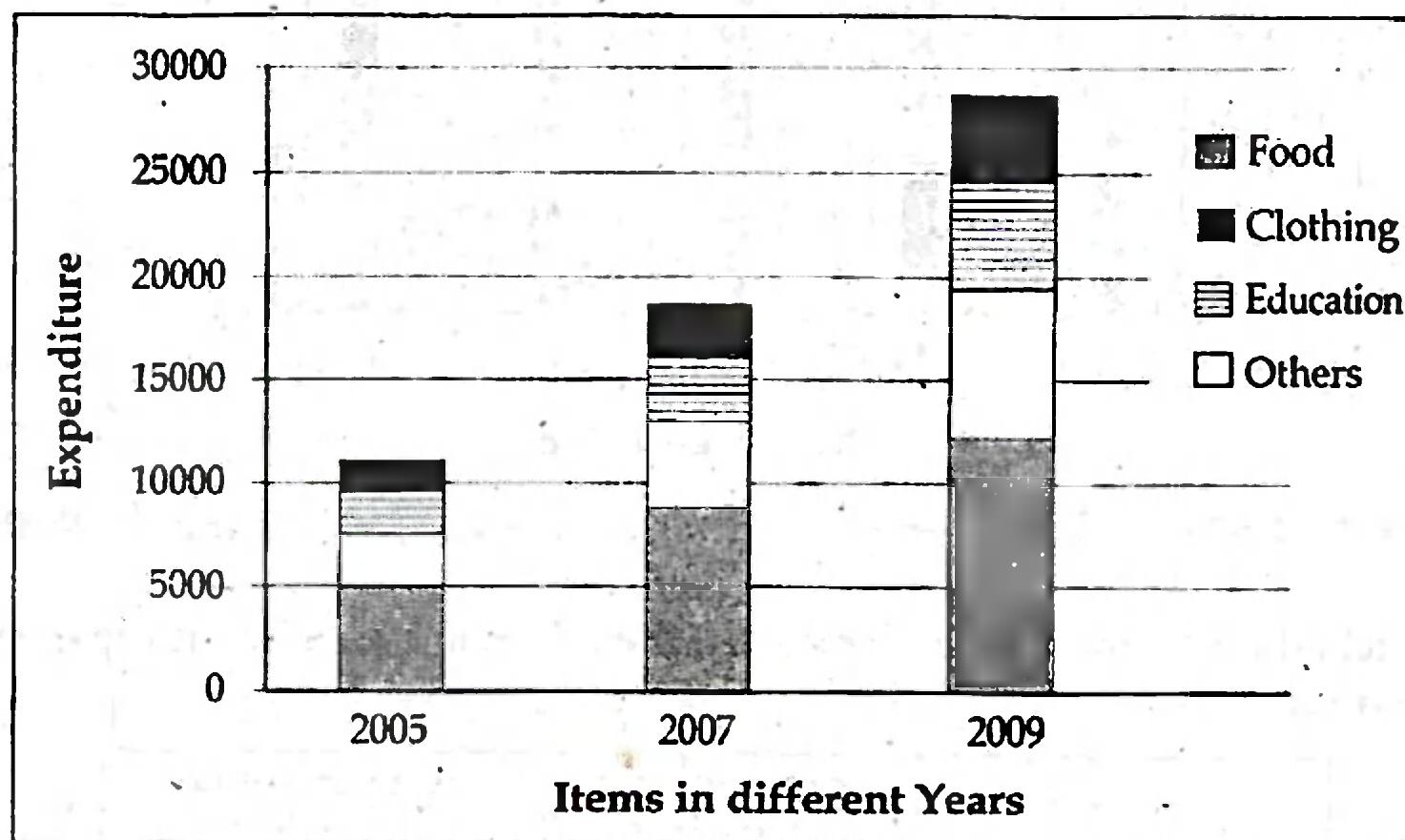


Fig 4.3.5. Component Bar diagram showing expenditure on different items in different years.

A matched problem to solve. The following table gives the budget in lakh taka of a company in different sectors for last four years:

	Year	2005	2007	2009	2010
Item	Raw materials	500	850	1200	1500
	Salary	150	250	400	550
	Others	250	400	600	800

Draw a component bar diagram with the above data.

4.4. Pie Diagram or Chart

A pie diagram is generally used to show how a whole is divided among several categories. The diagram is in the form of a circle and is also called a Pie, because the diagram looks like a pie and the components resemble

slices cut from it. The size of the slice represents the proportion of the component out of the total.

The pie chart is based on the fact that the total size of circle has 360° . The pie is divided into slices according to the percentage in each category. The percent frequency of a category is:

$$\text{Percent frequency} = \text{Relative frequency} \times 100$$

The angle of a category can be computed by the following formula:

$$\text{Angle} = \text{Relative frequency} \times 360^\circ$$

Example 4.4.1. Construct a Pie chart with the following frequency table.

Size of the shirt	S	M	L	XL	Total
Frequency (number of shirts)	15	26	12	7	60

Solution. The relative frequencies, percent frequencies and the angles corresponding to different sizes are calculated by the following formulae:

$$\text{Relative frequency} = \frac{\text{Frequency of a Size}}{\text{Total Frequency}}$$

$$\text{Percent frequency} = \text{Relative frequency} \times 100$$

The angle of a category can be computed by the following formula:

$$\text{Angle} = \text{Relative frequency} \times 360^\circ$$

For example, the relative frequency (S) = $15/60 = 0.25$

$$\text{Percent frequency (S)} = 0.25 \times 100 = 25$$

$$\text{Angle (S)} = 0.25 \times 360 = 90$$

The relative frequency, percent frequency and the angles corresponding to different sizes of shirts are calculated in similar manner and are shown in table 4.4.1. The corresponding pie-chart is shown in Figure 4.4.1.

Table 4.4.1. Relative and percent frequency and angles of different sizes of shirts sold by a store.

Size of shirt (class)	Frequency	Relative frequency	Percent frequency	Angles in degree
S	15	0.25	25	90.0
M	26	0.43	43	154.8
L	12	0.20	20	72.0
XL	7	0.12	12	43.2
Total	60	1.00	100.	360.0

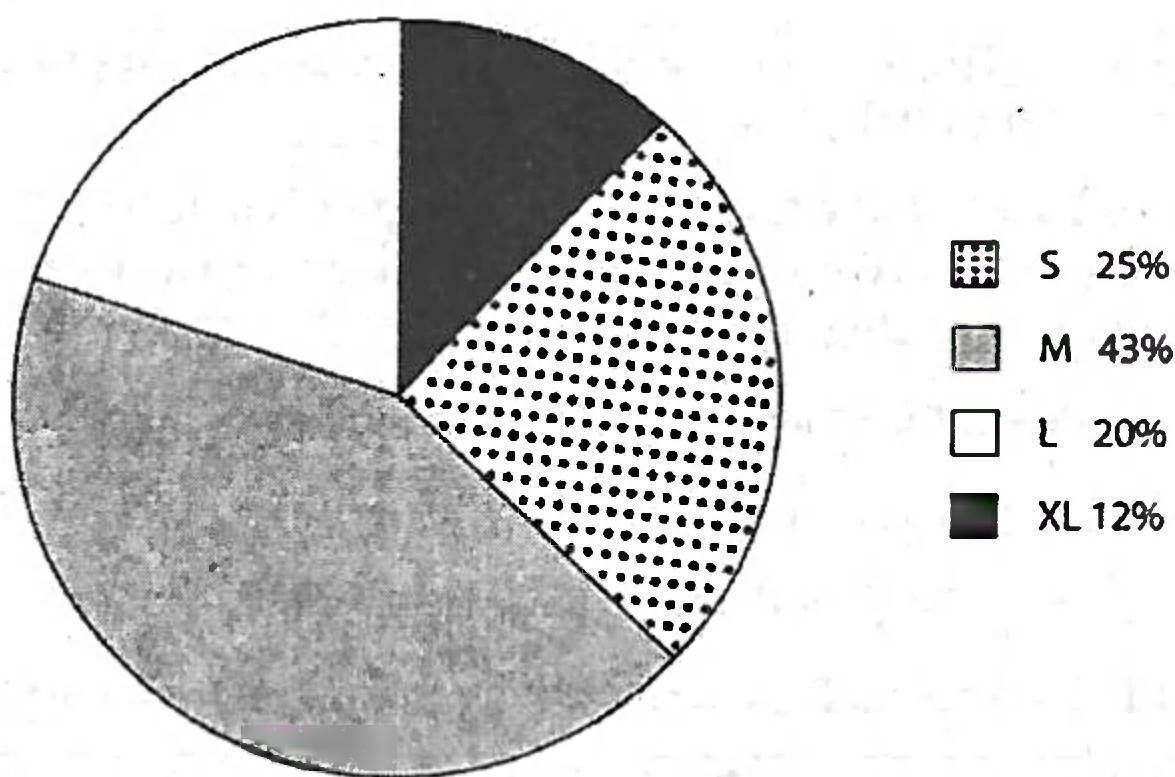


Fig. 4.4.1. Pie -chart of different sizes of shirts sold by a store.

Example 4.4.2. The following data gives the expenditure budget in crore taka of different sector of a country for the financial year 2005:

Sector	Agriculture	Industry	Education	Transport	Others	Total
Expenditure	80	70	40	25	55	270

Construct a pie-chart with the data.

Solution. The relative expenditures, percent expenditures and angles of different sectors are calculated by the following formulae:

$$\text{Relative expenditure} = \frac{\text{expenditure of any sector}}{\text{Total expenditure}}$$

$$\text{Percent expenditure} = \text{Relative expenditure} \times 100$$

The angle of a sector can be computed by the following formula:

$$\text{Angle} = \text{Relative expenditure} \times 360^\circ$$

$$\text{For example, the relative expenditure for education} = \frac{40}{270} = 0.1481$$

$$\text{Percent expenditure for education} = 0.1481 \times 100 = 14.81$$

$$\text{Angle for education} = 0.1481 \times 360^\circ = 53.33^\circ$$

The relative expenditures, percent expenditures and angles for different sectors are calculated in similar manner and are shown in table 4.4.2.

Table 4.4.2. Relative frequencies, percent frequencies and angles of different sectors.

Sector	Expenditure	Relative Expenditure	Percent Expenditure	Angles of different sectors in degree
Agriculture	80	0.30	29.63	108.00
Industry	70	0.26	25.93	93.60
Education	40	0.15	14.81	54.00
Transport	25	0.09	9.26	32.40
Other	55	0.20	20.37	72.00
Total	270	1.00	100.00	360

Pie chart of the expenditure of different sectors is exhibited in figure 4.4.2.

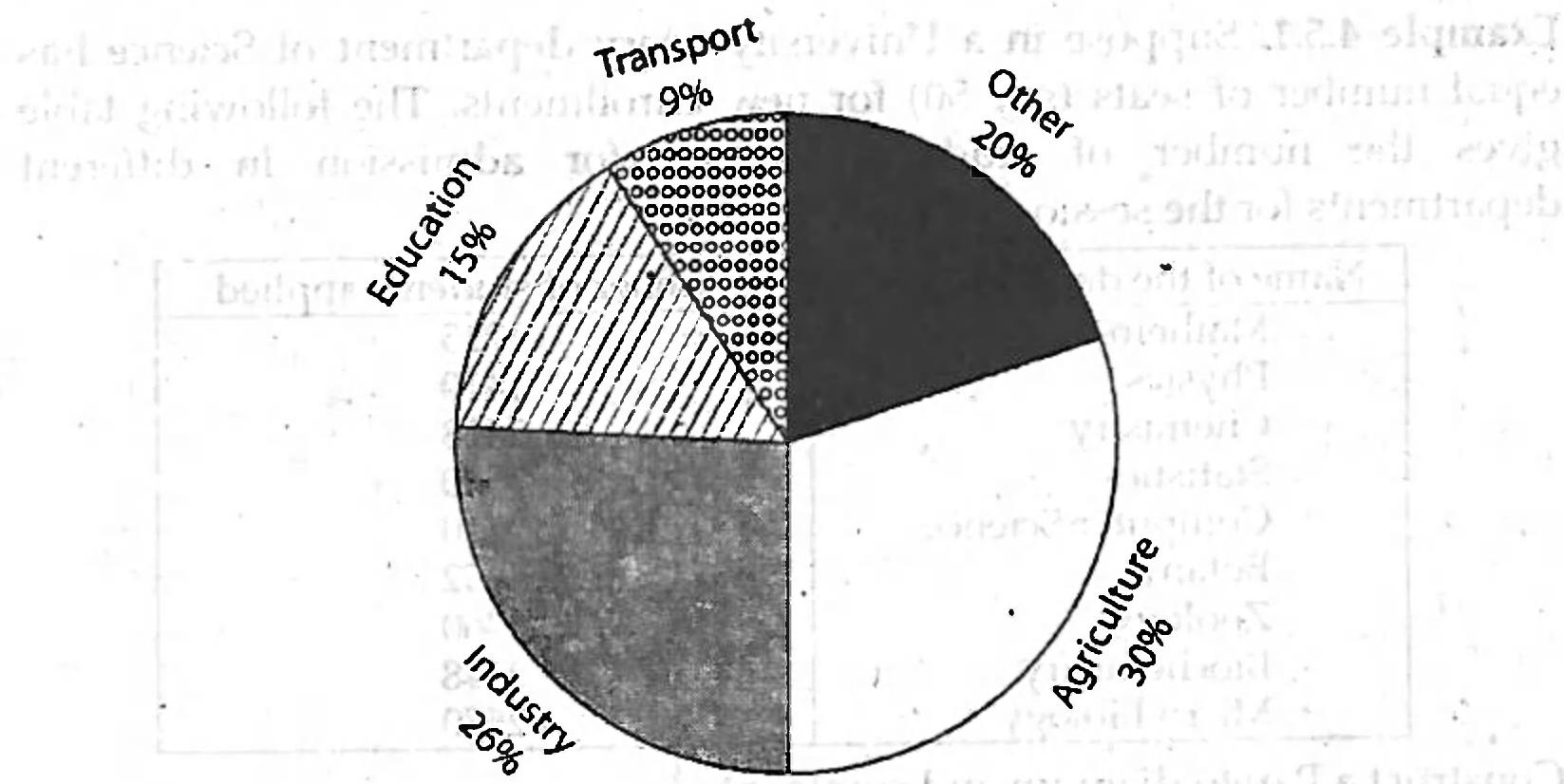


Fig. 4.4.2. Pie chart of expenditure of different sectors.

4.5. The Pareto Diagram

The graphical device for portraying categorical data that often provides more visual information than either the bar chart or the pie chart is the Pareto diagram. The Pareto diagram is a special type of vertical bar chart in which the categorized responses are plotted in the descending rank order of their frequencies and combined with a cumulative polygon on the graph. The main principle behind this graphical device is its ability to separate the 'vital few' from the 'trivial many' enabling one to focus on the important categories. Thus, the Pareto diagram achieves its greatest utility when the categorical variable of interest contains many categories. The Pareto diagram is widely used in analyzing process and product quality.

Construction of Pareto Diagram

- First different categories of the data are arranged in the descending rank order of their frequencies.

2. Then a vertical bar diagram is drawn on the rank ordered categories according to their percentage frequencies.
3. Construct a cumulative percentage frequency table with the rank ordered categories percentage frequency table.
4. Plot a point on the mid point of each bar according to the cumulative percentage frequency for each category.
5. Join the points by straight lines to get a cumulative polygon.
6. The resulting diagram thus obtained is called the Pareto Diagram.

Remark. The Pareto diagram is a very useful tool for presenting categorical data, particularly when the number of classification or grouping increases.

Example 4.5.1. Suppose in a University every department of Science has equal number of seats (say 50) for new enrollments. The following table gives the number of students applied for admission in different departments for the session 2007-2008.

Name of the departments	Number of students applied
Mathematics	1225
Physics	1410
Chemistry	2418
Statistics	1560
Computer Science	2710
Botany	2352
Zoology	2360
Biochemistry	2548
Micro-biology	2570

Construct a Pareto diagram and comment.

Solution. In order to obtain a Pareto diagram, a summary table is developed in which the subjects are arranged in descending order of their frequencies. Percentage frequency for each subject is computed for the ordered subjects and put them in the third column of the table. Cumulative percentage frequency for each subject of the ordered subjects is computed and put in the fourth column of the table. The results are shown in the table given below.

Subject	Frequency	Percentage frequency	Cumulative percentage frequency
Computer Science	2710	14.15	14.15
Micro-biology	2570	13.42	27.57
Biochemistry	2548	13.00	40.87
Chemistry	2418	12.62	53.49
Zoology	2360	12.32	65.81
Botany	2352	12.28	78.09
Statistics	1560	8.15	86.24
Physics	1410	7.36	93.60
Mathematics	1225	6.40	100.00
Total	19153		

In the construction of a Pareto diagram, the vertical axis on the left contains the frequencies or the percentage frequency. Here we put percentage frequency on the left vertical axis. The vertical axis on the right contains the cumulative percentage frequencies (from 100 on top to 0 on the bottom), and the horizontal axis contains the names of the ordered subjects according to their frequencies. The width of each bar and the space between two bars are kept equal. A point is put on the mid point of each bar corresponding to its cumulative percentage frequency. Now straight lines are drawn through the points to get the cumulative percentage frequency polygon. The resulting bar diagram with cumulative polygon is called the Pareto diagram and it is given in Fig. 4.5.1. It appears from the bar diagram that the most favorite subject to the admission seeking students is the computer science which is 14.15% and the least favorite subject is the mathematics which is 6.40% only. We see from the cumulative frequency polygon that 78.09% students applied the first six subjects of the nine subjects.

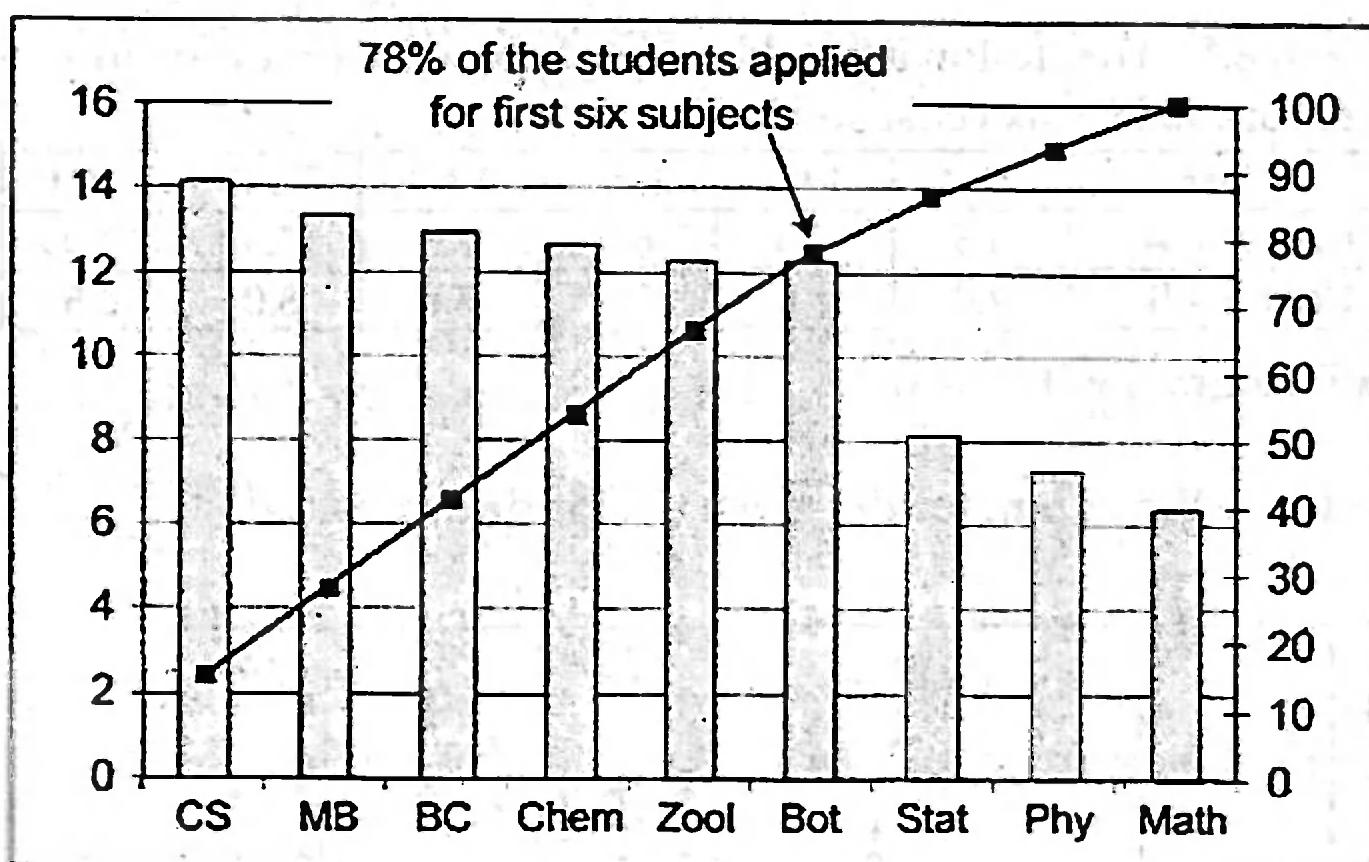


Fig. 4.5.1 Pareto Diagram for No. of Students.

~~A matched problem to solve.~~ A patient satisfaction survey conducted for a sample of 210 individuals discharged from a large urban hospital during the month of June led to a list of 384 complaints in the following categories.

Reason for complain	Number
Anger with other patients/visitors	13
Failure to respond to buzzer	71
Inadequate answers to questions	38
Lateness for tests	34
Noise	28
Poor food service	117
Rudeness of staff	62
All others	21
Total	384

- a. Form a Pareto diagram
 b. Which reasons of complaint do you think the hospital should focus on if it wishes to reduce the number of complaints? Explain.

4.6. Pictogram

Pictogram means presentation of data in the form of pictures. It is also known as picturegrams. It is quite a popular method used by governments and other organizations for information exhibition. Its main advantage is its attraction value. They stimulate interest in the information being presented.

News magazines are very fond of presenting data in this form. For example, in comparing the strength of the armed forces between U.S.A. and the Russia, they will simply show drawn pictures of soldiers where each picture may represent 10,000 soldiers. Similar comparison for missiles and tanks can also be done.

Example 4.6.1. The following table refers to number of people of two countries for the last six censuses:

Year	1961	1971	1981	1991	2001	2011
Country A	3.5	5.5	6.5	7.0	7.5	8.0
Country B	2.0	4.0	5.5	7.0	8.0	8.5

Draw a Pictogram with the data.

Solution. The Pictogram for the given data is drawn below:

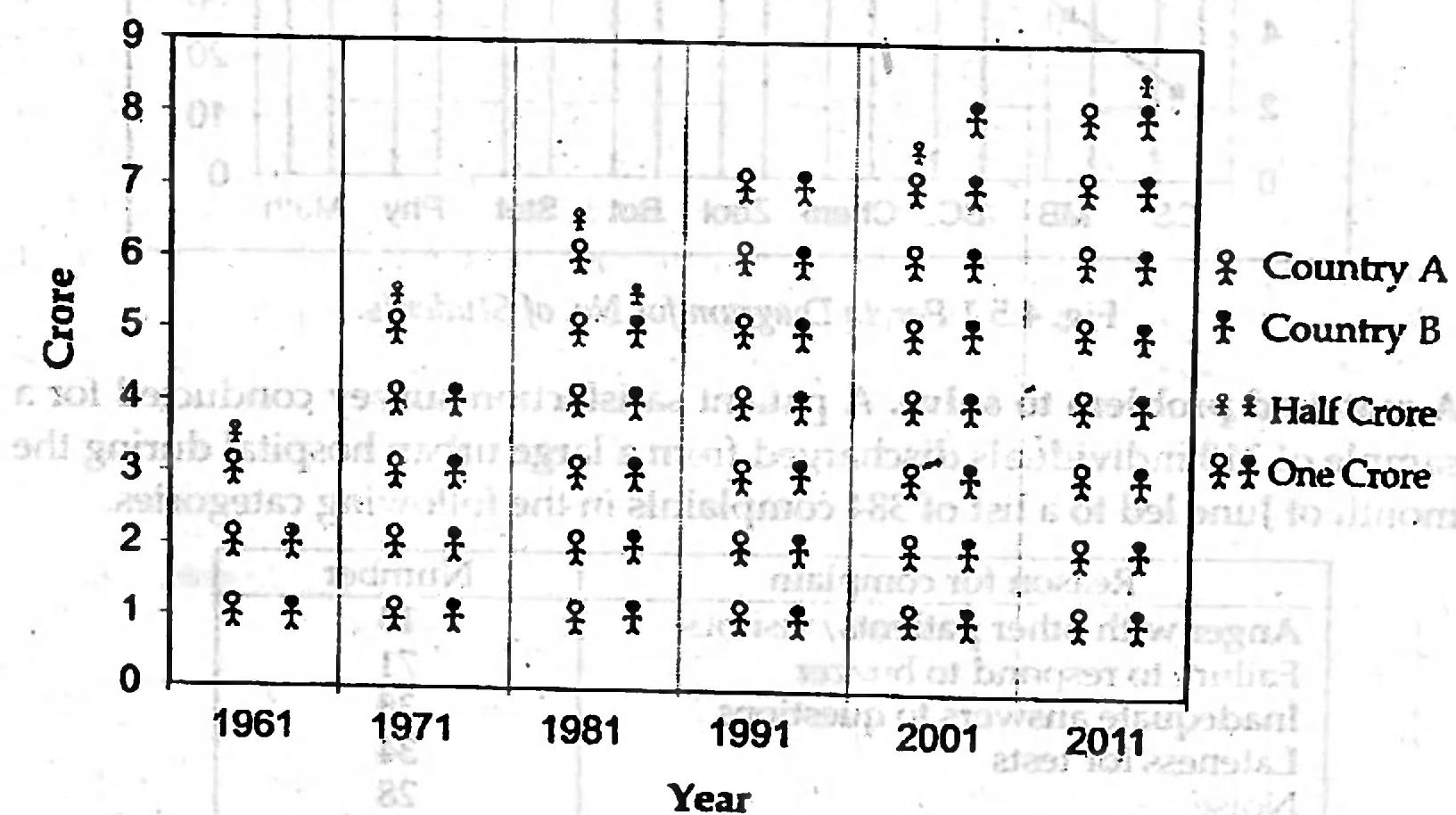


Fig. 4.6.1 Pictogram for population.

Merits

1. Pictogram is a very popular diagram to exhibit in a exhibition or fair.
2. Facts portrayed in pictorial form are generally remembered longer than facts presented in table or in non-pictorial chart.

Demerits

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

4.7. Condensing and Summarizing Quantitative Data

There are two types of quantitative data namely;

- (i) Discrete data and
- (ii) Continuous data

4.7.1. Frequency distribution of discrete data. We get discrete data from discrete variable and a discrete frequency distribution can be obtained from it. If the values of the discrete variable are finite and limited then each value of the variable can be considered as a class.

Frequency Distribution

Definition. Frequency distribution is a tabular summary showing the number of times each value of the variable occurs in the data.

Procedure of construction. Generally, a frequency distribution table consists of three columns.

1. The first column contains the classes. Here the number of classes will be the possible values of the discrete variable.
2. The second column contains tally marks. The number of observations in each class is denoted by tally marks.
3. The number of observations falls in each class is the class frequency. It is shown in the third column.

Now, we want to cite an example for discrete data:

Example 4.7.1. The management of a factory wants to know the family structure of the workers of his factory to educate their children. For this, they collected data on the number of children from the 45 workers of the factory. They are as follows:

2, 3, 5, 2, 1, 0, 3, 5, 3, 6, 4, 3, 5, 6, 1,
4, 4, 5, 4, 4, 1, 6, 2, 4, 5, 1, 4, 4, 2, 5,
4, 3, 5, 4, 4, 5, 3, 4, 4, 1, 6, 4, 5, 4, 4.

Construct a frequency table with the above data.

Solution. Here the number of children is a discrete variable X . It takes values 0, 1, 2, 3, 4, 5 and 6. The following steps are made to construct the frequency table:

1. The number of class will be 7 since the variable X can take only seven values. Number of children will be class heading.
2. The observations are read one by one and a tally mark is put (a tally mark is /) in the second column, against each value we read. The process will continue until we finish putting tally marks for all the values of the variable.
3. Now, we count the number of tallies in each class. This count is the frequency of that class.

The frequency distribution of the number of children of the 45 families is shown in table 4.7.1.

Table 4.7.1. Frequency distribution of number of children for 45 workers of a factory.

Number of children : X	Tally marks (# of families)	Frequency
0	/	1
1		5
2		4
3		6
4		16
5		9
6		4
Total		45

Table 4.7.1. gives us information about the distribution of the number of children of the 45 families of the workers of that factory. It is seen that 16 families have 4 children and only one family have no children. The picture will be clearer if we construct a relative frequency and percent frequency distribution of the data.

The relative frequency and percent frequency are calculated by the formulae:

$$\text{Relative frequency} = \frac{\text{Frequency of the value of a variable}}{\text{Total frequency}}$$

Percent Frequency = relative frequency \times 100. For example, the relative frequency for the value 5 is $9/45 = 0.2$. Percent frequency = $0.2 \times 100 = 20$.

Table 4.7.2. Relative frequencies and percent frequencies of 45 workers of a factory.

Number of children: X	Frequency	Relative frequency	Percent frequency
0	1	0.02	2
1	5	0.11	11
2	4	0.09	9
3	6	0.13	13
4	16	0.36	36
5	9	0.20	20
6	4	0.09	9
Total	45	1.00	100

It appears that 36% families have four children, 20% families have 5 children, 13% families have three children and 2% families have no children. That is most of the families have 3 to 5 children. Now the management of the factory will be able to recommend manageable size school building for the children of the workers.

A matched problem to solve. The managing director of a small factory wants to study the absent behavior of the workers of his factory. The following record gives the number of workers absent for last month:

0, 1, 5, 4, 3, 2, 2, 3, 5, 6,
5, 3, 4, 5, 4, 3, 3, 3, 4, 3,
4, 3, 4, 2, 1, 3, 3, 5, 2, 5,

Construct a frequency distribution with the above data and comment.

4.8. Construction of a Frequency Distribution for Continuous Data

It is obtained when the observations of a continuous variable are distributed among different classes of the variable. Actually, it is a table that divides the whole data into a small number of classes.

Important steps for constructing a frequency distribution table:

1. The number of classes depends on the range of the data. Range is the difference between the largest value and the smallest value of the data set.

$$\text{Range} = \text{largest value} - \text{smallest value}$$

2. Number of classes: Number of classes should not be too large or too small. As a general rule, the number of classes should range from 5 to 25. Another rule of thumb is that the number of classes should be around \sqrt{n} where n is the number of observations in the data. According to M.A. Struge's, the number of classes can be determined using formula

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

Actually, there is no hard and fast rule for the number of classes. It depends on the volume and nature of the data. However, it should be kept in mind that the number of classes should not be too more or too few.

3. Width of the class should be equal as far as possible. The width of the class is approximately the range divided by the number of classes:

$$\text{Width of class} = \text{range} / \text{number of classes}$$

In practice, the number of classes and the appropriate class width are determined by trial and error.

4. Tally Marks: Observations are counted and marked by tally marks.
5. Number of columns: Usually, there will be three columns in a frequency table: Class interval, tally marks and frequency.

Some more concepts in constructing frequency distribution for continuous data.

We shall explain all these concepts with the help of examples.

Class Limits. Each class of a frequency distribution has a lower value and an upper value. They are known as lower class limit and upper class limit. For example, take the class 10 - 20. The lowest value of this class is 10 and the highest value is 20. That is the lower class limit identifies the smallest possible value assigned to the class. The upper class limit identifies the largest possible value assigned to the class. Class limits must be chosen so that each observation of the data belongs to one and only one class.

There are two methods of classifying the data according to class intervals, namely

- (a) Exclusive method, and
- (b) Inclusive method.

(a) Exclusive method. When the class-intervals are so fixed that the upper limit of one class is the lower limit of the next class. In this case, upper limit of each class is excluded from the count in that class, the observation equal to upper limit is counted in the next class interval. Hence, it is known as the exclusive method of classification. A set of data classified by this method is given below:

Example 4.8.1. The following data relate to the audit-time of 20 clients:

10, 15, 20, 28, 13, 18, 24, 29, 12, 16,
23, 34, 14, 17, 22, 17, 21, 16, 18, 19

Table 4.8.1. Frequency distribution for the audit-time of 20 clients by exclusive method.

Audit time (hours)	Number of clients
10-15	4
15-20	8
20-25	5
25-30	2
30-35	1
Total	20

It is clear that the exclusive method ensures continuity of data in such a way that the upper limit of one class is the lower limit of the next class. The audit times between 10 to 14.99 hours are included in the class 10-15. But the audit time 15 hours would be included in the class 15-20. However, it is confusing to a layman who has no knowledge of statistics.

(b) Inclusive method. Under the inclusive method of classification, the upper limit of one class is included in that class itself. The frequency distribution of audit time data by inclusive method of classification can be shown as follows:

Table 4.8.2. Frequency distribution for the audit-time of 20 clients by inclusive method is.

Audit time (hours)	Number of clients
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1
Total	20

In the class 10-14, we include all the clients whose audit time is between 10 and 14 hours. Here 15 would be in the next class 15-19. It should be noted that both the inclusive and exclusive methods give us the same class frequencies.

Class boundaries. To ensure continuity of classes by inclusive method and to get rid of confusion from exclusive method we use class boundaries instead of class limits. The adjustment consists of finding the difference between the lower limits of one class with the upper limit of the previous class, dividing the difference by two, subtracting the value so obtained from all the lower limits and adding to all upper limits. The adjusted lower limits are called lower class boundaries and the adjusted upper limits are called upper class boundaries. This can be done by the formula as follows:

$$\text{Lower limit of one class} - \frac{\text{Upper limit of the previous class}}{2}$$

Correction term =

$$= \frac{d}{2}$$

For example, in table 4.8.2 the class 15–19 has lower limit 15 and the previous class 10–14 has upper limit 14. The correction term is $= \frac{15 - 14}{2} = 0.5$

To get class boundaries, we have to deduct 0.5 from the lower limits of all the classes and add up 0.5 to the all upper limits. Then the frequency distribution of table 4.8.3 will take the following form.

Table 4.8.3. Frequency distribution for the audit-time of 20 clients by inclusive method using class boundary is.

Audit time (hours)	Number of clients
9.5 - 14.5	4
14.5 - 19.5	8
19.5 - 24.5	5
24.5 - 29.5	2
29.5 - 34.5	1
Total	20

Remark. In order to avoid the ambiguity and to ensure the continuity, we have to take one more decimal point in the class boundaries than that exist in the data in such a way that the largest and the smallest values in the data set are included in the classes.

Sometimes, mid-value, relative frequency, cumulative frequency and percent cumulative frequency are also shown in a frequency table.

Mid-Point. In some applications, we want to know the mid-points of the classes of a frequency distribution for quantitative data. It is the value lying half-way between the lower and upper class limits/boundaries. It is considered as a representative value of the observations within that class. It is calculated by the following formula:

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

or, Mid-point =

$$\frac{\text{Upper limit of the class boundary} + \text{Lower limit of the class boundary}}{2}$$

For the purpose of further calculations in statistical work the mid-point of each class is considered to represent that class as a value of the variable.

Cumulative frequency. Sometimes one needs to know the answers to questions like 'how many workers of a factory earn more than Tk. 3000.00 per month' or 'how many workers earn less than Tk. 2500.00 per month'. To answer these questions it is necessary to know the cumulative frequencies. When frequencies are added, they are called cumulative frequencies. These frequencies are then listed in a table called a cumulative frequency table.

$$\text{Relative Cumulative frequency} = \frac{\text{cumulative frequency of a class}}{\text{Total frequency}}$$

Percent cumulative frequency = Relative Cumulative frequency $\times 100$.

Now we shall construct a frequency table with a continuous data.

Example 4.8.4. The management of a factory wants to know per month working pattern of the workers of their factory. In this connection, a survey was conducted on 50 workers of the factory. Following data give the number of hours worked per month of 50 workers of the factory:

140, 144, 187, 184, 87, 40, 122, 203, 148, 150,
 165, 133, 195, 151, 71, 94, 87, 42, 30, 62,
 103, 204, 162, 149, 79, 113, 69, 121, 93, 143,
 110, 175, 161, 157, 155, 108, 164, 128, 114, 178,
 130, 156, 167, 124, 164, 146, 116, 149, 104, 141.

Construct a frequency table using all the three methods.

Solution. In most of the cases in practice, we get data, discrete in nature, although they are in continuous. Here, we take the observations omitting decimal place and get the observations as integer. So, the data set is continuous here. We will construct frequency distributions by all the three methods: (a) Exclusive method, (b) Inclusive method and (c) Class boundary method. We take the following steps to construct the frequency distribution:

1. First we count the number of observations in the data set. It is $n = 50$.
2. The number of classes to be made is approximately $= \sqrt{50} \approx 7$.

According to Struge's formula, the number of class is

$$K = 1 + 3.3221 \log_{10} 50 = 1 + 3.322 \times 1.699 = 6.64 \approx 7$$

The number of classes should be 7 or around 7. Let us take it as 7.

3. Range of the data set is the difference between largest value and the smallest value. Thus,

Range = largest value - smallest value = 204 - 30 = 174

4. Width of each class = range / number of classes = 174 / 7 \approx 25

Once the number of classes and width of each class has been decided, the next step is to make classes

(a) **Classification by Exclusive method.** The classes will be 30 - 55, 55 - 80, 80 - 104, 105 - 129, 130 - 154, 155 - 179, 180 - 204 by this method. Here the observation 30 will be in the class 30 - 55 but 55 will be in the class 55-80 .

(b) **Classification by Inclusive method.** The classes will be 30 - 54, 55 - 79, 80 - 105, 105 - 130, 130 - 155, 155 - 180, 180 - 205 by this method. Here the observations both 30 and 54 in the class 30-54.

(c) **Classification by class boundary method.** To classify a value like 55 by exclusive method may create some problem, since both the intervals 30-55 and 55-80 contain 55. In order to avoid this kind of ambiguity, we classify the data by class boundary method stated before. We also use class boundary method to insure the continuity among the classes by inclusive method. The correction term is

$$\text{Correction term} = \frac{\text{Lower limit of the second class} - \text{Upper limit of the first class}}{2}$$

$$= \frac{55 - 54}{2} = 0.5.$$

We adjust the classes obtained by the inclusive method by deducting 0.5 from each lower limit and adding 0.5 to each upper limit of all the classes. Then the class boundaries of all the classes will be as follows:

$$29.5 - 54.5, 54.5 - 79.5, 79.5 - 104.5, 104.5 - 129.5,$$

$$129.5 - 154.5 - 179.5, 179.5 - 204.5.$$

We want to demonstrate all the three methods in a single table. Now we are ready to construct the frequency table. We make a table with 5 columns and 9 rows.

1. First we record the class interval of the seven classes by exclusive, inclusive and boundary methods in the first, second and third column respectively.
2. We read the observations one by one and put a tally mark in columns 2, 3 and 4 for each number. For example, first observation of row two is 165. For this put a tally mark against the class 155-180, 155-179 and 154.5-179.5 which are in the same row. We continue the process until putting tally marks for all the observations of the data set.
3. Now, we count the number of tallies in each class. This count is the frequency of that class. The frequencies of all the seven classes are

counted and put in column 5 of the table 4.8.4 For example; the number of tallies in class 155-180, 155-179 and 154.5-179.5 is 11. So the frequency of this class is 11. Actually, it is in the sixth class by the three methods.

Table 4.8.4. Frequency distribution of the number of hours worked by 50 workers.

Class Interval (Exclusive method)	Class interval (inclusive method)	Class interval (Class boundary method)	Tally Marks	Frequency
30 - 55	30 - 54	29.5 - 54.5		3
55 - 80	55 - 79	54.5 - 79.5		4
80 - 105	80 - 104	79.5 - 104.5		6
105 - 130	105 - 129	104.5 - 129.5		9
130 - 155	130 - 154	129.5 - 154.5		12
155 - 180	155 - 179	154.5 - 179.5		11
180 - 205	180 - 204	179.5 - 204.5		5
Total				50

The main advantage of using this summary table is that the major data characteristics become clearer than the original data. For example, 23 workers out of 50 work 130 to 180 hours per month. Only 3 workers work between 30 to 55 hours per month and 5 workers work between 180 to 205 hours per month.

On the other hand, the major disadvantage of this summary table is that the distribution of the individual values within a particular class interval is unknown without access to the original data.

It is seen that we get the same frequency distribution table by the three methods except the class interval.

Remarks.

1. It is advisable to construct frequency distribution by class boundary method for continuous variable.
2. Class boundary method is also important to represent quantitative data graphically which will be discussed latter on.
3. Once the frequency distribution is constructed, all the statistical analysis is performed using this grouped data. The important drawback of this method is that details information about the data is lost after grouping into different classes. Only the alternative is that we have to assume uniform frequency of all observations considered in a particular class. For example, suppose the frequency corresponding to the class interval 45-49 (inclusive) is 10. There are five observations considered in this interval, hence we have to assume that frequency of each observation is

$10/5 = 2$. That is why the mid-value of any class interval is considered as the representative value of that class.

To enhance the analysis of a frequency distribution, either the relative frequency distribution or the percent frequency distribution can be developed depending on proportions or percentages.

Example 4.8.5. Construct i) relative frequency, iii) percent relative frequency, iv) cumulative frequency, v) relative cumulative frequency and vi) percent relative cumulative frequency distribution by the following frequency distribution.

Class interval	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Frequency	3	4	6	6	12	11	5

Solution. The relative frequency of a class is calculated by the formula:

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{\text{Total frequency}}$$

$$\text{Percent frequency} = \text{Relative frequency} \times 100$$

When frequency of a class is added with all frequencies before that class is called cumulative frequencies of that class. The relative cumulative frequency and percent cumulative frequency are calculated by the following formulae:

$$\text{Relative Cumulative frequency} = \frac{\text{cumulative frequency of a class}}{\text{Total frequency}}$$

$$\text{Percent cumulative frequency} = \text{Relative Cumulative frequency} \times 100$$

The relative frequency, percent relative frequency, cumulative frequency, relative cumulative frequency and percentage cumulative frequencies are calculative by the above formula and shown in the third, fourth, fifth, sixth and seven columns of the table 4.8.5 given below.

Table 4.8.5.

Class interval	Frequency	Relative frequency	Percent frequency	Cumulative frequency	Relative cumulative frequency	Percent cumulative frequency
30 - 55	3	0.06	6	3	0.06	6
55 - 80	4	0.08	8	7	0.14	14
80-105	6	0.12	12	13	0.26	26
105-130	9	0.18	18	22	0.44	44
130-155	12	0.24	24	34	0.68	68
155-180	11	0.22	22	45	0.90	90
180-205	5	0.10	10	50	1.00	100

Total	50	1.00	100			
-------	----	------	-----	--	--	--

From the table, we can say that maximum 24% of the workers worked between 130 to 155 hours per month whereas 68% of the workers worked less than 180 hours per month and $(100 - 68) = 32\%$ of the workers worked 180 or more than 180 hours per month. We can get better interpretation from this table than the frequency distribution.

Ungrouped data. A set of raw data is called ungrouped data. That is, the data set on which no statistical work is done is called raw data.

Example 4.8.6. The prices in taka of 20 different brands of walking shoes are given below:

45, 70, 70, 55, 75, 73, 70, 65, 68, 60,
74, 83, 80, 58, 68, 85, 90, 64, 75, 82

This data set is called ungrouped data.

Grouped data. When a data set is given in the form of a frequency table then it is called grouped data.

Example 4.8.7. The frequency distribution given below gives the salary of 85 workers of a factory.

Salary per month in taka	3000-3500	3500-4000	4000-4500	4500-5000	5000-5500	5500-6000
Number of workers	10	14	25	18	13	5

This form of data is called grouped data.

In the next chapter, we shall describe both types of data numerically.

4.9. Stem and Leaf Display

Stem and leaf display is another form of presentation of quantitative data. It allows us to condense data, but still retain the individuality of the data. The idea was first proposed by Tukey (1977) and is based upon analogy to plants. Compare to other methods, stem and leaf plot is an easy and quick way of displaying data. The method is useful when data are not numerous. It is a simple and clear devise to construct a histogram-like picture of a frequency distribution. It allows us to show the range, concentration, presence of outliers, if any, and distribution of the data set at a glance.

Construction of Stem and leaf Plot. The stem of an observation is the leading digit or digits and the leaf of an observation is the trailing digit. All the values in the stem are listed in order in a column, a vertical line is drawn beside them and then all the corresponding leaf values are recorded for each

stem in a row, to the right of the vertical line. At last the digits of each leaf are arranged in ascending order to make the display more neat and clean.

Steps for construction of stem and leaf plot or display

1. Divide each observation into two parts: the stem and the leaf.
2. List the stem in a column, with a vertical line to their right.
3. For each observation, record the leaf portion in the same row as its corresponding stem.
4. Order the leaves from lowest to highest in each stem.
5. Mention the leaf unit to understand the actual observation.

Example 4.9.1. The prices in taka of 20 different brands of walking shoes are given below.

45, 70, 70, 55, 75, 73, 70, 65, 68, 60,
74, 80, 83, 58, 68, 85, 90, 64, 75, 82

Construct a stem and leaf plot to display the distribution of the data.

Solution. To create the stem and leaf plot, we divide each observation between the ones and the tens place. The digit to the left is the stem and the digit to the right is the leaf. For example, 45 is a value in the data, we put 4 as the stem and 5 as the leaf:

Stem	Leaf
4	5

The stems along with the leaves for each of the 20 observations are shown below:

Stem	Leaf
4	5
5	5 8
6	5 8 0 8 4
7	0 0 5 3 0 4 5
8	0 3 5 2
9	0

7 in stem and 5 in leaf means, the observation is 75.

Now we arrange the digits of each leaf in ascending order. After reordering the leaves, the stem – leaf plot will be as follows:

Stem	Leaf
4	5
5	5 8
6	0 4 5 8 8
7	0 0 0 3 4 5 5
8	0 2 3 5
9	0

By stem and leaf plot one can easily find the relative position of the values of the variable. It is seen that the lowest price of the walking shoe is taka 45 where as the highest price is taka 90. Taka 70 is the most popular price of the shoe.

Remarks. By stem and leaf plot display one can easily find the position measure such as median, quartiles, deciles, percentiles, even mode, range and quartile deviation of a distribution which will be discussed later on.

Example 4.9.2. The following data represent the amount of insurance (in units of thousand taka) purchased by 30 people from an insurance company in a given week:

31, 44, 51, 35, 76, 84, 110, 50, 56, 61,
40, 48, 61, 85, 90, 92, 40, 65, 120, 125,
100, 105, 115, 70, 77, 120, 75, 80, 92, 115

Construct a stem and leaf plot to display the data.

Construction. Let us construct the stem and leaf display of the above data. First we divide the observations into two parts. The last digit of each observation is leaf and the first one or two digits of each observation is the stem. For example, 31 and 110 are two values of the data set. The stems and leaves for the two values will be as follow:

Stem	Leaf
3	1
11	0

The stems along with the leaves for each of the 30 observations are shown below

Stem	Leaf
3	1 5
4	4 0 8 0
5	1 0 6
6	1 1 5
7	6 0 7 5
8	4 5 0
9	0 2 2
10	0 5
11	0 5 5
12	0 5 0

Here, 11 in stem and 5 in leaf means, the observation is 115.

Now we arrange the digits of each leaf in ascending order. After reordering the leaves, the stem and leaf display will be as follows.

Stem	Leaf
3	1 5
4	0 0 4 8
5	0 1 6
6	1 1 5
7	0 5 6 7
8	0 4 5
9	0 2 2
10	0 5
11	0 5 5
12	0 0 5

Comment. The distribution pattern is more or less uniform.

Example 4.9.3. The following data represent the lives of 40 similar car batteries recorded to the nearest tenth of a year:

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6	3.4	1.6
3.1	3.3	3.8	3.1	4.7	3.7	2.5	4.3	3.4	3.6
2.9	3.3	3.9	3.1	3.3	3.1	3.7	4.4	3.2	4.1
1.9	3.4	4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Construct a stem and leaf plot to display the data.

Solution. Here we first divide each observation into two parts. The stem for each observation is the digit left of the decimal point, while the digit right of the decimal point is the leaf. For example, 2.2 is a value of the data set. The stem and leaf of the value 2.2 is

Stem	Leaf
2	2

The stems along with the leaves for each of the 30 observations are shown below:

Stem	Leaf
1	6 9
2	2 6 5 9 6
3	5 2 7 0 4 1 3 8 1 7 4 6 3 9 1 3 1 7 2 4 8 2 9 0 5
4	1 5 7 3 4 1 7 2

Now we arrange the digits of each leaf in ascending order. After reordering the leaves, the stem and leaf display will be as follows:

Stem	Leaf
1	6 9
2	2 5 6 6 9
3	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4 5 5 6 7 7 7 8 8 9 9
4	1 1 2 3 4 5 7 7

4.10. Graphical Representation of Quantitative Data

Important graphs for representing frequency distribution of quantitative data are:

- (1) Dot plot,
- (2) Histogram,
- (3) Frequency polygon,
- (4) Frequency Curve,
- (5) Ogive polygon
- (6) Ogive curve and
- (7) Line graph of time series data
- (8) Scatter diagram

Now these graphs will be discussed one by one.

4.10.1. Dot Plot. One of the simplest graphical representations of numerical data is a dot plot. The horizontal X-axis shows the range the values for the observations. Each observation is represented by a dot placed above value of the variable.

Example 4.10.1. The marks obtained by 28 students in class test are as follows.

10, 11, 11, 15, 16, 16, 17, 18, 19, 19, 21, 23, 24, 25
 12, 13, 14, 14, 17, 15, 17, 18, 18, 19, 19, 20, 20, 21,

Construct a dot plot.

Solution. It is seen that the lowest value of the data set is 10 and the highest value of the data set is 25. Now plot the values of the variable along the X-axis and then plot each observation above the value of the variable. The fig. 4.10.1. is the dot plot of the above data set.

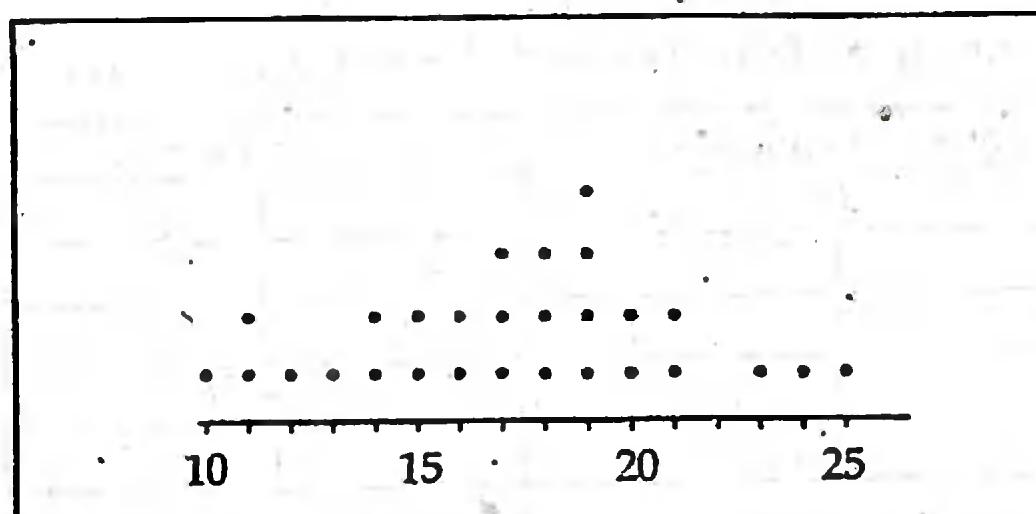


Fig. 4.7.1 Dot plot for marks

The four dots located above 19 on the X-axis indicate that there are four observations with a value of 19. Dot plots show the details of the data and are useful for comparing the distribution of the data for two more variables.

4.10.2. Histogram. Histogram is one of the most popular and widely used graphical methods for representing a frequency distribution. In this type of representation the given data are plotted in the form of a series of rectangles.

Class intervals or class boundaries are marked along the X-axis and the frequencies along the Y-axis according to a suitable scale. A histogram is constructed from a frequency distribution of grouped data.

In case of equal class interval, the height of each rectangle is equal to its frequency. But in case of unequal class interval, the area of each rectangle is proportional to its frequency. Each rectangle is joined with the other and the blank spaces if any between the rectangles would mean that the category is empty and there are no values in that class interval. The histogram is particularly appropriate when the variable is continuous. A discrete variable is also treated as a continuous variable while constructing histogram.

We can construct histogram in three different ways depending on the nature of the data.

- (a) Histogram for discrete frequency distribution.
- (b) Histogram for continuous frequency distribution with equal class interval.
- (c) Histogram for continuous frequency distribution with unequal class interval.

Histogram for discrete frequency distribution. In this case, the class intervals or values of the variable taken as class intervals are discrete in nature. So, first thing is to convert the class intervals into class boundaries to make them continuous. Then class boundaries are plotted along the X-axis and the frequency over each class boundary is plotted along the Y-axis. Now we shall cite one example.

Example 4.10.2. The following frequency distribution refers to the number of children of 45 workers of a factory:

Frequency distribution of number of children for 45 workers of a factory.

Number of children : X	Frequency
0	1
1	4
2	5
3	6
4	16
5	9
6	4
Total	45

Construct a histogram with the above frequency distribution.

Solution. Here the variable as well as the frequency distribution is discrete. Each value of the variable is considered as a class interval. So the class intervals are discrete. So we have to make each value of the variable into class boundary. The difference between two classes is 1. Here $d = 1/2 = 0.5$. To get the class boundary for each class, we have to subtract 0.5 from each value of the variable to get the value of the lower boundary and add 0.5 to each value of the variable to get the upper boundary of that class. For example,

$$\text{Lower boundary for the first class} = 0 - 0.5 = -0.5$$

$$\text{Upper boundary for the first class} = 0 + 0.5 = 0.5$$

$$\text{Similarly, lower boundary of the last class} = 6 - 0.5 = 5.5$$

$$\text{Upper boundary of the last class} = 6 + 0.5 = 6.5$$

Number of children: X	Class boundary	Frequency
0	-0.5 - 0.5	1
1	0.5 - 1.5	4
2	1.5 - 2.5	5
3	2.5 - 3.5	6
4	3.5 - 4.5	16
5	4.5 - 5.5	9
6	5.5 - 6.5	4

Now we plot the all the class boundaries along the X-axis and the corresponding class frequencies along the Y-axis to get the required histogram. The histogram of the above frequency distribution will take the following form:

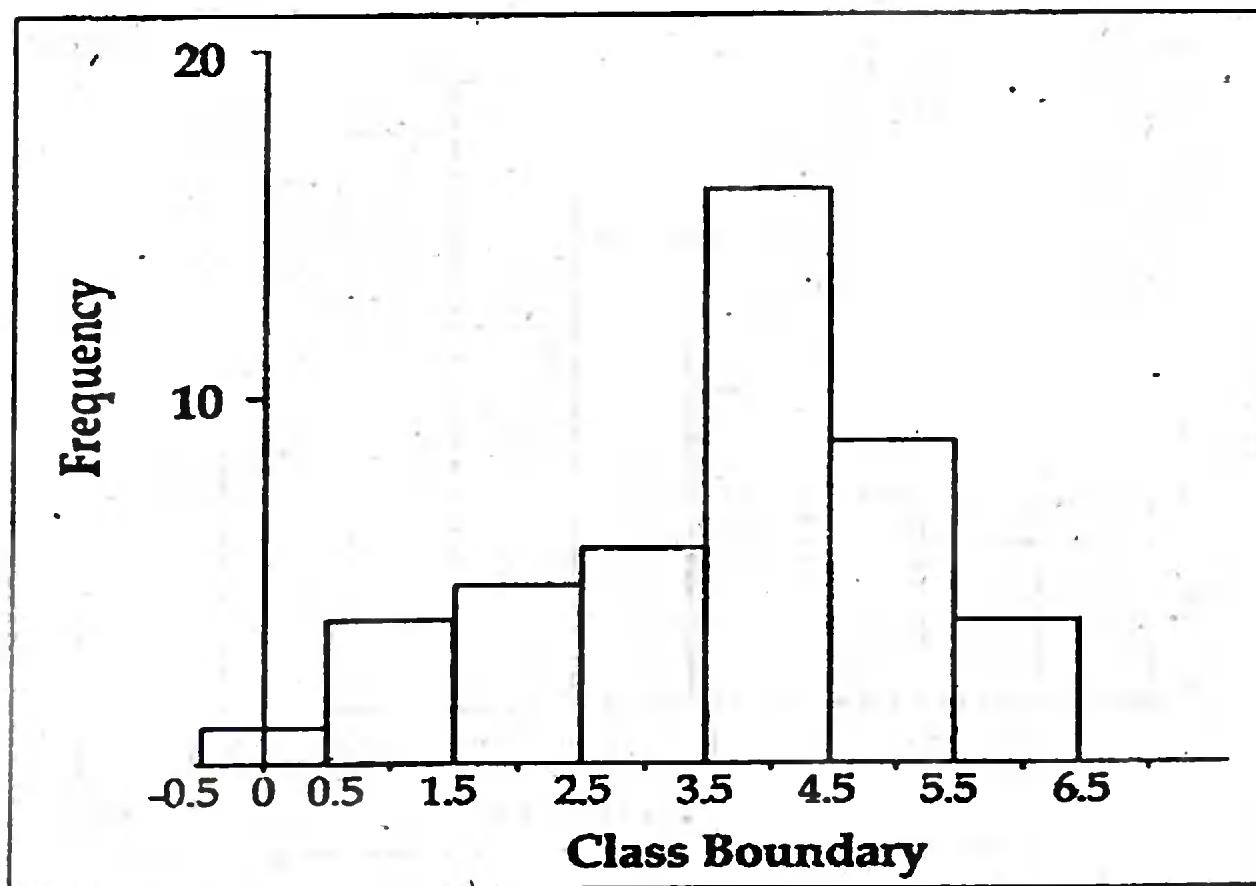


Fig 4.9.2. Histogram for number of Children.

(ii) Histogram for continuous data with equal class interval. In this case, the class intervals or class boundaries are plotted along the X-axis and the corresponding frequencies are plotted along the Y-axis and construct adjacent rectangles to get the required histogram.

Example 4.10.3. The following frequency distribution relate to the number of hours worked per month of 50 workers of a factory:

Frequency distribution of the number of hours worked per month of 50 workers of a factory.

Class interval (number of hours worked)	Frequency (number of workers)
30 – 55	3
55 – 80	4
80 – 105	6
105 – 130	9
130 – 155	12
155 – 180	11
180 – 205	5

Construct a histogram with the above data.

Solution. Here class intervals as well as the frequency distribution are continuous and the class intervals are equal. So the class intervals are plotted along the X-axis and the corresponding frequencies for each class are plotted along the Y-axis and construct the adjacent rectangles to get the required histogram. Figure 4.9.3 shows at glance how the distribution of the number of hours worked per month of 50 workers of a factory.

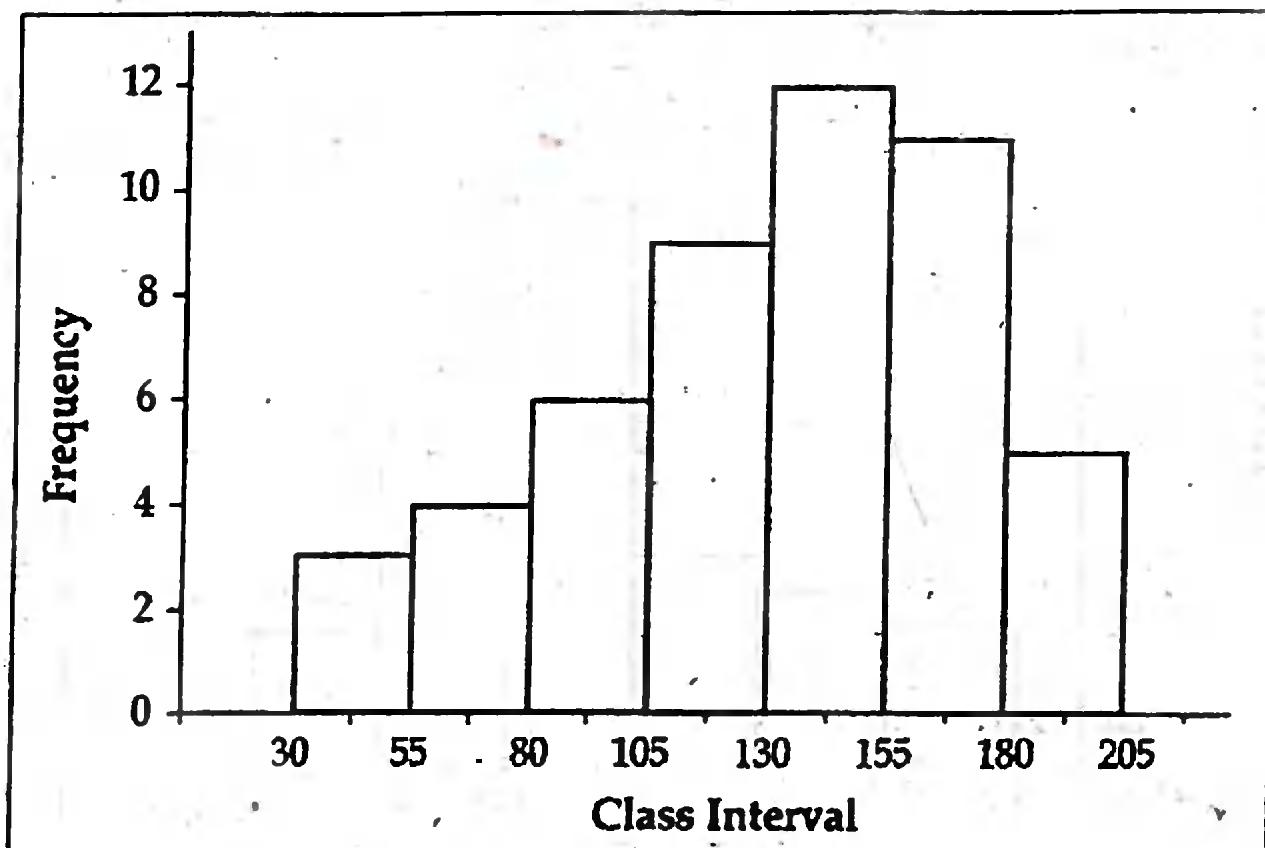


Fig. 4.9.3. Histogram for number of hours worked.

Histogram of continuous data with unequal class interval. When class intervals are unequal, the frequencies must be adjusted before constructing the histogram. Since the area of rectangle over a class is proportional to its frequency, the height of a rectangle is equal to the frequency divided by its width of the interval. This is sometimes called frequency density. Now class intervals are plotted along the X-axis and the height of rectangles are plotted along the Y-axis and construct adjacent rectangles to get the required histogram.

Example 4.10.4. The following data give the frequency distribution of weekly wage in Taka of 100 workers of a factory:

Weekly wage (in Taka)	Number of workers	Weekly wages (in Taka)	Number of workers
110-115	7	130-140	12
115-120	19	140-160	14
120-125	27	160-190	6
125-130	15		

Draw a histogram with the above data.

Solution. Here the width of class intervals unequal, varying from 5 to 30. The widths of the class intervals are shown in column 3 of table 4.8 given below. Now to construct histogram, heights of the rectangles are obtained by dividing the frequency of each class by its corresponding class widths. These values are shown in column 4 in table 4.9.3.

Table 4.9.3. Frequency distribution with class width and height of the rectangle.

Class interval	Class frequency	Class width	Height of rectangle
110 -115	7	5	1.4
115- 120	19	5	3.8
120 -125	27	5	5.4
125- 130	15	5	3
130 -140	12	10	1.2
140- 160	14	20	0.7
160 -190	6	30	0.2

Now class intervals are plotted along the X-axis and the heights of the rectangles are plotted along the Y-axis and construct adjacent rectangles to get the required histogram. The resulting histogram is shown in figure 4.9.3. shown below:

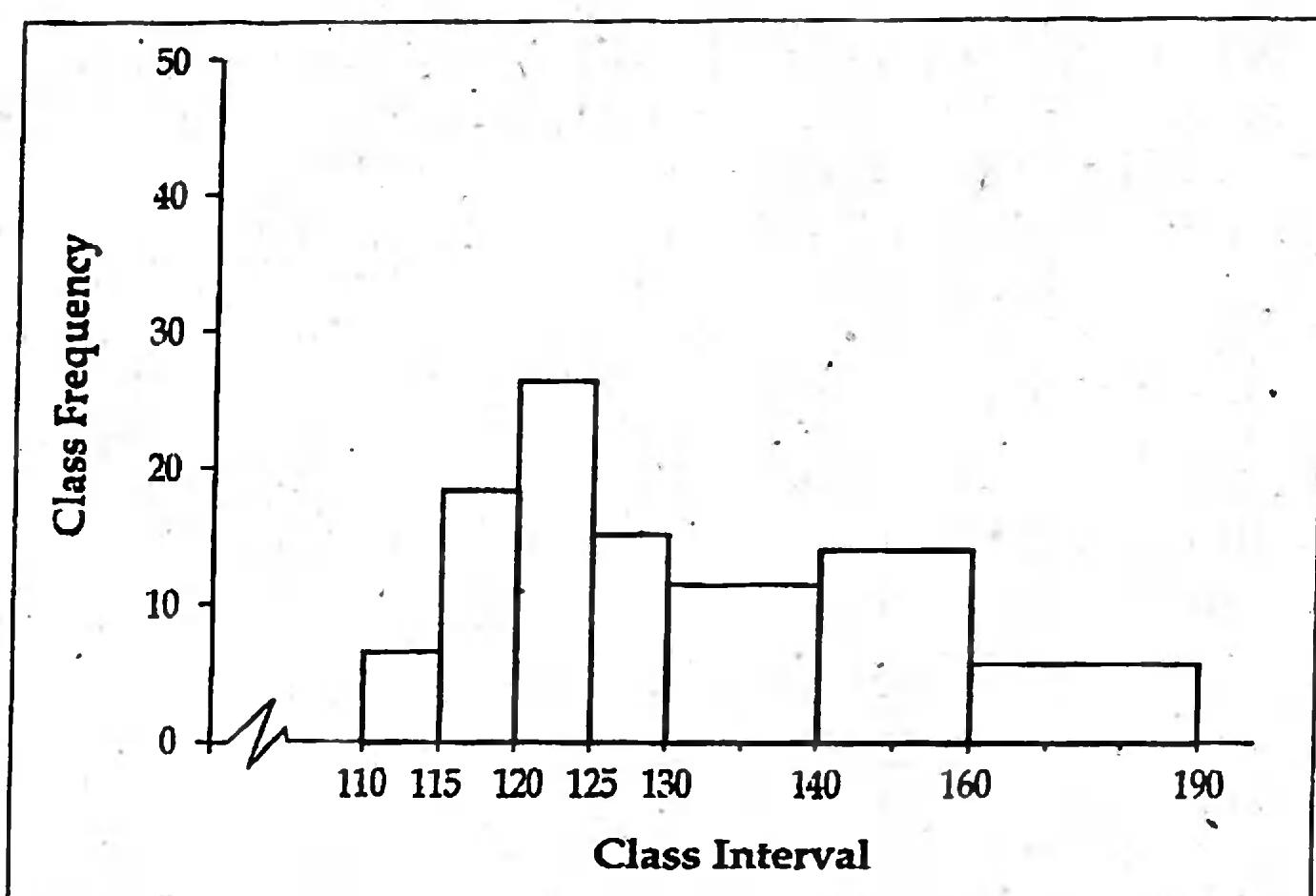


Fig. 4.9.4. Histogram for weekly wage of workers

Remarks:

1. Since the area of rectangle for each class interval is proportional to its relevant frequency; one can always construct histogram by taking class intervals along the X-axis and the heights of the rectangles along the Y-axis both for equal and unequal class intervals.
2. Since histogram is a graphical representation of frequency distribution for quantitative data, one can take frequency of each class along the Y-axis in case of equal class interval.
3. One can also construct histogram with the relative frequency or percent frequency in case of equal class interval. They also give the same type of graphical representation. Only the scale will differ.

4.10.3 Difference between Histogram and bar diagram. Although a bar diagram and a histogram are very similar to look at, but they are quite different and serve distinct purposes. The important differences are :

- (i) A bar diagram is a diagram whereas a histogram is a graph.
- (ii) A histogram is used for representing a frequency distribution for quantitative data but a bar diagram is used for representing a frequency distribution for qualitative data only.
- (iii) In a histogram the area of the rectangle is proportional to the relevant frequency, whereas in a bar diagram it is the height of the bar that counts.

- (iv) A bar diagram is one -dimensional figure, whereas a histogram is two-dimensional figure.
- (v) The rectangles in a histogram are all adjacent but the spacing of bars in a bar diagram is quite arbitrary.

4.11. Histogram and Stem-leaf Display

A histogram is constructed from a frequency distribution. Whereas a frequency distribution can easily be constructed from a stem and Leaf plot by counting the leaves belonging to each stem and noting that each stem defines a class interval. That is a stem-leaf plot contains the same information of a frequency distribution with its histogram. Moreover, stem-leaf display preserves the identity of the individual observations, which are absent both in frequency distribution and histogram. Actually it is a combined tabular and graphical display of statistical data, whereas histogram is a graphical representation of statistical data. A stem-leaf display looks like a horizontal histogram. Stem-leaf display turns out to be a usual histogram if the plots are rotated counterclockwise through an angle of 90° , a stem-leaf plot contains the same information of a frequency distribution with its histogram. Histogram is used for finding mode of a frequency distribution whereas stem and leaf plot can be used for finding all the position measures such as median, quartiles, deciles, percentiles and also mode.

4.12. Frequency Polygon

Frequency polygon is a graph used to represent a frequency distribution. Since it is a polygon, it has more than four sides. It is particularly effective in comparing two or more frequency distributions.

The values of discrete variable or the mid-values of class intervals are plotted along the X-axis and corresponding frequencies are plotted along the Y-axis. The latter points are then joined by straight lines, thus forming with the X-axis a polygon called frequency polygon. The frequency polygon should be brought down at each end to the X-axis by joining it to the mid-value of the next outlying interval.

There are two ways in which a frequency polygon may be constructed.

- (i) Frequency polygon from histogram
- (ii) Frequency polygon from frequency distribution.
- (i) Frequency polygon from histogram. First a histogram is to be constructed from the frequency distribution. Then midpoints of the upper horizontal side of each rectangle are joined by a straight line with the adjacent rectangle. The figure so formed is called frequency polygon.

Both the ends of the polygon are then extended to the base line. This extension is made with the object of making the area under polygon equal to the area under the corresponding histogram.

- (ii) **Frequency polygon from frequency distribution.** First mid-values of various class intervals are calculated. These mid-values are plotted along the X-axis and the frequency corresponding to each mid-values is plotted. Then all the points are joined by straight. The figure thus obtained is called the frequency polygon. The figure thus obtained by this method is the same as the other method. The only difference is that here we do not have to construct a histogram.

Frequency polygon can also be constructed for discrete frequency distribution. The different values of the variable are plotted along the X-axis and then the frequency corresponding to each value of the variable is plotted along the Y-axis. The points are then joined by straight lines and the resulting graph is called a frequency polygon.

Example 4.12.1. The following frequency distribution refers to the number of children of 45 workers of a factory:

Number of children : X	Frequency
0	1
1	4
2	5
3	6
4	16
5	9
6	4
Total	45

Construct a frequency polygon with the above frequency distribution.

Solution. Here the number of children is variable. The different values of the variable are plotted along the X-axis and then the frequency corresponding to each value of the variable is plotted along the Y-axis. The points are then joined by straight lines and the graphs so obtained is called frequency polygon. It is shown in figure 4.12.1.

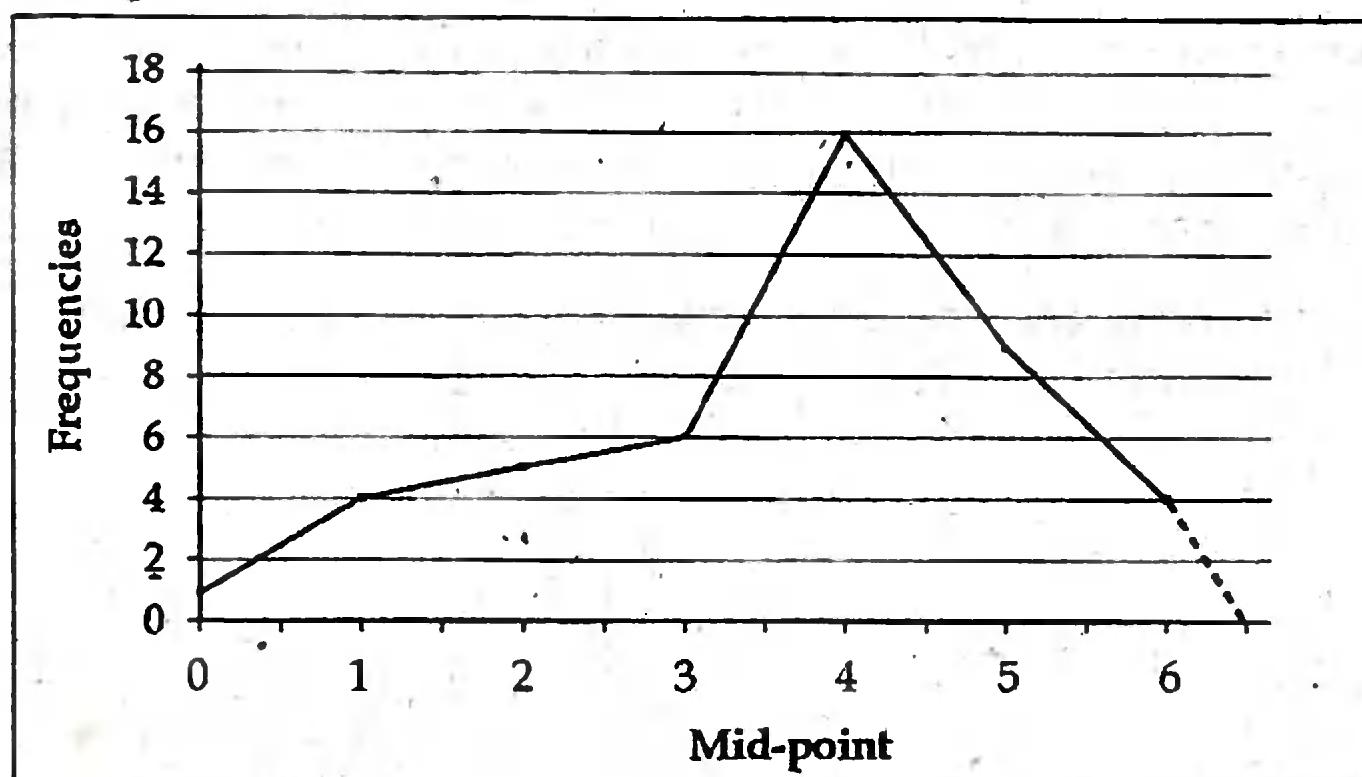


Fig. 4.12.1. Frequency Polygon for number of children.

Comment. It is clearly evident from the polygon that most of the workers' families have 4 children and only 1 family has no child.

Example 4.12.2. The following frequency distribution relates to the number of hours worked per month of 50 workers of a factory:

Class interval (number of hours worked)	Frequency (number of workers)
30 – 55	3
55 – 80	4
80 – 105	6
105 – 130	9
130 – 155	12
155 – 180	11
180 – 205	5

Construct

- (i) Histogram, frequency polygon and frequency curve from the above frequency distribution on a same graph.
- (ii) Frequency polygon and frequency curve on the same graph.

Solution. (i) First we construct a histogram by plotting class interval along the X-axis the frequency of each class along the y-axis. Then construct adjacent rectangles over the class intervals. The resulting graph gives us the required histogram and the graph looks like as in figure 4.12.2.

a) **Frequency polygon from histogram.** The frequency polygon is obtained by joining the mid-points of upper horizontal sides of each rectangle of the histogram by straight lines. It is shown in figure 4.12.2 with the histogram.

b) Frequency curve from histogram. Frequency curve is obtained by joining the mid-points of upper horizontal sides of each rectangle of the histogram by a smooth free curve. It is also shown in the same figure 4.12.2 with the histogram and frequency polygon.

Frequency polygon and frequency curve drawn from the histogram is shown below in figure 4.12.2.

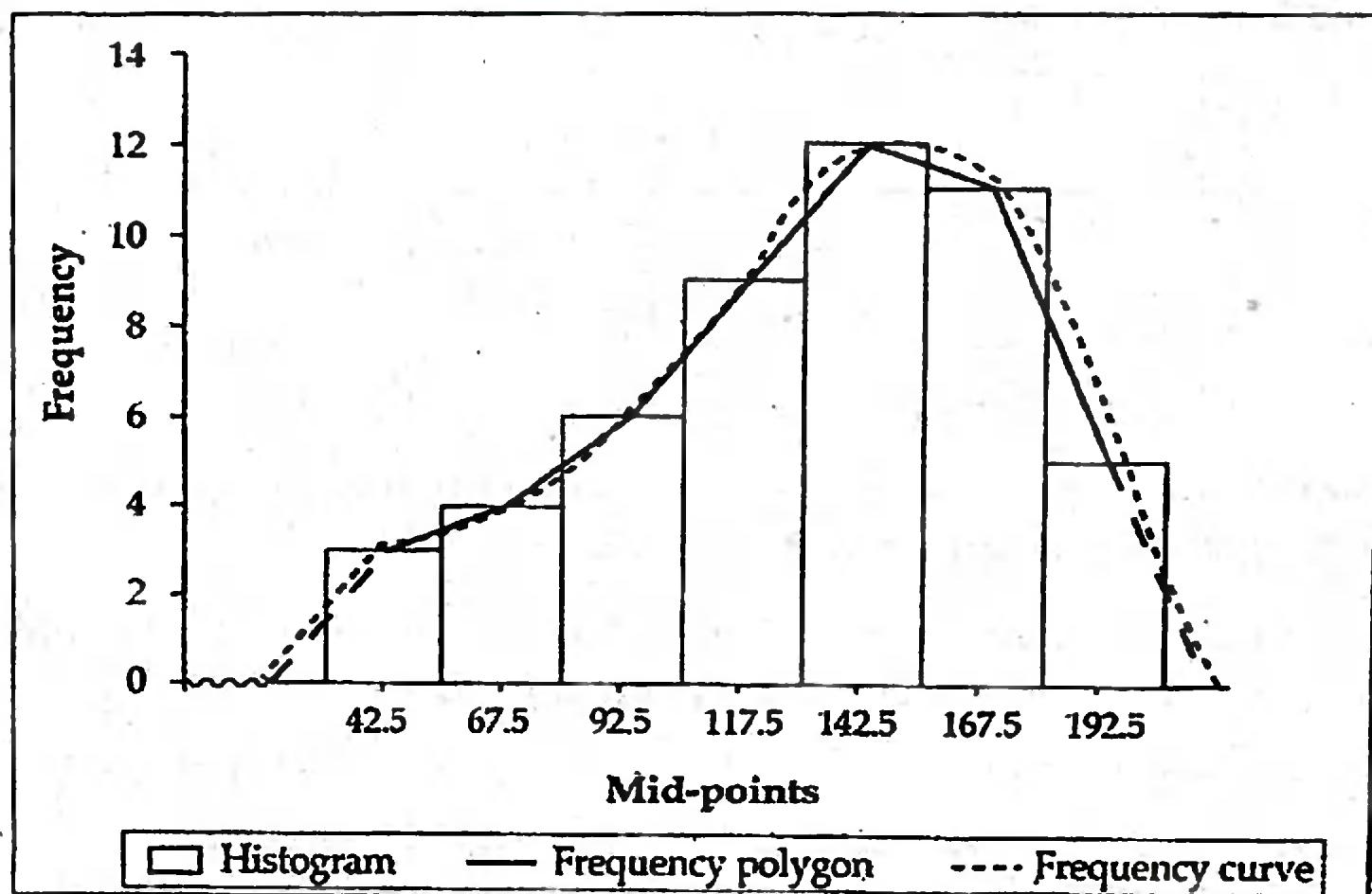


Fig. 4.12.2. Frequency polygon from histogram.

(ii) Frequency polygon and frequency curve from the frequency distribution. For drawing frequency polygon and frequency curve, we require mid-points of the classes and their corresponding class frequencies. For this we need the following frequency table:

Class interval (Number of hours worked)	Frequency (Number of workers)	Mid-points
30-55	3	42.5
55-80	4	67.5
80-105	6	92.5
105-130	9	117.5
130-155	12	142.5
155-180	11	167.5
180-205	5	192.5

(ii) Now we plot mid-points along the X-axis and the frequency along the Y-axis. Plot points directly above the mid-points at a height corresponding to the frequency of the class. Classes of zero frequency are added at each end of the frequency distribution. The frequency polygon is obtained by joining all the points by straight lines and is shown in figure 4.12.3.

(a) Frequency curve from the frequency polygon. First we draw a frequency polygon by plotting the mid-points along the X-axis and the corresponding frequencies above the mid-points by plotting points. A frequency curve is obtained by joining the plotted points by a free hand curve. It is shown in figure 4.12.3 with the frequency polygon.

The frequency polygon and frequency curve drawn from the distribution is as follows:

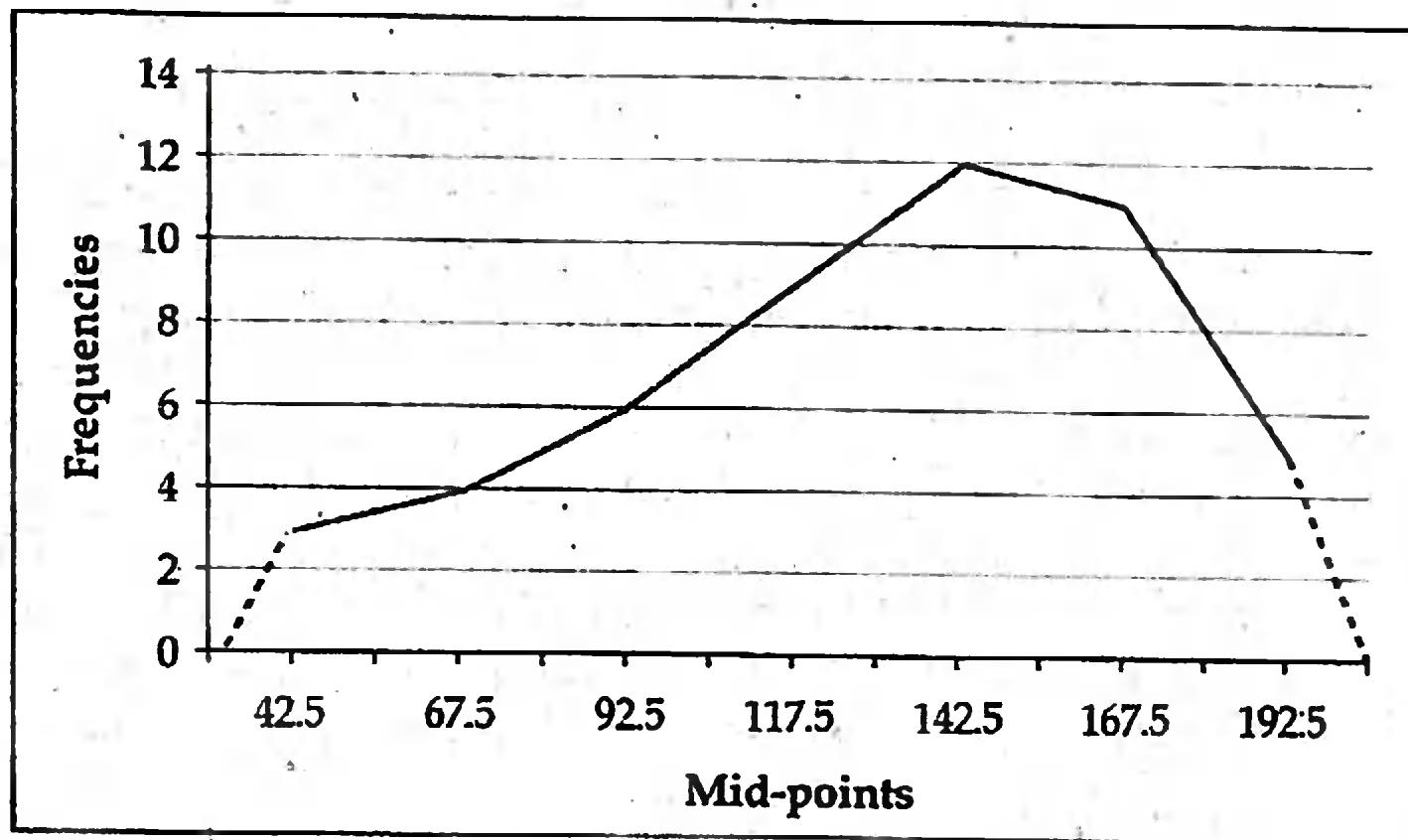


Fig. 4.12.3. Frequency polygon from frequency distribution.

Histogram and Frequency polygon. Histogram and frequency polygon both are used for representing a frequency distribution. If the variable under consideration is continuous, the histogram is decidedly superior; on the other hand if the variable is discrete, frequency polygon is to be preferred. Frequency polygon is used to compare two or more distributions, whereas histogram is used to present a single distribution.

4.13. Frequency Curve

A frequency curve is obtained by drawing a smooth freehand curve through the various points of a polygon. That is when a frequency polygon is smoothed; the resulting curve is called a frequency curve.

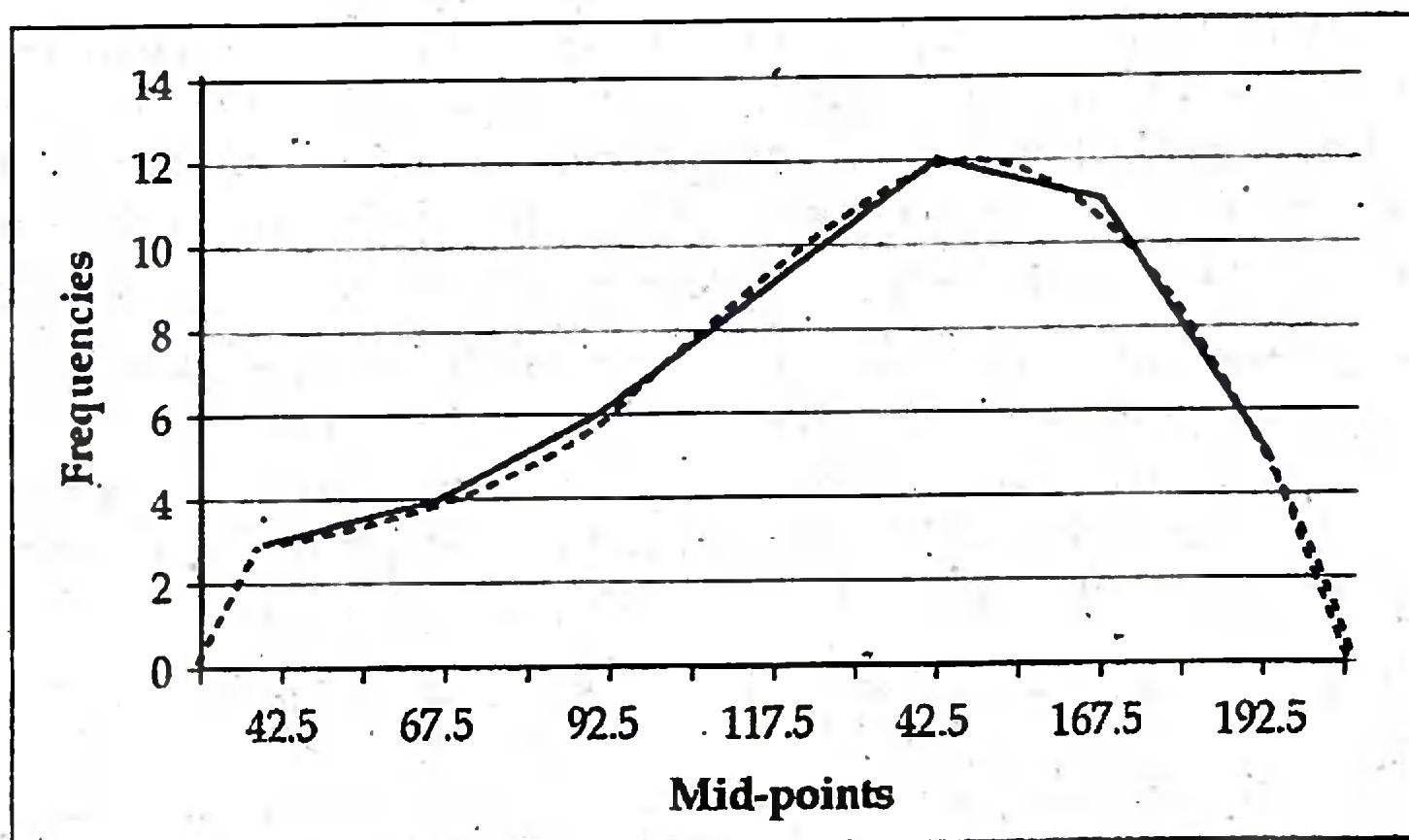


Fig. 4.12.4. Frequency curve and frequency polygon.

The object of drawing a smooth frequency curve is to eliminate as far as possible all accidental variations that might be present in the data. So for drawing a frequency curve it is necessary to first draw a frequency polygon and then smooth it out. As we know frequency polygon can be obtained from a histogram or by plotting the frequencies at the mid-values of the class-intervals. The smoothing of the polygon cannot be done properly without a histogram. Hence, it is desirable to draw a histogram first, then a polygon and lastly to smooth it to obtain the frequency curve. The curve is extended to the mid points of the class-intervals just outsides the histogram in both ends. The area under the curve should represent the total number of frequencies in the entire distribution.

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

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

Example 4.13.1. The following data refer to the length of service in years of 124 employees of a factory:

Length of service (in years)	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35
No. of Employees	6	12	24	35	26	15	6

Draw a histogram and a frequency polygon with the above data.

Solution. Table for constructing graphs.

Class Interval (Length of Service in years)	Frequency (Number of Employees)	Mid-points
0-5	6	2.5
5-10	12	7.5
10-15	24	12.5
15-20	35	17.5
20-25	26	22.5
25-30	15	27.5
30-35	6	32.5

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

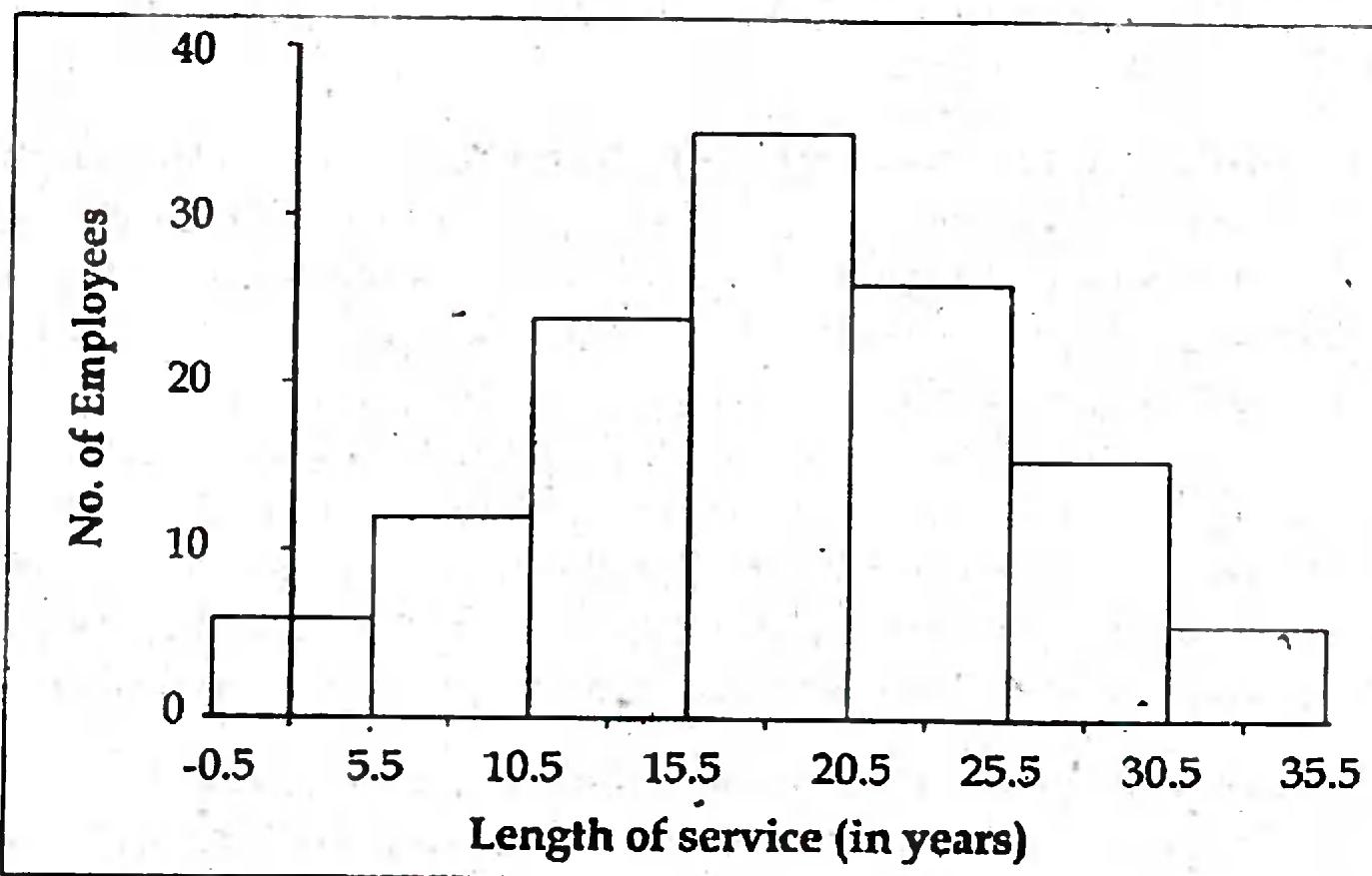


Fig. 4.13.1. Histogram for length of service of employees.

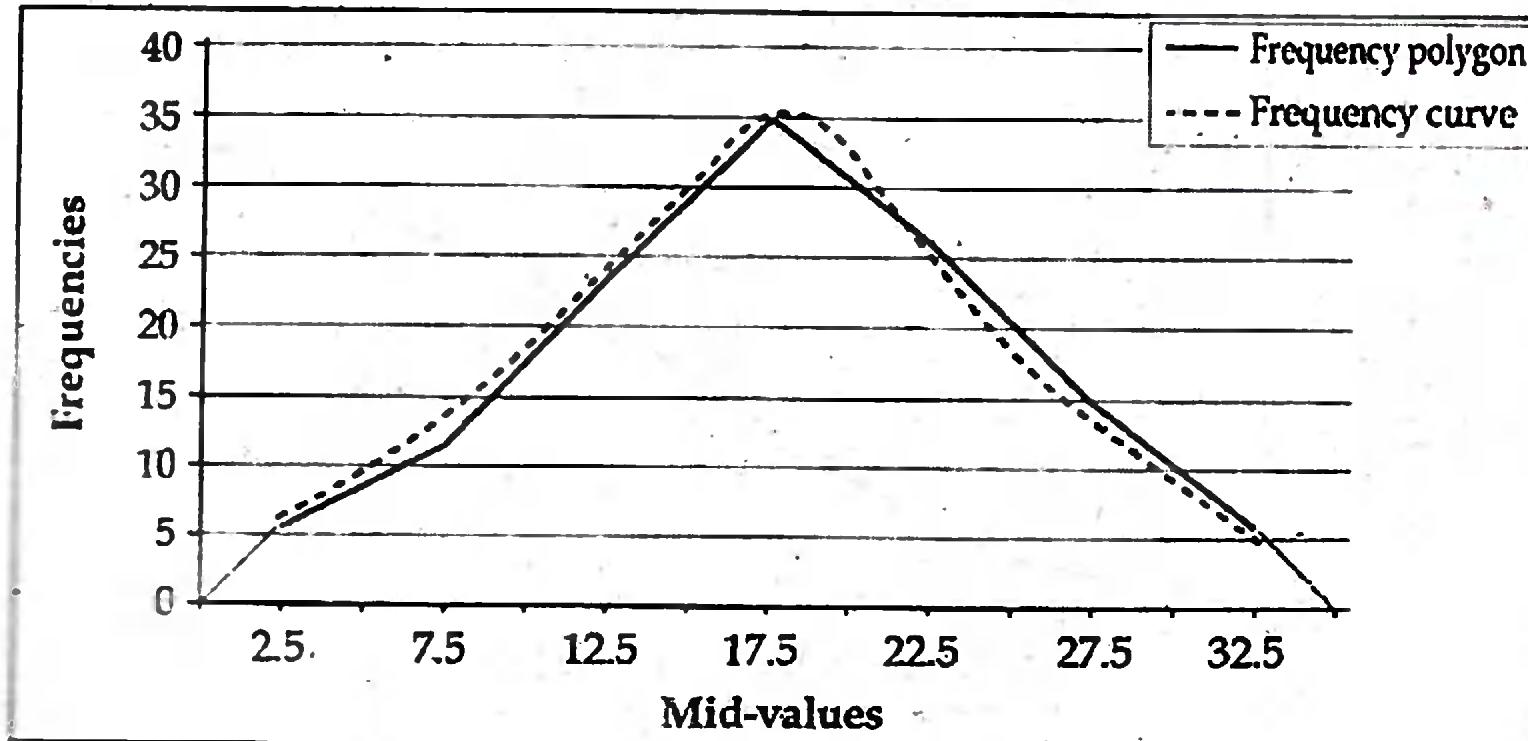


Fig. 4.13.2. Frequency polygon for length of service of employees.

Remarks. We will get the same frequency curve if we join all the points of the frequency polygon and then smooth it.

4.14. Cumulative Frequency polygon, cumulative frequency curve or Ogive

A cumulative frequency polygon or ogive curve is the graphic representation of a cumulative frequency distribution. First a cumulative frequency distribution is calculated from the original frequency distribution by adding the class frequencies.

There are two methods of constructing cumulative frequency polygon or ogive, namely:

- (i) The less than method, and
- (ii) The more than method.

4.14.1. Cumulative Frequency Polygon or Ogive by less than Method. In this method we start with the upper limits of the classes and go on adding the frequencies to get the "less than cumulative frequency table". Now upper limits (upper class boundaries) of the class intervals are plotted on the X-axis and the cumulative frequencies are plotted in the Y-axis. A point is then plotted directly above each upper class limit at a height corresponding to the cumulative frequency. One additional point is then plotted above the lower class limit for the first class at a height of zero cumulative frequency. These points are then joined by straight line and the resulting polygon is called a less than cumulative frequency polygon.

When the points are connected by a smooth freehand curve, the resulting curve is called a less than ogive or a less than cumulative frequency curve. That is a smooth cumulative frequency polygon is called a cumulative frequency curve or an ogive. It is an elongated S-shaped curve.

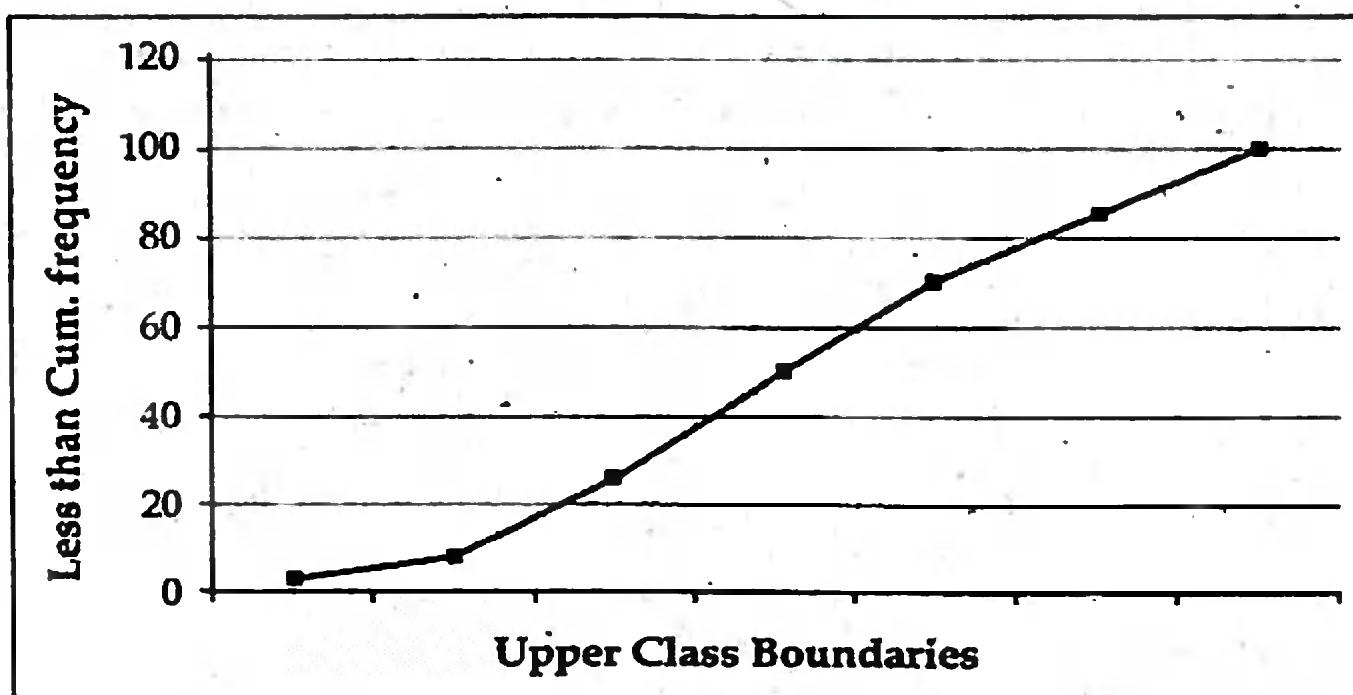


Fig. 4.14.1. Less than Ogive.

4.14.1. Cumulative Frequency polygon or ogive by more than method. In this method we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class to get the "more than cumulative frequency table". In this table total frequency is put corresponding to the lower limit of the first class. Now lower limits (lower class boundaries) of the class intervals are plotted on the X-axis and the cumulative frequencies are plotted in the Y-axis. A point is then plotted directly above each lower class limit at a height corresponding to the cumulative frequency at that lower class limit. One additional point is to be plotted above the upper class limit of the last class at a height of zero cumulative frequency. These points are then joined by straight line and the resulting polygon is called a more than cumulative frequency polygon.

When the points are connected by a smooth freehand curve, the resulting curve is called 'a more than ogive' or 'a more than cumulative frequency curve'. It is a declining curve.

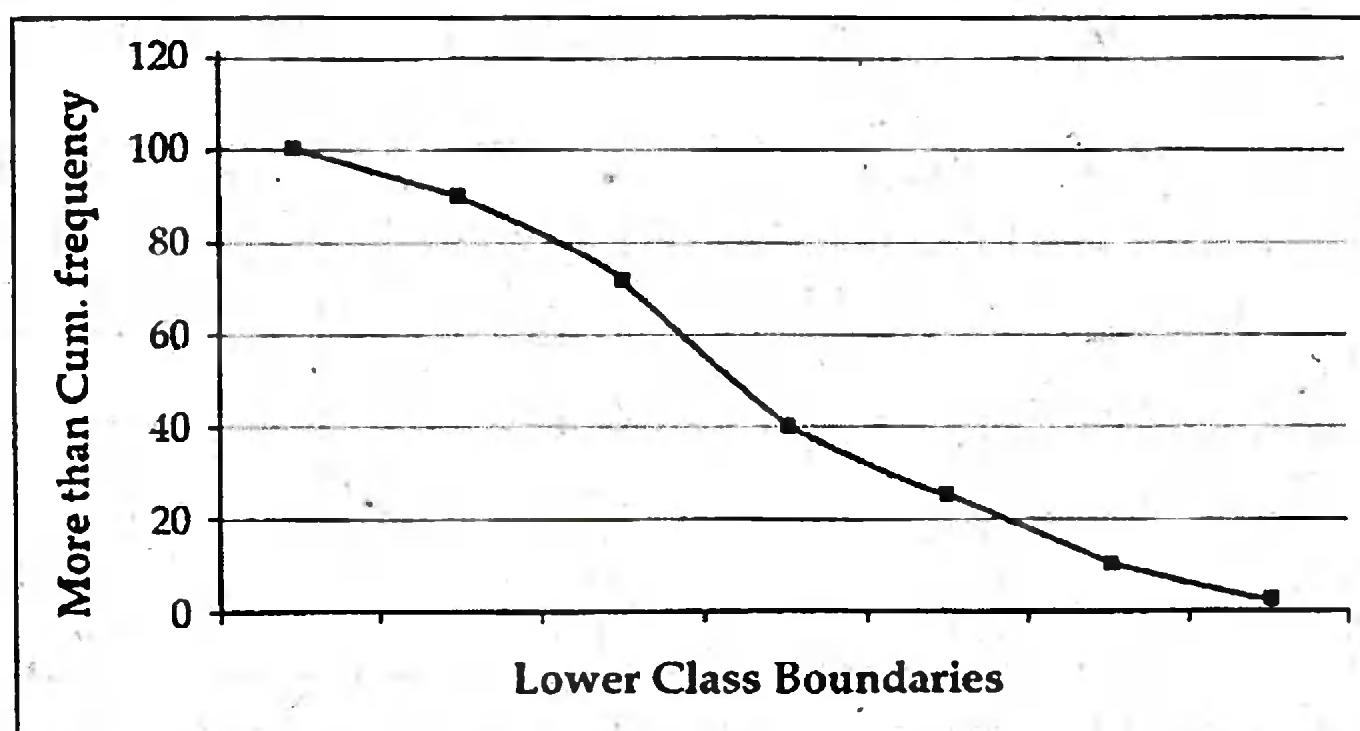


Fig. 4.14.2. More than Ogive.

Remarks.

1. Actually ogive means less than ogive. It is mainly used for finding different position measures such as median, quartiles, deciles and percentiles and will be discussed in the next chapter.
2. Ogive can also be drawn from the relative cumulative frequency or percent cumulative frequency distribution in a similar way. They will take the same form and serve the same purpose.
3. Less than and more than ogives are required to identify the position of median only. If we draw less than and more than ogive on the same graph paper, they will take χ -shape and the intersecting point between the two ogives helps us to identify the median and it will be discussed in the next chapter.

Example 4.14.1. The following data refer to the length of service in years of 124 employees of a factory.

Length of service (in years)	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35
No. of Employees	6	12	24	35	26	15	6

- (i) Construct a less than and a more than cumulative frequency tables,
- (ii) Draw a less than and a more than cumulative frequency polygons and
- (iii) Also draw their respective ogive curves with the above frequency distribution.

Solution. To draw the cumulative frequency polygon or the ogive by the less than or more than methods, we have to construct less than or more than cumulative frequency tables. First we shall construct a less than cumulative frequency table.

(i) Less than cumulative frequency table of the service of 124 employees of a factory.

Length of service	Cumulative frequency	Relative cumulative frequency	Percent cumulative frequency
Less than 0	0	0	0
Less than 5	6	0.05	5
Less than 10	18	0.15	15
Less than 15	42	0.39	39
Less than 20	77	0.62	62
Less than 25	103	0.83	83
Less than 30	118	0.95	95
Less than 35	124	1.00	100

More than cumulative frequency table of the service of 124 employees of factory.

Length of service	Cumulative frequency	Relative cumulative frequency	Percent cumulative frequency
More than 0	124	1.00	100
More than 5	118	0.95	95
More than 10	106	0.85	85
More than 15	82	0.66	66
More than 20	47	0.38	38
More than 25	21	0.17	17
More than 30	6	0.05	5
More than 35	0	0.00	0

(ii)

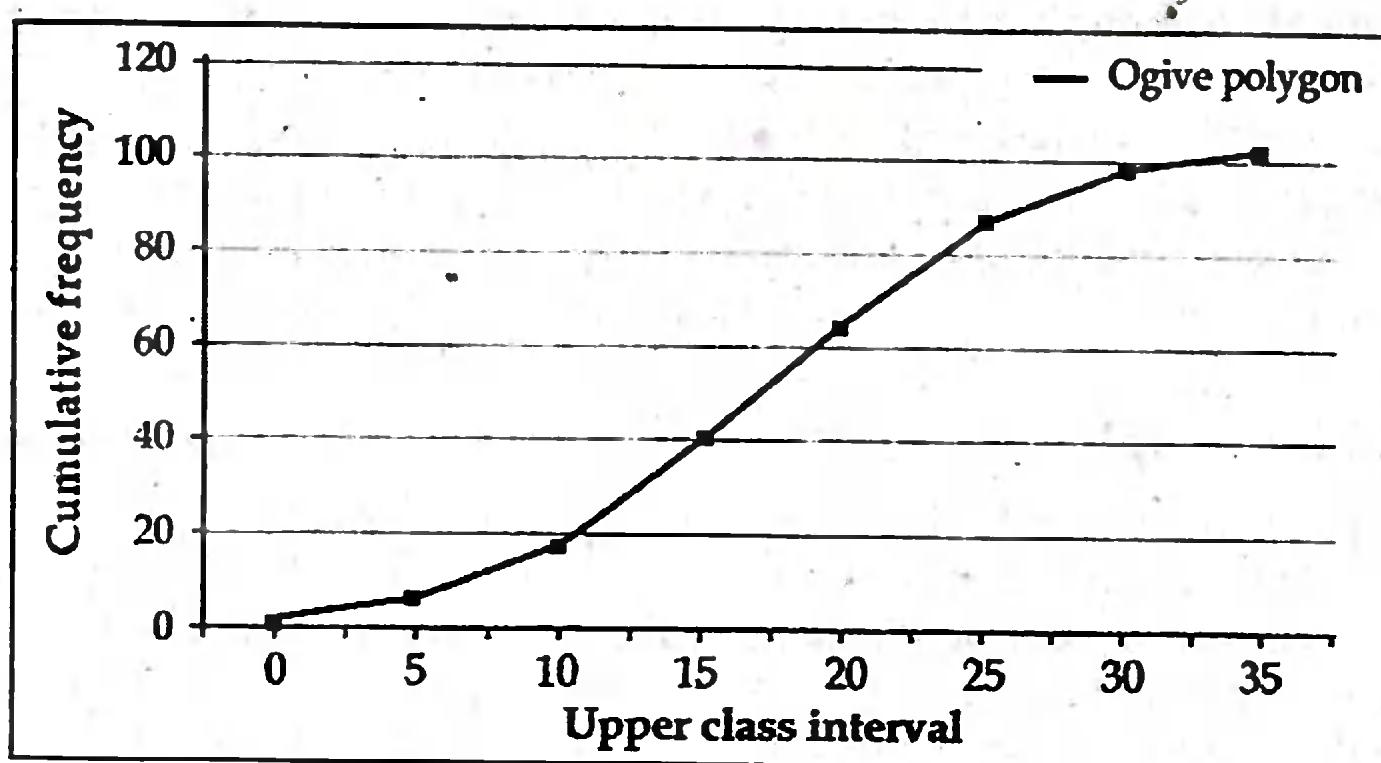


Fig. 4.14.3(a). Less than ogive.

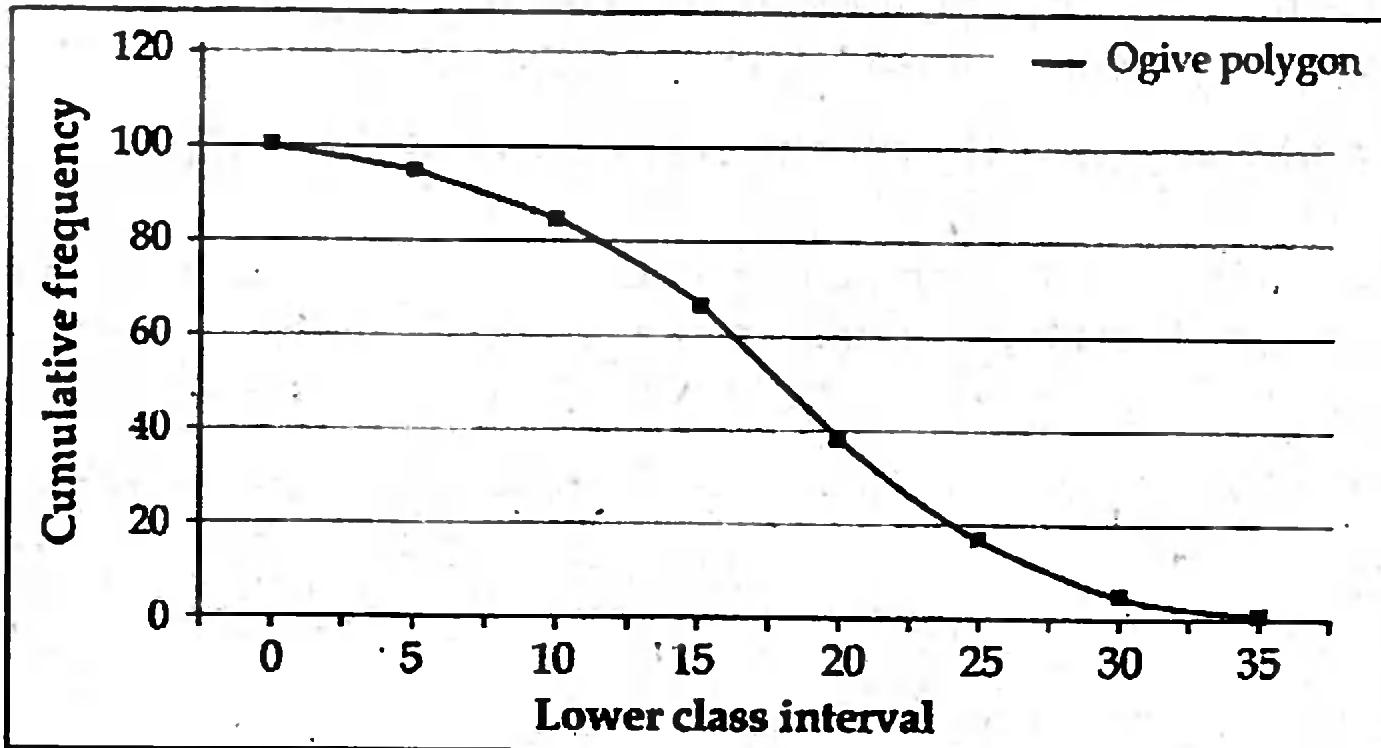


Fig. 4.14.3(b). More than ogive.

(iii)

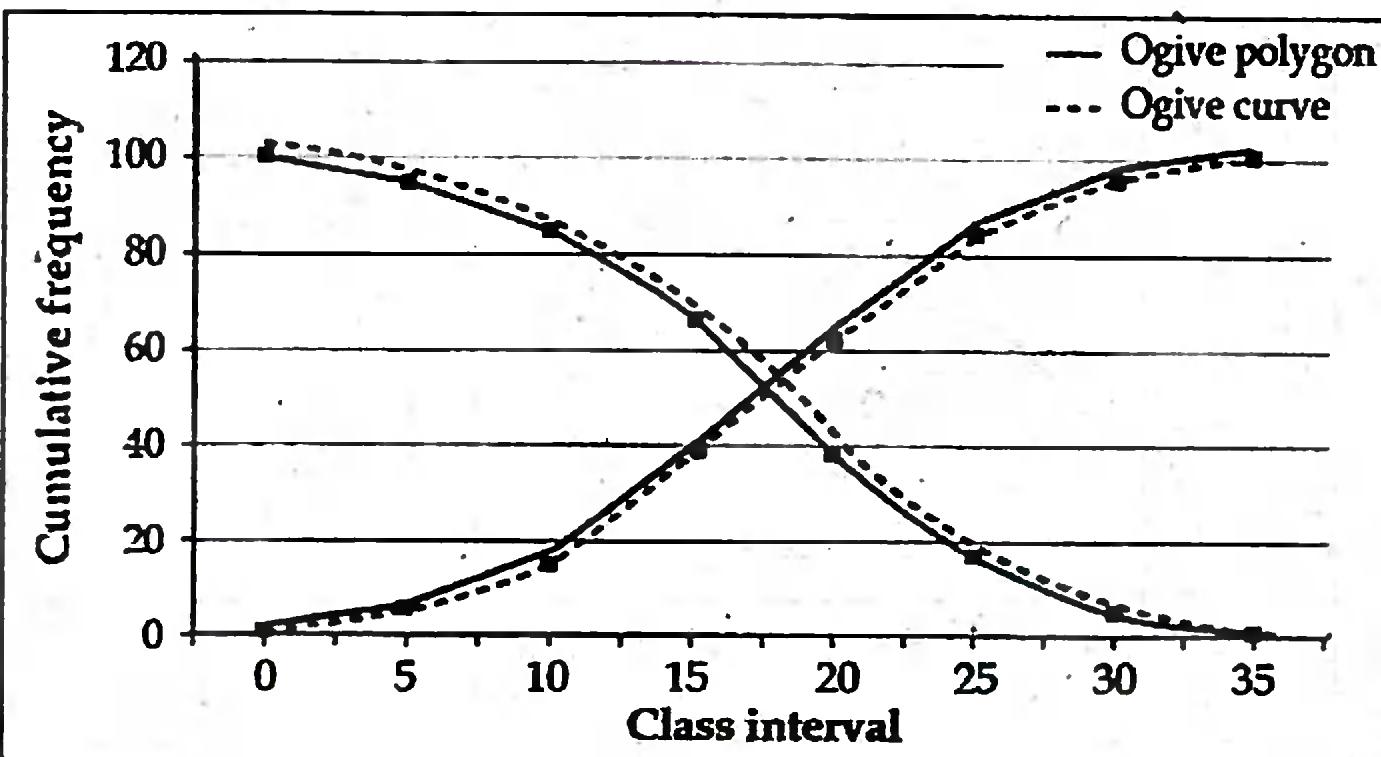


Fig. 4.14.3(c). Less than and more than ogive curve for length of service of employees.

Comment. Although a frequency curve or an ogive is often approximated by smoothing the corresponding frequency polygon or cumulative frequency polygon, it should be emphasized that the concept of a frequency curve or ogive is applicable only to a population rather than a sample. Moreover, strictly speaking, the concept of a continuous curve applies to a continuous variable only, but often it is used in connection with discrete variables as well.

4.15. Line Graph

Line graph is another type of graph to represent a special type of numerical data called the time series data. When the values of a variable are available at different points of time, the set of values or data is known as a time series data. In this graph, time is plotted along the X-axis and the values of the variable are plotted along the Y-axis. Points are plotted directly above each time at a height corresponding to the values of the variable. The points are then connected by straight lines and the resulting graph is called line graph. Often smooth curve is drawn through these points.

Utility of line graph. The situations in which the line graph is particularly useful are:

1. It is very helpful to compare several time series data on the same graph.
2. It is used to estimate or forecast the values of a variable graphically.
3. It gives the movement or the trend of the values of a variable.

Example 4.15.1. The sales of a firm for the past 12-year are given below:

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Sales (in thousand Taka)	235	355	395	420	455	490	430	485	520	575	570	590

Draw a line graph with the above data.

Solution. The line diagram for the given sales is drawn by joining the points obtained for sales of each year against year and shown below:

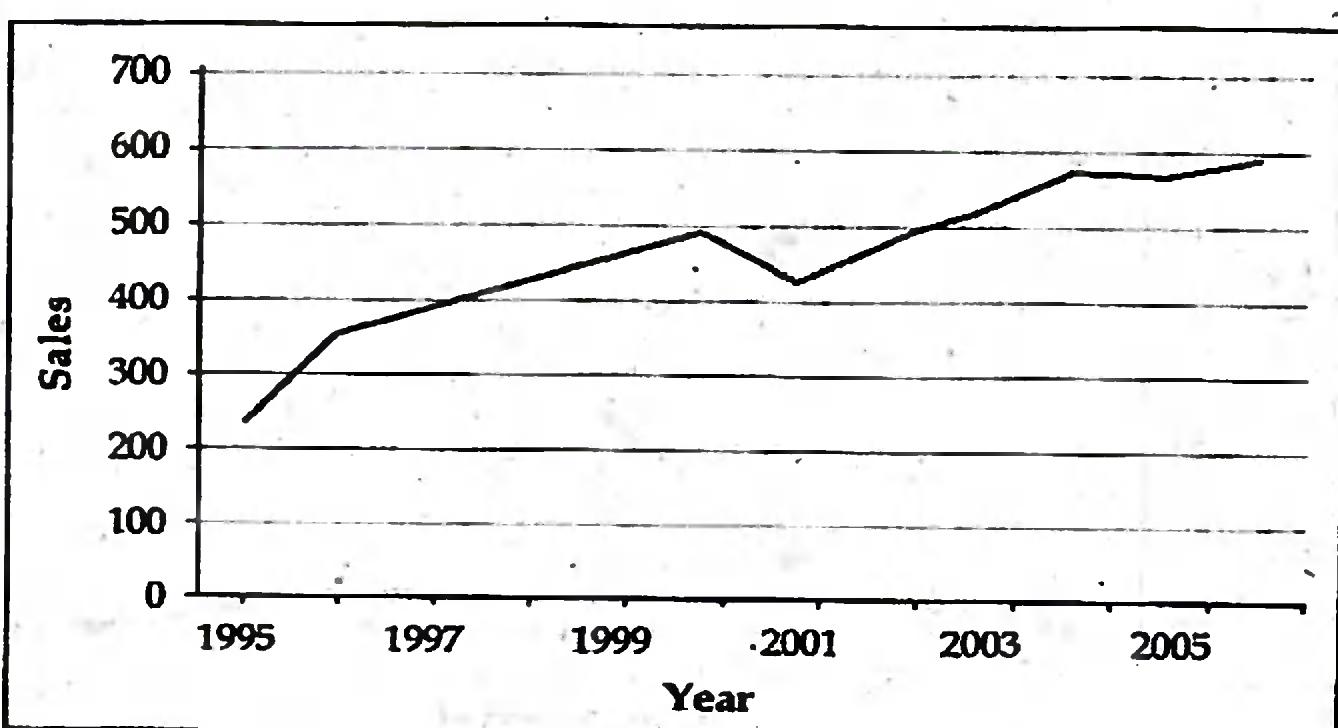


Fig. 4.15.1. Line diagram for the sales of different years.

4.16. Some Additional Examples

Example 4.16.1. The following data refer to the ages of 60 employees of a firm.

33, 41, 21, 25, 36, 38, 35, 36, 35, 37, 42, 30, 35, 37, 36, 38, 30, 54, 40, 48, 15, 28, 51, 42, 25, 41, 30, 27, 42, 36, 28, 26, 37, 54, 44, 31, 36, 40, 36, 22, 30, 31, 19, 48, 16, 42, 32, 21, 22, 40, 43, 42, 39, 38, 37, 33, 49, 47, 46, 48.

- (i) Construct a frequency table with suitable class interval,
- (ii) Draw histogram, frequency polygon, ogive polygon and ogive curve with the frequency table,
- (iii) Find percent frequency of the age of the workers and comment.

Solution. The range of the data set is

$$\text{Range} = 54 - 15 = 39$$

$$\text{The number of classes} = \sqrt{60} = 7.75 \approx 8$$

$$\text{Width of the class} = \frac{39}{8} = 4.88 \approx 5$$

The frequency distribution of the ages of the workers by the inclusive method is given below.

Class interval (Age)	Tally marks	Frequency
15 – 19		3
20 – 24		4
25 – 29		6
30 – 34		9
35 – 39		17
40 – 44		12
45 – 49		6
50 – 54		3

Table for finding histogram, percent frequency, frequency polygon and ogive polygon and ogive curve

Class interval	Frequency	Mid-Point	Class boundary	Relative frequency	Percent frequency	Cumulative frequency
15-19	3	17	14.5 – 19.5	.05	5	3
20-24	4	22	19.5 – 24.5	.07	7	7
25-29	6	27	24.5 – 29.5	.10	10	13
30-34	9	32	29.5 – 34.5	.15	15	22
35-39	17	37	34.5 – 39.5	.28	28	39
40-44	12	42	39.5 – 44.5	.20	20	51
45-49	6	47	44.5 – 49.5	.10	10	57
50-54	3	52	49.5 – 54.5	.05	5	60

The histogram, frequency polygon and ogive for ages of employees are constructed below.

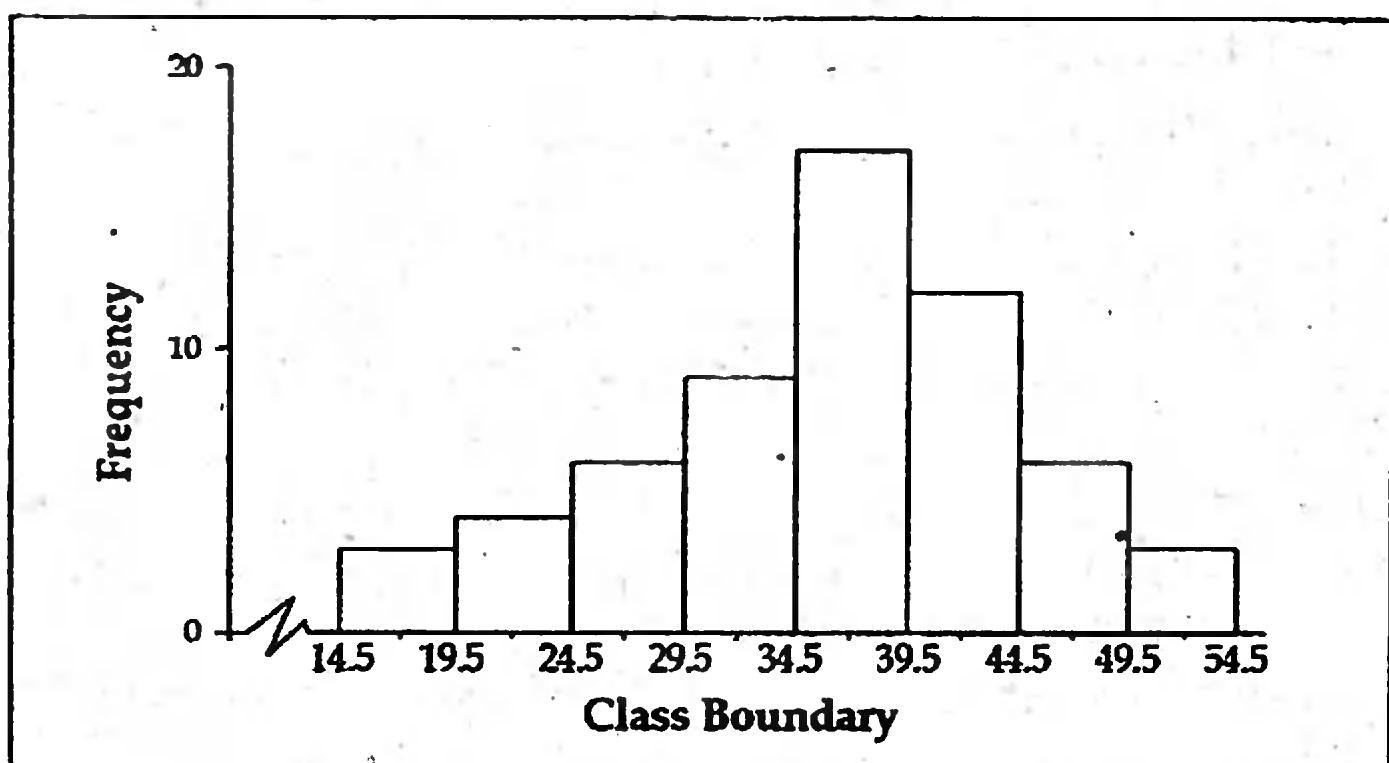


Fig. 4.16.1. Histogram for ages of employees.

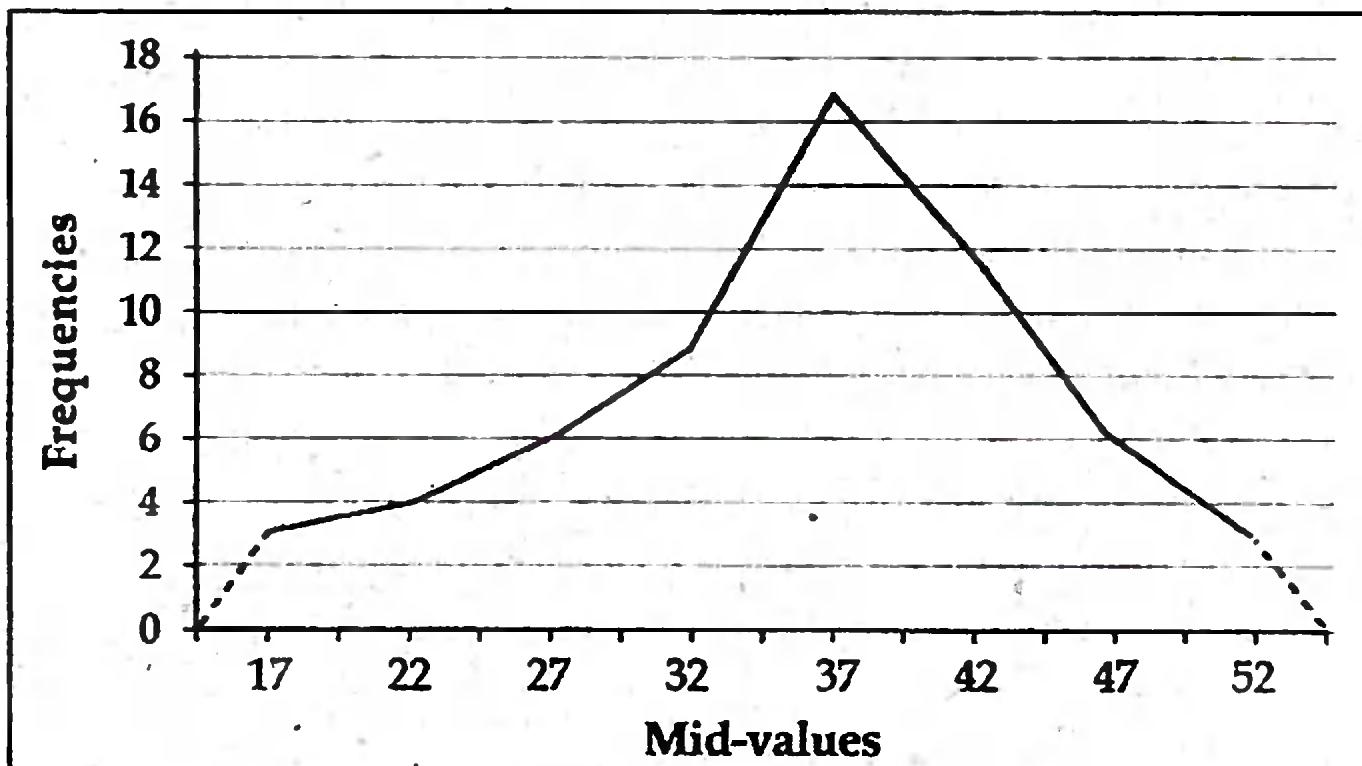


Fig. 4.16.2. Frequency polygon for ages of employees.

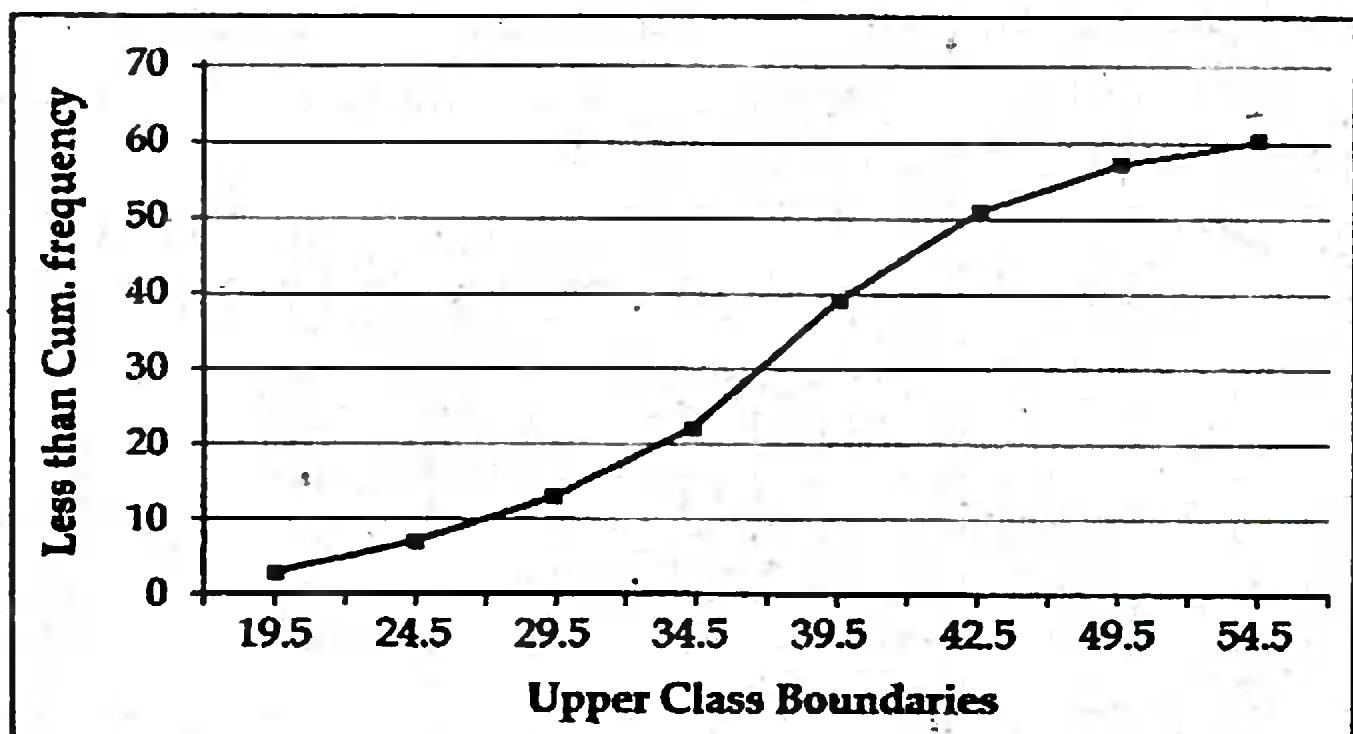


Fig. 4.16.3. Ogive for ages of employees.

Example 4.16.2. The following table gives the monthly income in taka per family in a working class locality.

Monthly income (in Taka)	No. of families (f)
1500 - 2000	15
2000 - 2500	25
2500 - 3000	31
3000 - 3500	42
3500 - 4000	27
4000 - 4500	16
4500 - 5000	8

Draw (i) a histogram, (ii) frequency polygon, (iii) Less than and More ogives by (a) cumulative frequency distribution and (b) relative cumulative frequency distribution.

Solution. (i) Since the class intervals are equal, plot class intervals along the X-axis and the frequency of each class along the Y-axis. Then construct adjacent rectangles over the class intervals. The resulting graph gives us the required histogram and the graph looks like as in figure 4.16.4.

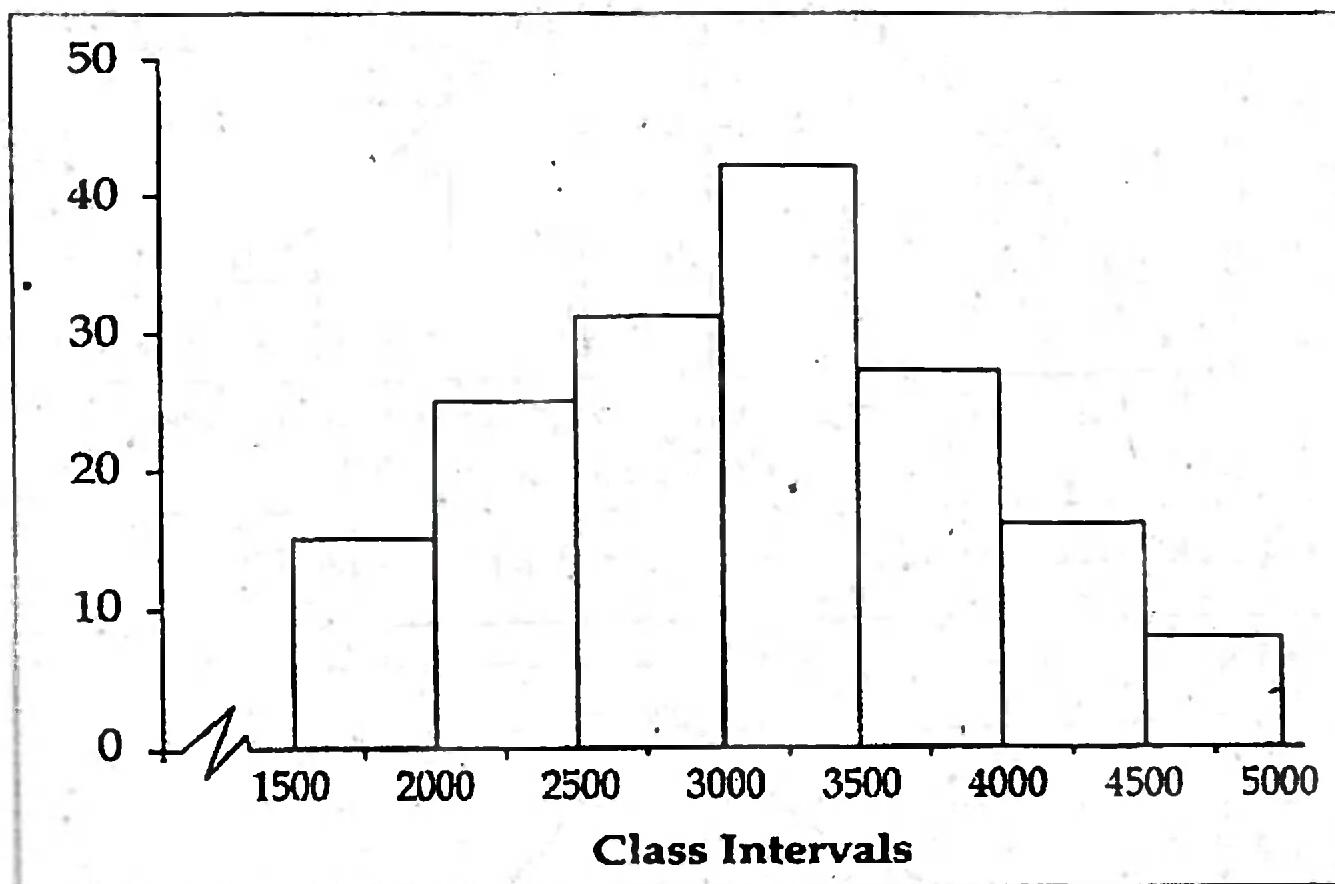


Fig. 4.16.4. Histogram for Income of families

(ii) Frequency polygon from histogram. The frequency polygon is obtained by joining the midpoints of upper horizontal sides of each rectangle of the histogram by straight lines. It is shown in figure 4.16.5 with the histogram.

(ii) Frequency polygon from the frequency distribution. For drawing frequency polygon, we require mid-values of the classes and their corresponding class frequencies. For this we need the following frequency table.

Monthly income (in Taka)	No. of families (f)	Mid-value (X)
1500 - 2000	15	1750
2000 - 2500	25	2250
2500 - 3000	31	2750
3000 - 3500	42	3250
3500 - 4000	27	3750
4000 - 4500	16	4250
4500 - 5000	8	4750

Now we plot mid-values along the X-axis and the frequency along the Y-axis. Plot points directly above the class mid-values at a height corresponding to the frequency of the class. Classes of zero frequency are added at each end of the frequency distribution. The frequency polygon is obtained by joining all the points by straight lines and is shown in figure 4.16.6.

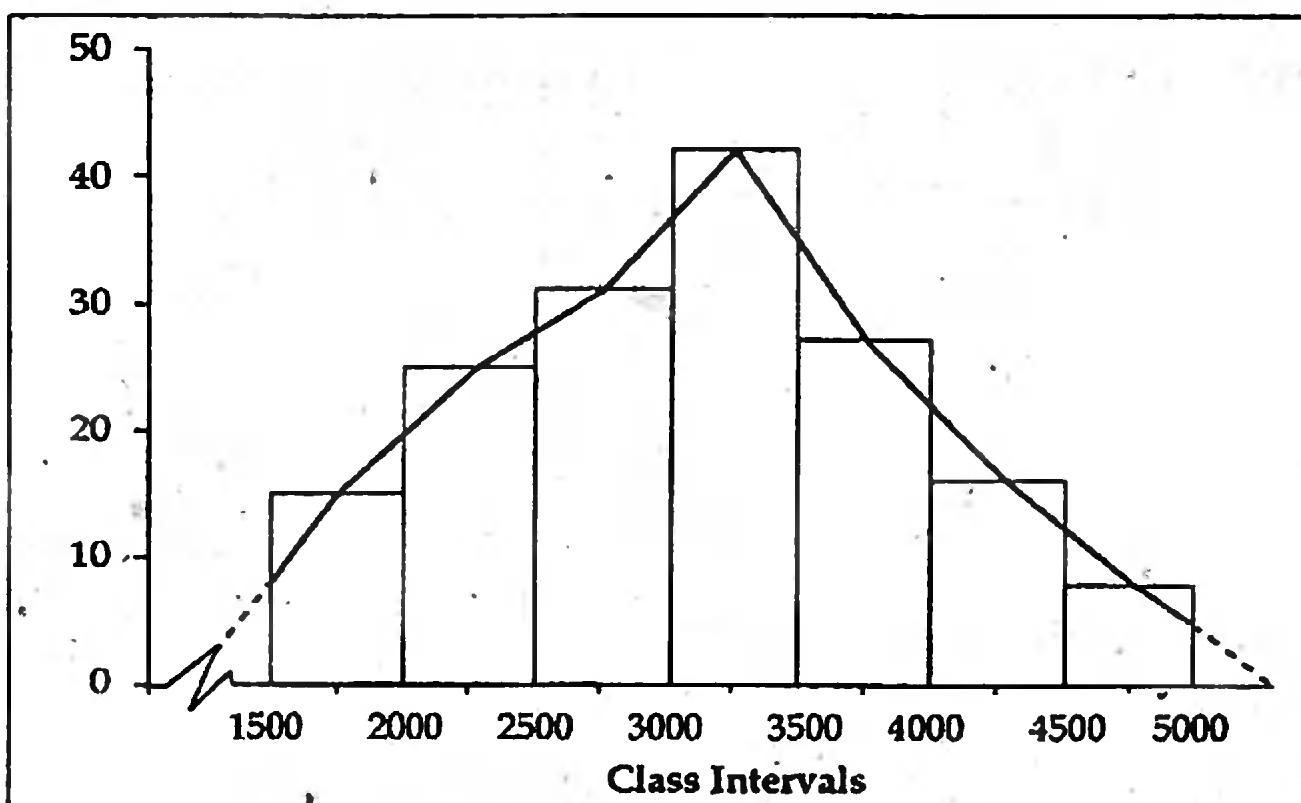


Fig. 4.16.5 Frequency polygon for income of families from Histogram.

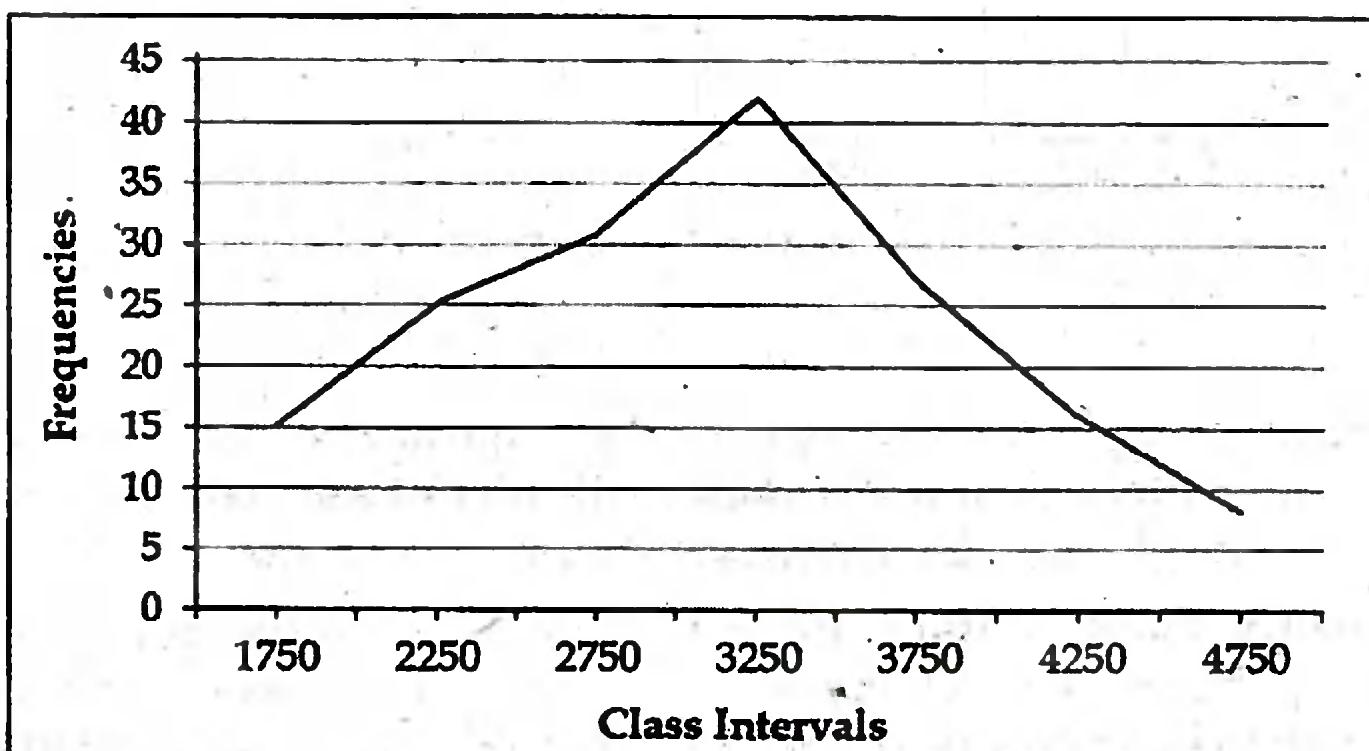


Fig. 4.16.6 Frequency polygon for income of families using frequency distribution.

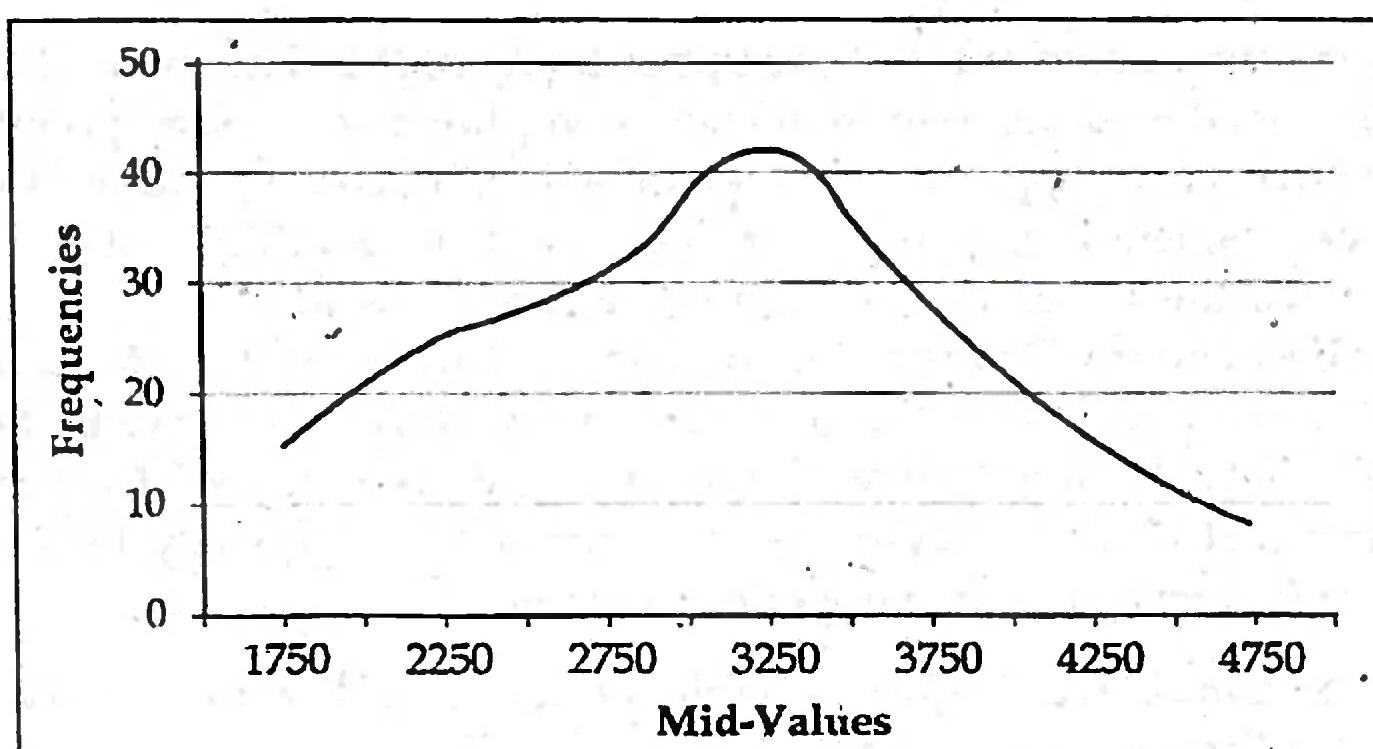


Fig. 4.16.7 Frequency curve for income.

- (iv) For drawing two ogives, we need less than and more than cumulative frequency tables.

Less than cumulative frequency table of monthly income of 164 working families.

Monthly income (in Taka)	Cumulative Frequency	Relative cumulative frequency	Percent cumulative frequency
Less than 1500	0	0.00	0
Less than 2000	15	0.09	9
Less than 2500	40	0.24	24
Less than 3000	71	0.43	43
Less than 3500	113	0.69	69
Less than 4000	140	0.85	85
Less than 4500	156	0.95	95
Less than 5000	164	1.00	100

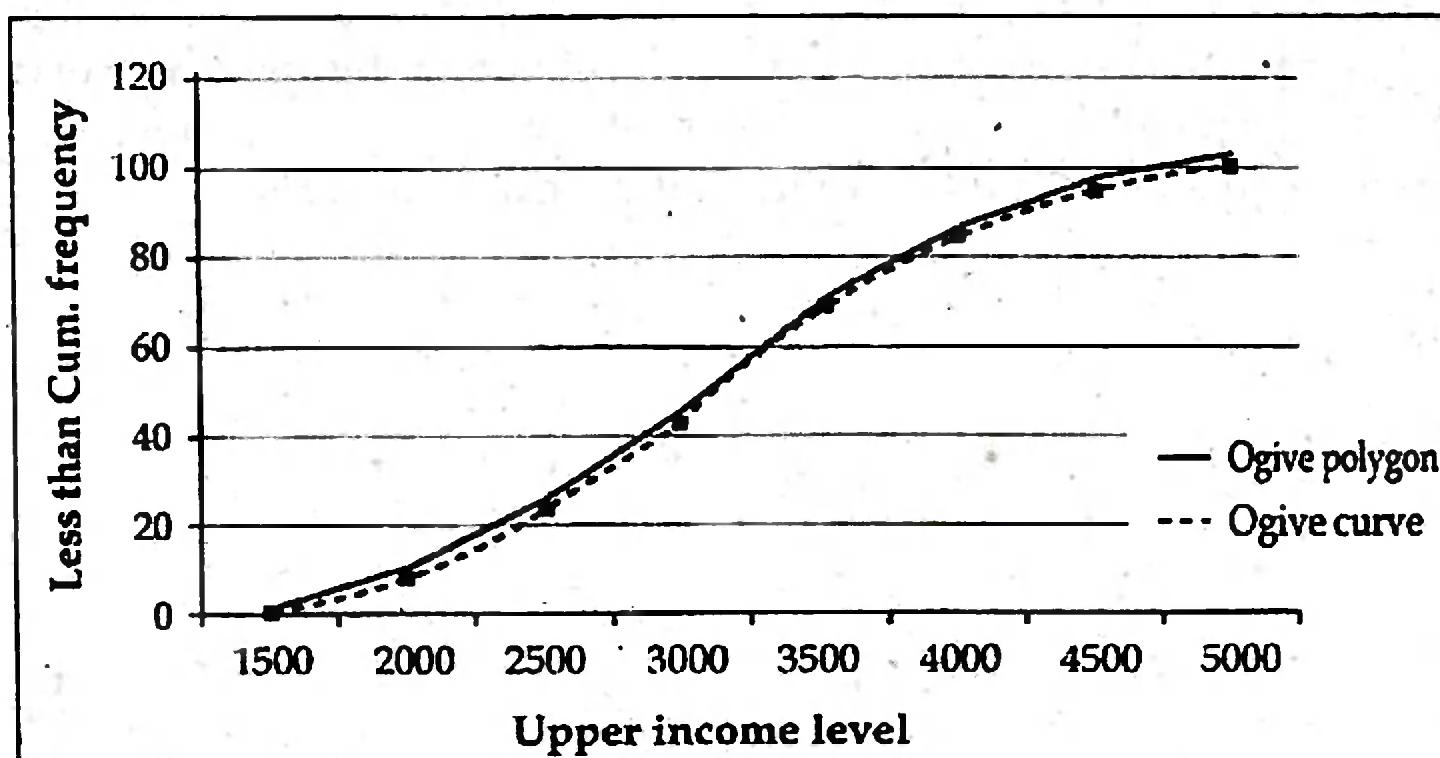


Fig. 4.16.8. Less than ogive for income.

To get a less than ogive, we plot the upper limits of the class interval along the X-axis and the cumulative frequency along the Y-axis. Now points are plotted above each upper class limit at a height corresponding to the cumulative frequency of that upper class limit. One additional point is plotted above the lower class limit for the first class at a height of zero cumulative frequency. Points are then connected freehand to get a smooth curve, which is called a less than ogive or simply ogive. The curve is shown in figure 4.16.8. A very similar shape of curves is obtained by putting relative cumulative frequency or percent cumulative frequency instead of cumulative frequency over the upper class limits.

More than cumulative frequency table of the monthly income of 164 working families.

Monthly income (in Taka)	No. of families (f)	Relative cumulative frequency	Percent cumu- lative frequency
More than 1500	164	1.00	100
More than 2000	149	0.91	91
More than 2500	124	0.76	76
More than 3000	93	0.57	57
More than 3500	51	0.31	31
More than 4000	24	0.15	15
More than 4500	8	0.05	5
More than 5000	0	0.00	0

To get a more than ogive, we plot the lower limits of the class interval along the X-axis and the cumulative frequency along the Y-axis. Now points are plotted above each lower class limit at a height corresponding to the cumulative frequency of that lower class limit. One additional point is plotted above the upper class limit for the last class at a height of zero cumulative frequency. Points are then connected freehand to get a smooth curve, which is called a more than ogive. The curve is shown in figure 4.16.9. A very similar shape of curves is obtained by putting relative cumulative frequency or percent cumulative frequency instead of cumulative frequency on the lower limits of the class intervals.

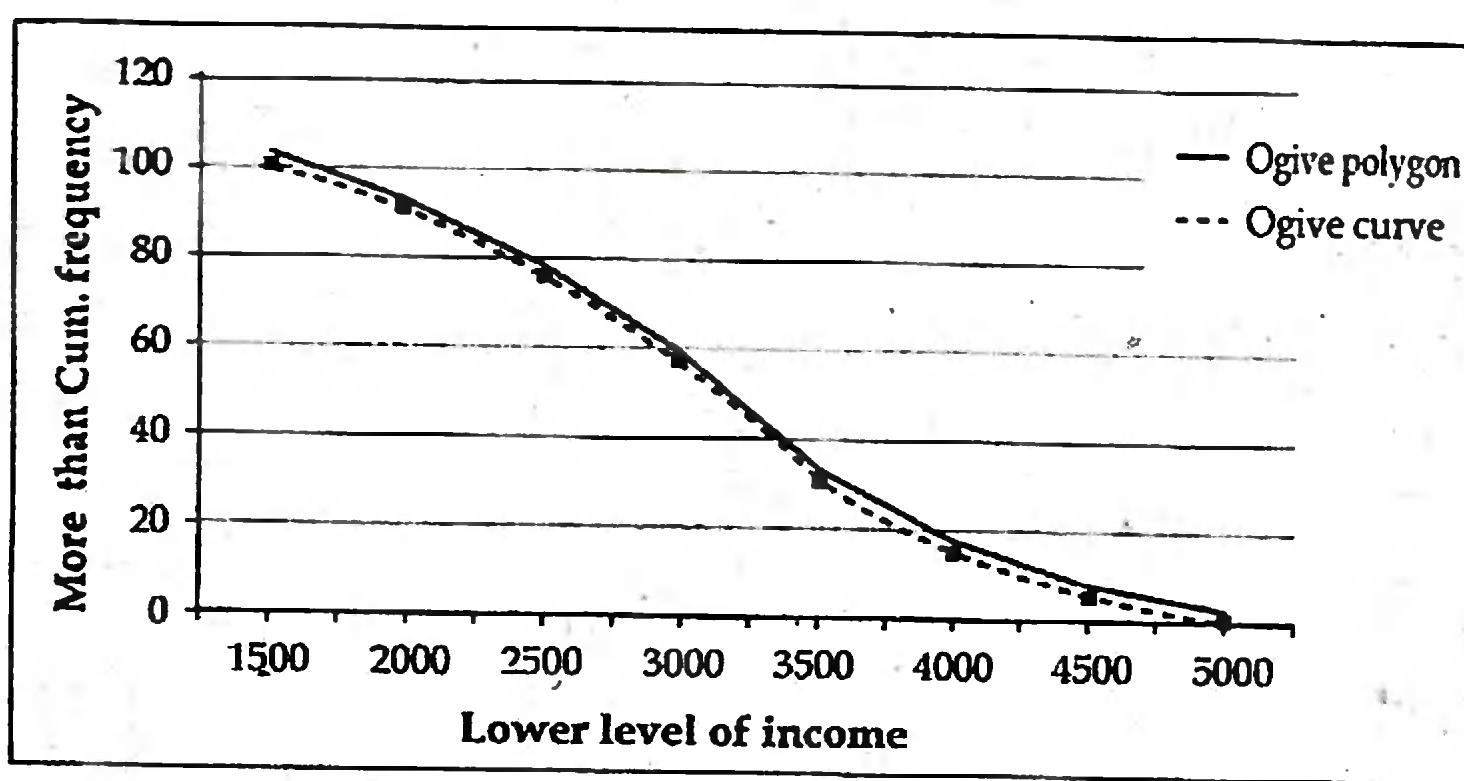


Fig. 4.16.9. More than ogive for income.

4.17. Cross Tabulation of Statistical Data

So far we have focused on tabular and graphical methods that are used to summarize the data for one variable at time. Often a manager or a decision maker is interested in tabular and graphical methods that will assist in the understanding of the relationship between two variables. Cross tabulation and scatter diagrams are two such methods. Cross tabulation can be constructed in three ways by (i) two qualitative variable; (ii) one qualitative and the other quantitative variables and (iii) two quantitative variables.

4.17.1. Contingency table or Cross table by two sets of qualitative data. When the data of two qualitative variables are presented in a table form to summarize the data. This type of table is called a contingency table. The main purpose of contingency table is to study the association between two qualitative variables.

Example 4.17.1. A study was conducted on 1100 workers of a factory. Smoking habit and their health conditions are presented in the following contingency table:

Table showing smoking habits of 1100 workers with their health status.

Smoking habit	Health condition			Total
	Excellent	Good	Poor	
Smoker	25	125	500	650
Non-smoker	150	250	50	450
Total	175	375	550	1100

Now we shall cite one example of a cross table with a qualitative and a quantitative variable.

Example 4.17.2. The following data refers to the quality of foods and the meal prices in Taka of 172 restaurants.

Quality of foods	Meal Price				Total
	30 - 50	51 - 70	71 - 90	91 - 110	
Excellent	3	5	15	25	48
Very Good	12	20	10	8	50
Good	20	16	3	1	40
Satisfactory	25	8	1	0	34
Total	60	49	29	34	172

Cross table of two quantitative variables is called a correlation table. The main purpose of this table is to study the relationship between the two variables. Now, we want to cite one example.

Example 4.17.3. The following bi-variate frequency table refers to the ages of husbands and their wives:

Age of husbands /Age of wives	15-25	25-35	35-45	45-55	55-65	65-75	Total
10-20	3	5	1				9
20-30	1	10	6	3			20
30-40		2	14	6	2		24
40-50			3	10	4	1	18
50-60				2	2		4
60-70				1	3		4
Total	4	17	24	19	9	6	79

4.18. Scatter Diagram

A scatter diagram is a graphical presentation of the relationship between two quantitative variables.

Example 4.18. The following data represents the money spent on advertising of a product and the respective profits realized from each advertising period for a given product. The amounts are in thousands of taka.

Advertisement cost in thousands Tk (x)	7	8	9	10	11	12	13	15
Profits in thousand Tk. (Y)	10	11	10	13	16	18	19	21

Draw a scatter diagram for the data.

Solution. The scatter is drawn by plotting profits against advertisement as shown below:

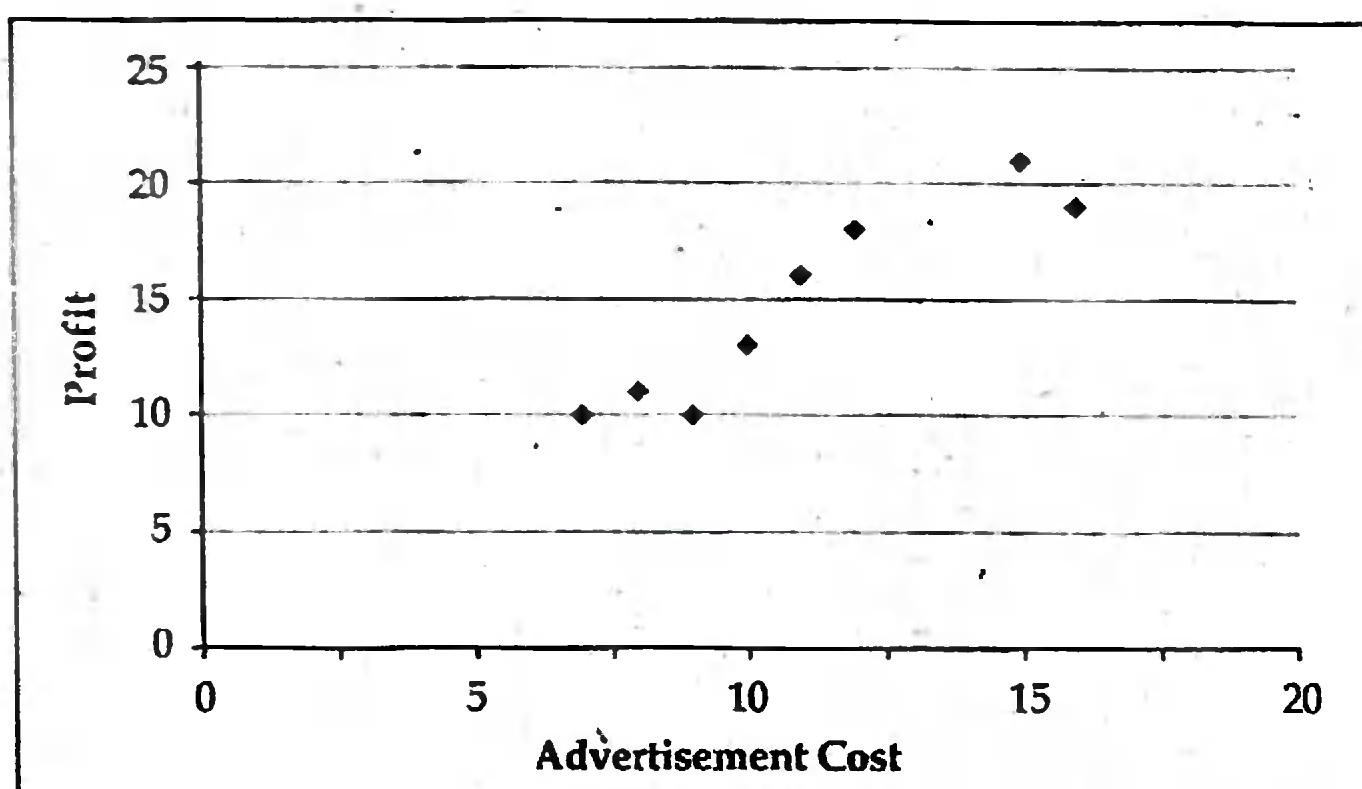


Fig. 4.18. Scatter diagram for advertisement cost and profit.

From the scatter plot it is evident that in most of the cases profit has been increased as advertisement cost increases and vice-verse that means both variables change in the same directions.

Exercise

1. Define the following: Frequency, Frequency distribution, proportion, percent, ratio and rates with examples.
2. Discuss the procedure of construing a frequency distribution of qualitative data Explain the terms class frequency, relative frequency, percent frequency.
3. What are the important diagrams for representing a frequency distribution of qualitative data? Discuss different types of bar diagrams.
4. How can you construct a pie chart? Cite a situation where you can apply this chart.
5. State the important steps for constructing a frequency distribution of continuous data.
6. What are the important graphs for representing a frequency distribution of quantitative data? Discuss histogram and ogive curve.
7. Discuss how to construct a histogram. Distinguish between a histogram and a bar diagram.
8. What is a stem leaf display? Distinguish between a stem leaf display and histogram.
9. Distinguish between a (i) frequency polygon and frequency curve, (ii) ogive polygon and ogive curve.

Application

11. A country has four political parties say A, B, C, D. An opinion survey was conducted on 50 people randomly. The data were obtained as follows:

B, C, A, B, A, C, D, A, A, B, C, A, B, C, D,
 A, A, B, C, A, B, D, A, A, B, B, D, C, A, D,
 B, A, B, D, C, C, B, A, A, B, A, D, C, A, B,
 A, C, B, A, D,

- Construct a frequency distribution with this qualitative data and comment.
 - Which one is the most popular party?
 - What is the percent of people liked party B?
 - Construct a bar diagram
 - Construct a pie chart.
12. The following data give the number of shoes in different sizes sold by a shop on last one week:

L, M, S, M, M, M, S, L, L, S, M, M, S, L, M, M, S, L, S, M, M, M, S, L, S, S, S, S, M, M, M, L, S, M, M, M, L, S, M, S, M, L, M, L, M, M, L, M, M, L, S, S, M

- Construct a frequency table.
 - Which one is the most preferable size?
 - Also construct a bar diagram.
13. The following table gives the yearly budget of a country in Billion taka for different sector:

Sector	Agriculture	Industry	Education	Transport	Others	Total
Estimated Expenditure (in billion taka)	80	70	40	25	55	270

Construct a pie chart.

14. The population of some Islamic countries for the year 1993 is given below:

Name of country	Bangladesh	Egypt	Iran	Pakistan	Turkey
Population(in Million)	113	56	61	123	59

- Construct a bar diagram
 - Draw a pictogram.
15. The table given below gives population of Bangladesh in million for the period 1988 to 1993

Year	1988	1989	1990	1991	1992	1993
Population (in million)	103	105	107	109	111	113

- Construct a bar diagram.

- (ii) Construct a pictogram with the above data.
16. The following table gives the number of workers in different categories of four garments factory:

Worker	Factory 1	Factory 2	Factory 3	Factory 4
Male	100	250	300	250
Female	600	750	850	900
Girl	250	150	250	250
Boy	50	50	100	100
Total	1000	1200	1400	1500

Construct a component bar diagram with the help of the above data.

17. The table given below gives the population in million of four big cities of a country for last three censuses:

Year	City			
	A	B	C	D
1991	5.5	4.0	3.5	6.0
2001	6.5	6.0	4.5	8.5
2011	8.0	7.5	5.5	11.0

Construct a multiple bar diagram with above data.

18. The following table gives the number of students in different faculties for the period 2007-2009 :

Year	Faculty			
	Science	Business	Arts	Social Sciences
2007	550	440	630	450
2008	600	500	650	500
2009	650	550	700	550

- (i) Construct a multiple bar diagram and comment
(ii) Construct a component bar diagram with the above data.

19. A social scientist wants to study the family structure of the workers of a factory whose ages are over 40 years. A survey was conducted on 50 workers and the results were recorded as follows:

2, 5, 3, 5, 4, 5, 6, 7, 8, 8,
 7, 6, 5, 6, 5, 6, 6, 5, 2, 3,
 4, 5, 2, 3, 7, 6, 5, 7, 6, 3,
 4, 7, 6, 4, 5, 6, 4, 5, 4, 6,
 5, 6, 5, 6, 7, 6, 5, 6, 7, 5,

- (i) Construct a frequency distribution with the above data.
(ii) Find percent frequency and comment on your results.
(iii) Construct a histogram with the frequency distribution.

20. The following data refer to the ages of 50 employees of a firm:

33, 41, 21, 25, 36, 38, 35, 36, 35, 37, 42, 30, 35, 37, 36, 38, 30, 54, 40, 48, 15, 28, 51, 42, 25, 41, 30, 27, 42, 36, 28, 26, 37, 54, 44, 31, 36, 40, 36, 22, 30, 31, 19, 48, 16, 42, 32, 21, 22, 40.

- (i) Construct a frequency table with suitable class interval by exclusive method.
 - (ii) Draw histogram, frequency polygon, ogive curve with the frequency table.
 - (iii) Find percent frequency of the age of the workers and comment.
21. The following data give the ages of 30 workers of a factory.

25, 23, 30, 60, 25, 30, 21, 27, 34, 41,
40, 54, 47, 59, 41, 47, 59, 58, 54, 47,
58, 53, 60, 57, 54, 45, 48, 54, 57, 48,

Construct a stem and leaf diagram for this data.

22. The following data represent the amount of insurance (in units of Tk. 1000) purchased by 25 people from an insurance company in a given week:

30, 45, 100, 42, 47, 95, 50, 65, 100, 33,
85, 90, 72, 66, 76, 80, 65, 95, 64, 45,
95, 100, 86, 72, 69

Display the data by a stem leaf plot.

23. The data refer to the daily wages of 58 workers of a sugar factory:

50, 75, 100, 125, 150, 175, 200, 250, 56, 86, 105, 130, 155, 178, 210, 265, 65, 90, 107, 145, 165, 190, 215, 270, 70, 95, 108, 109, 135, 174, 196, 98, 138, 170, 188, 218, 274, 96, 115, 148, 164, 193, 120, 124, 120, 124, 148, 146, 142, 147, 135, 144, 142, 147, 138, 136, 140, 124.

- (i) Construct a frequency distribution by suitable class interval.
- (ii) Draw histogram, frequency polygon, frequency curve, ogive polygon and ogive curve.

Thus the value of a qualitative variable can only be classified into categories called classes. We can summarize such data numerically in three ways:

- (i) By computing class frequency - the number of observations in the data set that fall into each class;
- (ii) By computing the class relative frequency - the proportion of the total number of observations falling into each class.
- (iii) By computing the percent relative frequency - the percentages of observation falling into each class.

CHAPTER - 5

DESCRIBING DATA WITH NUMERICAL MEASURES

5.1. Introduction

So far, we have learned the techniques of data collection, and condensing and summarizing data in the form of frequency distribution table and presenting data in the form of different diagrams and graphs. Diagrams and graphs, discussed in the preceding chapter, are the powerful and effective media for presenting statistical data, they can only represent a limited amount of information, and they are not much helpful when intensive analysis of the data is required. Now, we shall deal with some arithmetic procedures that can be used for analyzing, interpreting quantitative data both for ungrouped and grouped data. These measures and procedures relate to some properties and characteristics of data, which include measures of central tendency, or location of data, measures of dispersion of data in itself and around some central values, and the shape characteristics of the data. Broadly speaking there are four important characteristics of a set of data or its frequency distribution. These are :

- (a) Location or central tendency,
- (b) Dispersion,
- (c) Skewness, and
- (d) Kurtosis.

In this chapter we shall discuss how to measure the first type of characteristics of a distribution. Other measures will be discussed in subsequent two chapters.

5.2. Measures of Location or Central Tendency

When we have a set of quantitative data it is observed that most of the values of the data set cluster around some central value. This tendency of a set of quantitative data is called central tendency. It is more or less a central value and one of the principal characteristics of a frequency distribution. An average is a measure of central values of a set of data whether it is a sample or population. For example, we often talk of average income, average hourly product of a firm, average weight, average age of employees etc. Thus an average is a single value, which is considered as the most representative or typical value for the respective set of data.

According to Simpson and Kafea "A measure of central tendency is a typical value around which other figures congregate."

The purpose of an average is to get one single value that describes the characteristics of the entire data numerically and facilitates comparison with other distribution of similar nature.

The most commonly used measures of central tendency or averages are :

- i) Mean,
- ii) Median, and
- iii) Mode.

Again there are three types of mean. They are:

- iv) Arithmetic mean,
- v) Geometric mean, and
- vi) Harmonic mean.

All these measures have the same unit as that of the variable. For example, if the variable height is measured in centimeter, the mean or median will also be in centimeter.

5.2.1. Characteristics of a good measure of location or central tendency.
According Yule and Kendall, a good measure of central tendency should have the following characteristics:

- i) It should be easy to understand
- ii) It should be easy to compute
- iii) It should be rigidly defined
- iv) It should be based on all the observations
- v) It should be capable of further algebraic treatment
- vi) It should have sampling stability
- vii) It should not be affected by the presence of extreme values.

5.3. Arithmetic Mean for Ungrouped Data

It is the most widely used average in statistics. It is commonly known as mean.

Definition. The arithmetic mean is the total or sum of the values of a set of observations divided by the total number of observations.

Actually, it is the center of gravity of a set of observations. Now, we shall state some formulae for finding arithmetic mean in different situations.

We may get data from a population or from a sample. When we have population data, we can compute population mean by the following formula:

5.3.1. Population Mean. If X_1, X_2, \dots, X_N are N values of a finite population, then the population mean denoted by μ (mu) is a parameter defined by :

$$\mu = \frac{\sum X_i}{N}$$

Where N is the population size or the total number of observations in the population. It is customary to represent the parameter by Greek letters. Population mean is denoted by the Greek letter μ .

Example 5.3.1. Suppose in a small city there are five drugstores. The numbers of employees at the five drugstores are 3, 5, 6, 4, and 7. Find the mean number of employees for the five stores.

Solution. We can consider it as a finite population with 5 observations. Here $N = 5$. The population mean is

$$\mu = \frac{3+5+6+4+7}{5} = \frac{25}{5} = 5 \text{ employees.}$$

5.3.2. Sample Mean or Sample arithmetic mean. If x_1, x_2, \dots, x_n are n observations of a sample, then the sample mean or sample arithmetic mean is a statistic defined by :

$$\bar{x} = \frac{\sum x_i}{n}$$

Here n is the sample size. A statistic is usually represented by ordinary letters of the English alphabet. A sample mean is denoted by \bar{x} .

Example 5.3.2. Suppose a factory has 150 workers. The monthly incomes (in taka) of 10 workers selected randomly from this factory are as follows:

3449, 3447, 3468, 3493, 3572, 3516, 3502, 3492, 3446, 3475

Find the arithmetic mean using the appropriate symbol.

Solution. It is a sample data, since only 10 workers are selected from the 150.

The sample arithmetic mean or simply sample mean is :

$$\begin{aligned}\bar{x} &= \frac{3449 + 3447 + 3468 + 3493 + 3572 + 3516 + 3502 + 3492 + 3446 + 3475}{10} \\ &= \frac{34940}{10} = \text{Tk. } 3494.\end{aligned}$$

Example 5.3.3. Suppose a Biology professor teaches a class of 10 students. He took a test in which the scores are:

$$60, 95, 70, 85, 68, 77, 64, 56, 78, 84$$

Find the arithmetic mean of the score using the appropriate symbol.

Solution. It is a population data set since there are only 10 students in the class. The population arithmetic mean is

$$\mu = \frac{60 + 95 + 70 + 85 + 68 + 77 + 64 + 56 + 78 + 84}{10} = \frac{737}{10} = 73.7$$

As mentioned earlier, observations are generally available in the raw form, are called ungrouped data or raw data. The data in example 5.1 and 5.2 are ungrouped data. When the observations are available in the form of a frequency distribution it is called grouped data. Now we shall define sample arithmetic mean from grouped data. Very often, we get frequency distribution of a discrete variable without class interval. In this case, sample arithmetic mean can be computed from the formula give below.

5.3.3. Sample arithmetic mean for grouped data in case of discrete variable. Suppose x_1, x_2, \dots, x_k are the k values of a variable with corresponding frequencies f_1, f_2, \dots, f_k then the arithmetic mean is computed by the formula

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}, \text{ where } n = \sum f_i.$$

One can compute \bar{x} by using the following table.

Value of the : x	Frequency : f	fx
x_1	f_1	$f_1 x_1$
x_2	f_2	$f_2 x_2$
x_3	f_3	$f_3 x_3$
.	.	.
x_k	f_k	$f_k x_k$
Total	$\sum f_i = n$	$\sum f_i x_i$

Thus the arithmetic mean is calculated as

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}$$

Example 5.3.4. The following frequency distribution refers to the number of children per family of 75 workers of a factory.

Number of children: x	0	1	2	3	4	5	6
Number of families: f	3	5	10	16	25	12	4

Compute the average number of children per family of the workers of the factory.

Solution. This is a frequency distribution with discrete variable. Number of children is a variable. Here $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 5$, and $x_7 = 6$ and $f_1 = 3$, $f_2 = 5$, $f_3 = 10$, $f_4 = 16$, $f_5 = 25$, $f_6 = 12$, $f_7 = 4$.

Value of the variable : x	Frequency : f	Fx
0	3	0
1	5	5
2	10	20
3	16	48
4	25	100
5	12	60
6	4	24
Total	$\sum f = 75$	$\sum Fx = 257$

Here $k = 7$, $n = \sum f_i = 75$, $\sum f_i x_i = 257$.

Hence the average number of children per family is

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n} = \frac{257}{75} = 3.43.$$

5.3.4. Arithmetic mean from a frequency distribution with class interval or from a grouped data or from a continuous variable.

Suppose x_1, x_2, \dots, x_k are the k mid-points of k classes with their corresponding frequencies f_1, f_2, \dots, f_k , then the arithmetic mean is defined as:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}, \text{ where } n = \sum f_i.$$

Here mid-point of each class is taken as the representative value of the class. One can compute \bar{x} by using the following table.

Mid-point of class interval : x	Frequency: f	fx
x_1	f_1	$f_1 x_1$
x_2	f_2	$f_2 x_2$
x_3	f_3	$f_3 x_3$
.	.	.
x_k	f_k	$f_k x_k$
Total	$\sum f_i = n$	$\sum f_i x_i$

Thus the arithmetic mean is calculated as : $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}$

Example 5.3.5. The following data refer to the number of years worked by 37 employees of a firm

Number of years worked	5 - 10	10 - 5	15 - 20	20 - 25	25 - 30	30 - 35
Number of employees	5	7	11	8	4	2

Find the average number of years worked by the employees of the firm.

Solution. The arithmetic mean of the number of years worked by the employees is the required answer. Here the mid-points of the class intervals are considered as values of the variable. To get the arithmetic mean, we required the following table.

Class interval	Mid-point : x	Frequency : f	fx
5 - 10	7.5	5	37.50
10 - 15	12.5	7	87.50
15 - 20	17.5	11	192.50
20 - 25	22.5	8	180.00
25 - 30	27.5	4	110.00
30 - 35	32.5	2	65.00
Total		37	672.50

Here $\Sigma f = n = 37$, $\Sigma fx = 672.50$

$$\text{Arithmetic mean} = \bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n} = \frac{672.50}{37} = 18.18 \text{ years}$$

That is the average number of years worked by the employees of the firm is 18.18.

This method is called direct method for finding arithmetic mean from the grouped data.

Remarks. For grouped data mid value of each class is taken as the representative value of that class as the data are available in the form of frequency distribution, the exact frequency with which each value of the variable occurs in the distribution is not known. The arithmetic mean obtained by this method is an approximate value even if the frequency distribution is obtained from a population data. But this is not always true in case of discrete variable; very often discrete data can be grouped with the values of the variable and their frequencies. So, sometimes arithmetic mean obtained from a population in case of discrete grouped data is found to be exact.

Example 5.3.6. The following frequency distribution table gives the family structure of 121 families of a factory:

Family Size	1	2	3	4	5	6	7	8	9
Number of families	7	11	16	17	26	31	11	1	1

Find the average number of family members.

Solution. Table for calculation of arithmetic mean

Family Size : x	Number of families : f	fx
1	7	7
2	11	22
3	16	48
4	17	68
5	26	130
6	31	186
7	11	77
8	1	8
9	1	9
$\Sigma f = 121$		$\Sigma fx = 555$

$$\text{Arithmetic mean} = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{555}{121} = 4.59.$$

The average number of persons per family is 4.59.

Example 5.3.7. The frequency table given below gives marks of 60 students in statistics in an examination.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Student	4	4	9	20	12	6	3	2

Compute arithmetic mean or average marks of the students in statistic.

Solution. Table for computing arithmetic mean.

Marks	Number of students : f	Mid-points : x	fx
0-10	4	5	20
10-20	4	15	60
20-30	9	25	225
30-40	20	35	700
40-50	12	45	540
50-60	6	55	330
60-70	3	65	195
70-80	2	75	150
$\Sigma f = 60$			$\Sigma fx = 2220$

$$\text{Arithmetic mean of the marks} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{2220}{60} = 37.$$

It is an approximate arithmetic mean of the marks of the student since we do not know the exact marks of the student.

The result shows that the performance of the student in statistics is not satisfactory.

Example 5.3.8. The exact data set of the example 5.3.7 is

6, 10, 58, 56, 0, 25, 32, 35, 35, 9
 78, 17, 60, 50, 35, 38, 30, 10, 48, 5
 68, 48, 35, 30, 31, 41, 23, 23, 50, 72
 19, 25, 35, 40, 46, 42, 45, 25, 60, 41
 35, 36, 38, 35, 33, 46, 28, 31, 35, 42
 46, 38, 39, 45, 48, 50, 28, 29, 31, 55

Find arithmetic mean of the original data set and comment.

Solution. Here $\sum x = 2204$ and $n = 60$.

The actual arithmetic mean of the data set is

$$\bar{x} = \frac{\sum x}{n} = \frac{2204}{60} = 36.73.$$

Comment. It is seen that the actual arithmetic mean is 36.73 but the arithmetic mean of the same data in case of grouped data is exactly 37 which is higher than the actual value. Sometimes it may be lower than the actual value. It depends on how the mid-point of each class represents the actual observations within the class.

5.3.6. Short-Cut Method for Calculating Arithmetic Mean. The method of computation for finding arithmetic mean discussed so far is time-consuming and laborious, which usually needs calculator. Now, we shall discuss a method, which is known as short-cut method. In this method, the computation of mean can be done even manually. It saves both time and labour for computing arithmetic mean compared to the direct method discussed above. This method is useful when the data set is very large.

In this case, we introduce a new variable defined by: $d_i = \frac{x_i - A}{c}$.

Here A is called assumed mean, c is the width of the class interval, n is the total number of observations, and f_i is the frequency corresponding to the value x_i . If we subtract or add any constant from a variable, its origin changes, and if the variable is divided or multiplied by some constant, the

scale changes. Hence, the system is called change of origin and scale of measurement.

Then, arithmetic mean becomes :

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times c = A + c \bar{d}.$$

This formula is known as short-cut method for computing arithmetic mean.

Remark. A is usually taken as a middle value of the variable which has the highest frequency or near to the highest frequency just to get the maximum benefit of the calculation.

Example 5.3.9. The following frequency distribution refers to the number of hours worked per month by 50 workers of a factory.

Number of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Find the average number of hours worked per month by the workers of that factory by using the (i) Direct method and (ii) Short-cut method.

Solution. Table for Calculation of Arithmetic Mean.

Number of hours worked	Mid-point x	f	fx	d = $\frac{x - 117.5}{25}$	fd
30 - 55	42.5	3	127.5	-3	-9
55 - 80	67.5	4	270.0	-2	-8
80 - 105	92.5	6	555.0	-1	-6
105 - 130	117.5	9	1057.5	0	0
130 - 155	142.5	12	1710.0	1	12
155 - 180	167.5	11	1843.5	2	22
180 - 205	192.5	5	962.5	3	15
Total		50	6525		26

i) Direct Method. The formula for computing arithmetic mean by direct method is

$$\bar{x} = \frac{\sum f_i x_i}{n}.$$

Here x_i = mid point of the i th class

f_i = frequency of the i th class

n = total frequency or total number of observations.

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{6525}{50} = 130.50 \text{ hours per month.}$$

That is, the worker of the factory worked on an average 130.50 hours per month.

(ii) Short-cut method. The formula for finding arithmetic mean by short-cut method is :

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times c$$

Here, $d_i = \frac{x_i - A}{c}$, A = assumed mean and c = size of the class interval or width of the class interval.

Here, we take $A = 117.5$ as it is in the middle most value of x , and $c = 25$ as the width of the class interval is 25.

$$\sum f d = 26, n = 50$$

$$\begin{aligned} \text{So, } \bar{x} &= A + \frac{\sum f_i d_i}{n} \times c \\ &= 117.5 + \frac{26}{50} \times 25 \\ &= 117.5 + 13 \approx 130.5 \text{ hours.} \end{aligned}$$

It is seen that both the methods give the same results but the short-cut method is easier than the direct method from the computation point of view.

Example 5.3.10. The following frequency table gives the weekly number of hours worked including overtime by 70 workers of a factory :

Number of hours worked	0-10	10-20	20-30	30-40	40-50	50-60	60-71
Number of workers	5	12	15	25	8	3	2

Compute arithmetic mean or average working hours by (a) direct method and (b) short-cut method.

Solution. (a) Table for computation of average working hours by direct method.

Number of hours worked	Number of workers : f	Mid-Point : x	fx
0-10	5	5	25
10-20	12	15	180
20-30	15	25	375
30-40	25	35	875
40-50	8	45	360
50-60	3	55	165
60-70	2	65	130
	$\Sigma f = 70$		$\Sigma fx = 2110$

$$\text{Arithmetic Mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{2110}{70} = 30.14 \text{ hours.}$$

Hence average working hours = 30.14 or 30 hours approx.

(b) Table for Computation of average working hours by short-cut method.

Number of hours worked	Number of workers : f	Mid-point x	$d = \frac{x - 35}{10}$	fd
0 - 10	5	5	-3	-15
10 - 20	12	15	-2	-24
20 - 30	15	25	-1	-15
30 - 40	25	35	0	0
40 - 50	8	45	1	8
50 - 60	3	55	2	6
60 - 70	2	65	3	6
$\sum f = 70$				$\sum fd = -34$

$$\text{Arithmetic mean} = \bar{x} = A + \frac{\sum fd}{\sum f} \times c; \text{ Here } A = 35 \text{ and } c = 10$$

$$= 35 + \frac{-34}{70} \times 10 = 35 - 4.86 = 30.14 \text{ hours}$$

Comment. The arithmetic means calculated by both the methods are same since they are calculated for the same grouped data.

5.3.7. Weighted Arithmetic Mean. In ordinary arithmetic mean, equal importance is given to all the observations. But in practice relative importance of all the observations are not the same. In such situation we compute weighted arithmetic mean. The term 'weight' stands for the relative importance of the different observations. Weighted mean is especially useful in problems relating to the construction of index numbers and standardized birth and death rates.

Definition. Suppose x_1, x_2, \dots, x_k are k values of a variable x whose relative importance are measured by the weights w_1, w_2, \dots, w_k respectively, then the weighted arithmetic mean is computed by the following formula :

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

It is to be noted that there is similarity between the formula of weighted mean and the mean of a frequency distribution. Actually, the formula for mean of a frequency distribution can be considered as a special case of weighted mean. Frequency of a class is considered as weight of the mid-point of that class.

Example 5.3.11. A contractor employs three types of worker say male, female and children. To a male worker he pays Tk. 125 per day, to female worker Tk. 100 per day and to a child worker Tk. 75 per day. The numbers of male, female and child workers hired by the contractor are 15, 25 and 35 respectively. What is the average wage per day paid by the contractor?

Solution. The simple arithmetic mean of the wage is

$$\bar{x} = \frac{125 + 100 + 75}{3} = 100 \text{ Tk. per day.}$$

It is not the correct answer to the problem. If the numbers of male, female and child workers are the same, this answer would be correct.

For example, if the contractor hired 20 workers in each category, then the weighted mean is

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{20 \times 125 + 20 \times 100 + 20 \times 75}{20 + 20 + 20} = \frac{6000}{60} = 100 \text{ Tk. per day.}$$

It is the same as the simple mean.

Here the numbers of male, female and child workers are different. The appropriate mean is the weighted mean. It is calculated as follows:

Worker	Wage per day : x	No. of workers : w	wx
Male	125	15	1875
Female	100	25	2500
Boy	75	35	2625
		75	7000

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{7000}{75} = 93.33 \text{ Tk. per day.}$$

The weighted mean is less than the simple mean, since the weight of the child workers is more than the male and female.

Example 5.3.12. The unit price and the quantity of 7 food items consumed by a family per month are as follows.

Food items	Price per kg (in taka)	Quantity consumed (in kg)
Rice	25	20
Wheat	18	8
Sugar	30	3
Potato	15	5
Cereal	45	3
Salt	10	2
Oil	45	5

Compute the average price of the food items consumed by the family per month. Also compute the average price of the food items considering equal weight to each item and comment.

Solution. The weighted arithmetic mean of the price of the food items is

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Table for calculation of weighted arithmetic mean.

Food item	Price : x	Quantity : w	Wx.
Rice	25	20	500
Wheat	18	8	124
Sugar	30	3	90
Potato	15	5	75
Cereal	45	3	135
Salt	10	2	20
Oil	45	5	225
Total	188	46	1169

The weighted arithmetic mean of the price of the food items is

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{225}{46} = 20.88 \text{ Tk.}$$

But the simple arithmetic mean of the price of the food items considering equal weight to each item is

$$\bar{X} = \frac{\sum x}{n} = \frac{188}{7} = 26.86 \text{ Tk.}$$

Comment. Here weighted arithmetic mean is the appropriate average. Weighted mean provides a more realistic answer than the ordinary mean.

5.3.8. Merits and demerits of Arithmetic Mean.

Merits. It is the most popular and widely used average in practice. It has the following merits:

1. It is easy to understand
2. It is easy to calculate
3. It is based on all the observations
4. It is rigidly defined
5. It is capable of further algebraic treatment.
6. It is less affected by sampling fluctuation.

It is the best measure of average among all the averages. However it has some limitations too.

Demerits or limitations of Arithmetic mean

1. It is affected by extreme values.
2. It cannot be computed in case of open-ended class interval of a frequency distribution.
3. It is not a good measure of central tendency in case of highly skewed distribution.
4. It cannot be calculated for qualitative data.
5. It cannot be found graphically.

5.3.9. Some Mathematical Properties of Arithmetical Mean. Three most important mathematical properties of arithmetic mean are:

1. The algebraic sum of the deviations of all the observations about the arithmetic mean is always zero. Symbolically, if x_1, x_2, \dots, x_n are n observations of a set of data and if \bar{x} is the arithmetic mean, then $\sum(x_i - \bar{x}) = 0$.

Example 5.3.13. Let 3, 4, 5, 6, 7 be a set of data.

$$\text{It is easily seen that } \bar{x} = \frac{\Sigma x}{n} = \frac{25}{5} = 5$$

X	(x - \bar{x})
3	$3 - 5 = -2$
4	$4 - 5 = -1$
5	$5 - 5 = 0$
6	$6 - 5 = 1$
7	$7 - 5 = 2$
$\Sigma x = 25$	$\Sigma(x - \bar{x}) = 0$

This property of arithmetic mean enables one to check its accuracy. If this property holds, the computation of arithmetic mean is considered to be accurate, otherwise inaccurate. This implies that arithmetic mean is amenable for further algebraic treatment.

2. The sum of the squared deviations of all the observations from the arithmetic mean is minimum.

Symbolically, $\sum(x - \bar{x})^2 \leq \sum(x - a)^2$, where a is any arbitrary value other than \bar{x} .

Example 5.3.14. Suppose 4, 6, 8, 7, 10 are five values of a variable. Show that sum of square of deviation from the arithmetic mean is minimum.

Solution. The arithmetic mean of the five values is

$$\bar{x} = \frac{5+6+8+7+9}{5} = \frac{35}{5} = 7.$$

Let us take two more values 6 and 8 for finding the deviations.

x	x-7	(x-7) ²	x-6	(x-6) ²	x-8	(x-8) ²
5	-2	4	-1	1	-3	9
6	-1	1	0	0	-2	4
8	1	1	2	4	0	0
7	0	0	1	1	1	1
9	2	4	3	9	2	4
Total	0	10	5	15	-2	18

It is easily seen that, $\sum (x - 7)^2 \leq \sum (x - 6)^2 \leq \sum (x - 8)^2$

Hence is the proof.

3. If \bar{x}_1 and \bar{x}_2 are two arithmetic means of two related sets of observations, and n_1 and n_2 are the number of observations, then the combined arithmetic mean of the two sets is obtained by the following formula:

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

\bar{x} = combined mean of two sets,

\bar{x}_1 = Arithmetic mean of the first set,

\bar{x}_2 = Arithmetic mean of the second set,

n_1 = Number of observation in the first set,

n_2 = Number of observation in the second set.

If there are k such groups with means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ with number of observations n_1, n_2, \dots, n_k respectively, then, the combined mean of the k groups is given by the formula

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum n_k \bar{x}_k}{\sum n_k}$$

Example 5.3.15. In a garments factory, the mean wages of male and female workers per month are Tk. 4,000.00 and Tk. 3,500 respectively. The numbers of male and female workers are 65 and 125 respectively. Find the average monthly wage of the workers.

Solution. Here $\bar{x}_1 = 4000$, $\bar{x}_2 = 3500$, $n_1 = 65$ and $n_2 = 125$.

The combined mean,

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{65 \times 4000 + 125 \times 3500}{65 + 125}$$

$$= \frac{260000 + 437500}{190} = \frac{697500}{190} = \text{Tk. } 3671.05 \text{ per month.}$$

Example 5.3.16. In a factory, there are three types of workers-male, female and children. The average wage of a male worker per day is Tk. 200.00; to a female worker Tk.150 and to a child worker is Tk.100. The numbers of male, female and child workers are 25, 50 and 15 respectively. Find the average wage of the workers per day.

Solution. Here $\bar{x}_1 = 200$, $\bar{x}_2 = 150$, $\bar{x}_3 = 100$, $n_1 = 25$, and $n_2 = 50$ and $n_3 = 15$

Then the average wage of the workers is

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3} = \frac{25 \times 200 + 50 \times 150 + 15 \times 100}{25 + 50 + 15}$$

$$= \frac{5000 + 7500 + 1500}{90} = \frac{14000}{90} = \text{Tk. } 155.56 \text{ per day.}$$

5.4. Median

Definition. Median is the middle most value of a set of observations when the values are arranged in order of magnitude. That means, it divides the whole ordered observations into two equal parts, half of the observations are greater than or equal to it. It is also called a position or location measure of central tendency.

5.4.1. Median from ungrouped data. First arrange the observations in ascending or descending order of magnitude (both arrangement would give the same answer).

Rule 1. Now if the number of observations n is odd, then there will be a single middle value, which is the median, and its position will be $(\frac{n+1}{2})$ th ordered observation of the series.

Rule 2. If the number of observations n of data set is even, then the position of the median will be the arithmetic mean of the $(\frac{n}{2})$ th and $(\frac{n}{2} + 1)$ th ordered observations.

Another rule for computing median. When $n/2$ of the ordered array is integer, then median is the mean of the $(n/2)$ th and $\{(n/2) + 1\}$ th ordered

observations. But when $n/2$ is not an integer, round up it to the next higher integer and the observation corresponding to this integer is the median.

Example 5.4.1. The following data give the monthly wages in taka of 7 workers of a factory :

Wage (in Taka): 2700, 2750, 2680, 2790, 2760, 2720, 2740.

Compute median wage of the workers.

Solution. First we arrange the data set in ascending order of magnitude. The ordered array is :

2680, 2700, 2720, 2740, 2750, 2760, 2780.

Here n is odd. The median of the data set is, $\frac{n+1}{2}$ th ordered observation =

$\frac{7+1}{2}$ th ordered observation = 4th ordered observation = Tk. 2740 per month.

Example 5.4.2. The following data refer to the profits of a store in thousand taka for the last 12 months:

3, 6, 8, 9, 6, 10, 5, 12, 9, 8, 11, 7

Compute median profit of the store.

Solution. First we arrange the observations in ascending order of magnitude. The ordered array is

3, 5, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12

Here $n = 12$ is even.

$$\text{Median} = \frac{\frac{n}{2} \text{th observation} + \frac{n+1}{2} \text{th observation}}{2}$$

$$= \frac{6\text{th observation} + 7\text{th observation}}{2} = \frac{8+8}{2} = \text{Tk. 8 Thousand}$$

Interpretation. 50% of the monthly profit is Tk. 8 thousand or less.

Matched problems to solve. Find median form the following two sets of observations :

- | | |
|-----------------------|----------|
| (i) 8, 4, 5, 7, 9, 10 | Ans. 7.5 |
| (ii) 5, 9, 2, 8, 6 | Ans. 6 |

5.4.2. Computation of Median from grouped data of discrete variable. Median from a discrete frequency distribution can be obtained by using the rule1 and rule 2. But to locate the position of median we have to construct a cumulative frequency table.

We shall cite one example to clarify the matter.

Example 5.4.3. A survey was conducted on 100 school teachers to know the median number of children of their family. The results of the survey are presented in the following frequency table:

No. of children	0	1	2	3	4	5
No. of family	5	15	25	35	16	4

Solution. First we have to construct a cumulative frequency table for locating the position of median.

Variable (No. of children) : x	Frequency (No. of family) : f	Cumulative frequency
0	5	5
1	15	20
2	25	45
3	35	80
4	16	96
5	4	100

Here $n = 100$ is even, and then the position of median will be the arithmetic mean of the 50th and 51st ordered observations. It is noted that the frequency distribution is always constructed orderly. It is seen from the third column (cumulative frequency column) that the positions of both the observations 50th and 51st correspond to the no. of children 3. Hence median is $(3+3)/2 = 3$. That means, 50% of the teachers have children 3 or less.

The position of median can also be easily located from the stem and leaf plot.

Example 5.4.4. The prices in taka of 20 different brands of walking shoes are given below:

45, 70, 70, 55, 75, 73, 70, 65, 68, 60,
74, 83, 80, 58, 68, 85, 90, 64, 75, 82

Construct a stem and leaf plot to display the distribution of the data and compute median price of the walking shoes.

Solution. The stem and leaf plot of the problem has been displayed in problem 4.9.1 of chapter 4. The stem and leaf plot of the problem is

Stem Leaf

4	5
5	5 8
6	0 4 5 8 8
7	0 0 0 0 3 4 5
8	0 2 3 5
9	0

Here $n = 20$ is even. So the median will be the mean of the 10th and 11th ordered observations. Since the observations are orderly arranged in stem and leaf plot, it is easily seen that the 10th and 11th observations are both 70. The median of the data set is Tk. 70. Hence, the median price of the walking shoes is Tk. 70. That means, the prices of 50% of the walking shoes are less than or equal to Tk. 70.

Example 5.4.5. The following data represent the amount of insurance (in units of thousand taka) purchased by 30 people from an insurance company in a given week..

31, 44, 51, 35, 76, 84, 110, 50, 56, 61,
40, 48, 61, 85, 90, 92, 40, 65, 120, 125,
100, 105, 115, 70, 77, 120, 75, 80, 92, 115

Construct a stem and leaf plot to display the data and find median.

Solution. The stem and leaf plot of the problem given in example 4.28 is displayed below:

Stem	Leaf
3	1 5
4	0 0 4 8
5	0 1 6
6	1 1 5
7	0 5 6 7
8	0 4 5
9	0 2 2
10	0 5
11	0 5 5
12	0 0 5

3 in stem and 1 in leaf means 31, and so on.

Here, $n = 30$ is even. So the median is the mean of the ordered observations 15th and 16th. Here, the observations are 76 and 77. Hence the median is :

$$\frac{76+77}{2} = \text{Tk } 76.5 \text{ thousand.}$$

That means, 50% of the people purchased insurance Tk 76.5 thousand or less.

Example 5.4.6. The following data represent the lives of 41 similar car batteries recorded to the nearest tenth of a year.

2.2 4.1 3.5 4.5 3.2 3.7 3.0 2.6 3.4 1.6
3.1 3.3 3.8 3.1 4.7 3.7 2.5 4.3 3.4 3.6
2.9 3.3 3.9 3.1 3.3 3.1 3.7 4.4 3.2 4.1
1.9 3.4 4.7 3.8 3.2 2.6 3.9 3.0 4.2 3.5
4.8

Construct a stem and leaf plot to display the data and locate median.

Solution. The stem and leaf plot display of the problem is

Stem	Leaf
1	6 9
2	2 5 6 6 9
3	0 0 1 1 1 2 2 2 3 3 4 4 4 5 5 6 7 7 7 8 8 9 9
4	1 1 2 3 4 5 7 7 8

1 in stem and 6 in leaf means 1.6 and so on.

Here $n = 41$ is odd. So the position of the median is $\frac{41+1}{2} = 21$ th ordered observation. It is 3.4. Hence the median life of the battery is 3.4 years. That is 50% of the battery has life 3.4 years or less.

5.4.3. Computation of median from a grouped data of continuous variable. First we have to construct a cumulative frequency table. Then we have to identify the median class. Median class is the most important class for computing median. The class which contains $\frac{n}{2}$ th observation is called the median class. Here we always use $\frac{n}{2}$ instead of $\frac{n+1}{2}$ to locate median because $\frac{n}{2}$ divides the whole area of the curve into two equal parts in case of continuous variable.

The formula for computing median is

$$M_e = L + \frac{n/2 - F}{f} \times c .$$

Here M_e = median, L = lower limit of the median class, n = Total of observations, F = Cumulative frequency of pre-median class, f = Frequency of the median class, c = Width of the median class.

Example 5.4.7. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory.

Number of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Compute median of the frequency distribution.

Solution. First we construct a cumulative frequency table with the frequency distribution. The cumulative frequency distribution table is

Class interval	Frequency : f	Cumulative frequency : F
30 - 55	3	3
55 - 80	4	7
80 - 105	6	13
105 - 130	9	22
130 - 155	12	34
155 - 180	11	45
180 - 205	5	50

Here $n = 50$, $\frac{50}{2} = 25$ th observation lies in the class 130-155. Hence the median class is 130-155. That is 25th observation is in cumulative frequency 34 and the corresponding class is 130-155.

Here, $L = 130$, $\frac{n}{2} = 25$, $F = 22$, $f = 12$ and $c = 25$

$$M_e = L + \frac{n/2 - F}{f} \times c = 130 + \frac{25 - 22}{12} \times 25$$

$$= 130 + 6.25 = 136.25 \text{ hours per month.}$$

Interpretation. 50% of the workers worked for 136.25 hours or less per month.

5.4.4. Locating Median graphically.

Median from ogive Curve

Generally median is obtained graphically from ogive curve.

It can be determined graphically by applying any of the following two methods:

1. Draw two ogives- one by 'less than' method and the other by 'more than' method. Draw a perpendicular on the X-axis from the point where the two curves intersect each other. The point where this perpendicular touches the X-axis gives the value of the median.
2. Draw one ogive usually by 'less than' method. Plot the upper limit of the variable on the X-axis and the cumulative frequency on the Y-axis. Locate a point by $n/2$ on the Y-axis and from this point draw a horizontal line parallel to the X-axis on the cumulative frequency curve. Draw a perpendicular on the X-axis from the point where it meets the ogive. The point at which the perpendicular cuts the X-axis is the median.

Example 5.4.8. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

- (i) Draw ogives by 'less than' method and by 'more than' and locate median from them.
- (ii) Draw an ogive by 'less than' method and locate median

Solution. First we construct a 'less than' and 'more than' cumulative frequency table.

No. of hours	Less than Cumulative frequency	No. of hours	More than cumulative frequency
Less 30	0	More than 30	50
Less 55	3	More than 55	47
Less 80	7	More than 80	43
Less 105	13	More than 105	37
Less 130	22	More than 130	28
Less 155	34	More than 155	16
Less 180	45	More than 180	5
Less 205	50	More than 205	0

Now plot the class intervals on the X-axis and the cumulative frequency on the Y-axis.

- (i) Draw two ogives by 'less than' and by 'more than' methods on the same graph paper. Now draw a perpendicular from the intersecting point A on the X-axis. The point at which the perpendicular cuts the X-axis is the median. Here it is

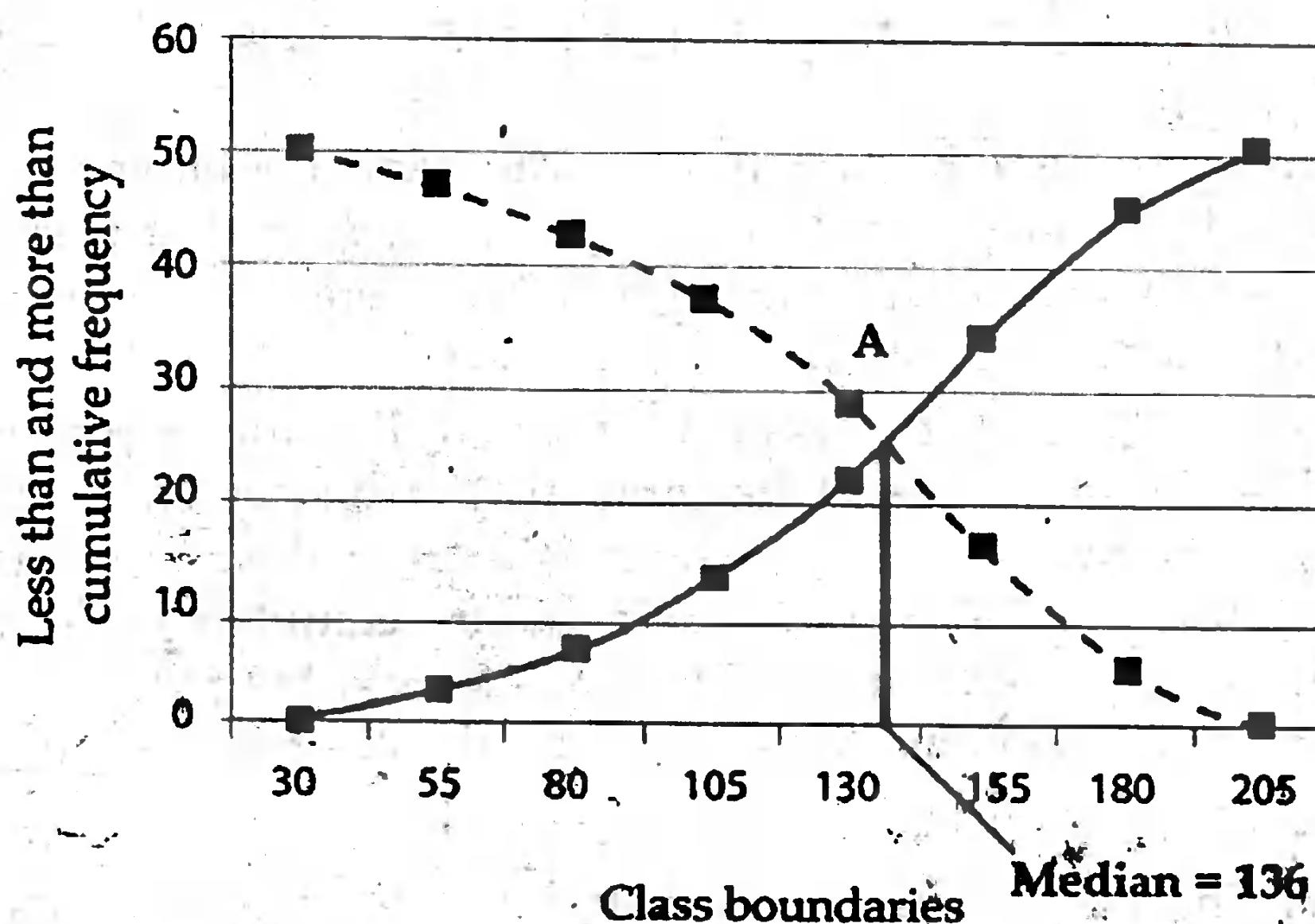


Fig. 5.1. Median from two ogives.

It is observed that two ogives intersect at A, and the value at which the perpendicular from A to X-axis cuts the X-axis is about 136. So median is 136.

- (ii) Now plot the class intervals on the X-axis and the 'less than' cumulative frequency on the Y-axis. Plot points above the class intervals according to their cumulative frequencies. Join the point's free hand to get the required ogive. Locate a point $n/2 = 50/2 = 25$ on the Y-axis and from this point draw a line parallel to the X-axis on the ogive. Now draw perpendicular on the X-axis from the point at which the line cuts on the ogive. The point at which the perpendicular cuts the X-axis is the median. Here it is

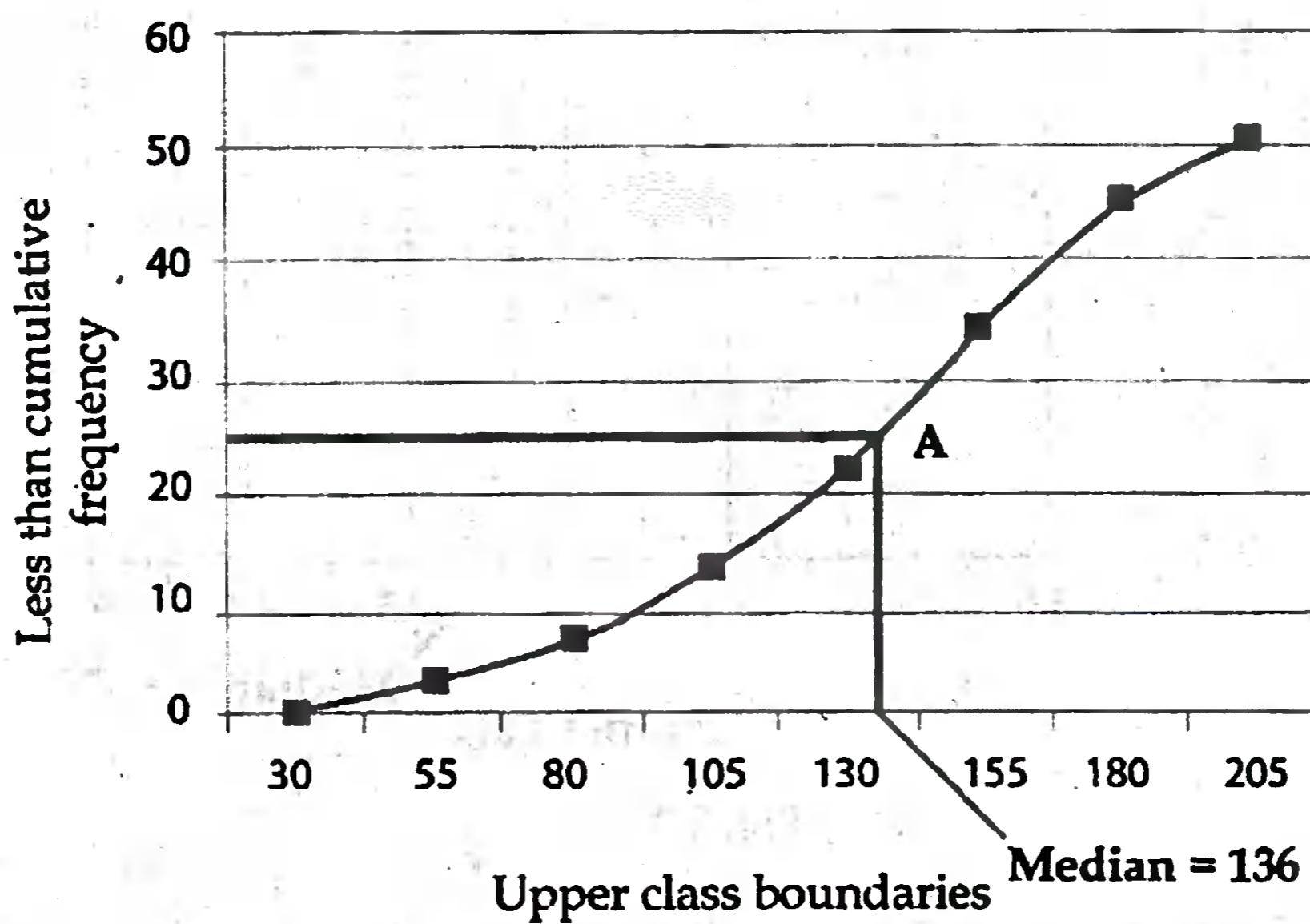


Fig. 5.2. Median from less than ogive.

Median from Histogram

The median of a frequency distribution can also be located from the histogram if the frequency of all classes exists. Then median is that value of the variable, which divides the whole area of the histogram into two equal parts. Now we shall cite an example and show how median can be located from a histogram.

Example 5.4.9. Draw a histogram and find median from the following frequency distribution

Class interval	Frequency
3.5 - 4.5	3
4.5 - 5.5	1
5.5 - 6.5	2
6.5 - 7.5	4
7.5 - 8.5	3

8.5 - 9.5	2
Total	15

Solution. We first draw the histogram with the frequency distribution. It is easily seen that the total area of the histogram is 15, which is just the sum of the frequencies, since all the rectangles have a base of width 1. The area to the left of the median must be half the total area, that is $15/2 = 7.5$.

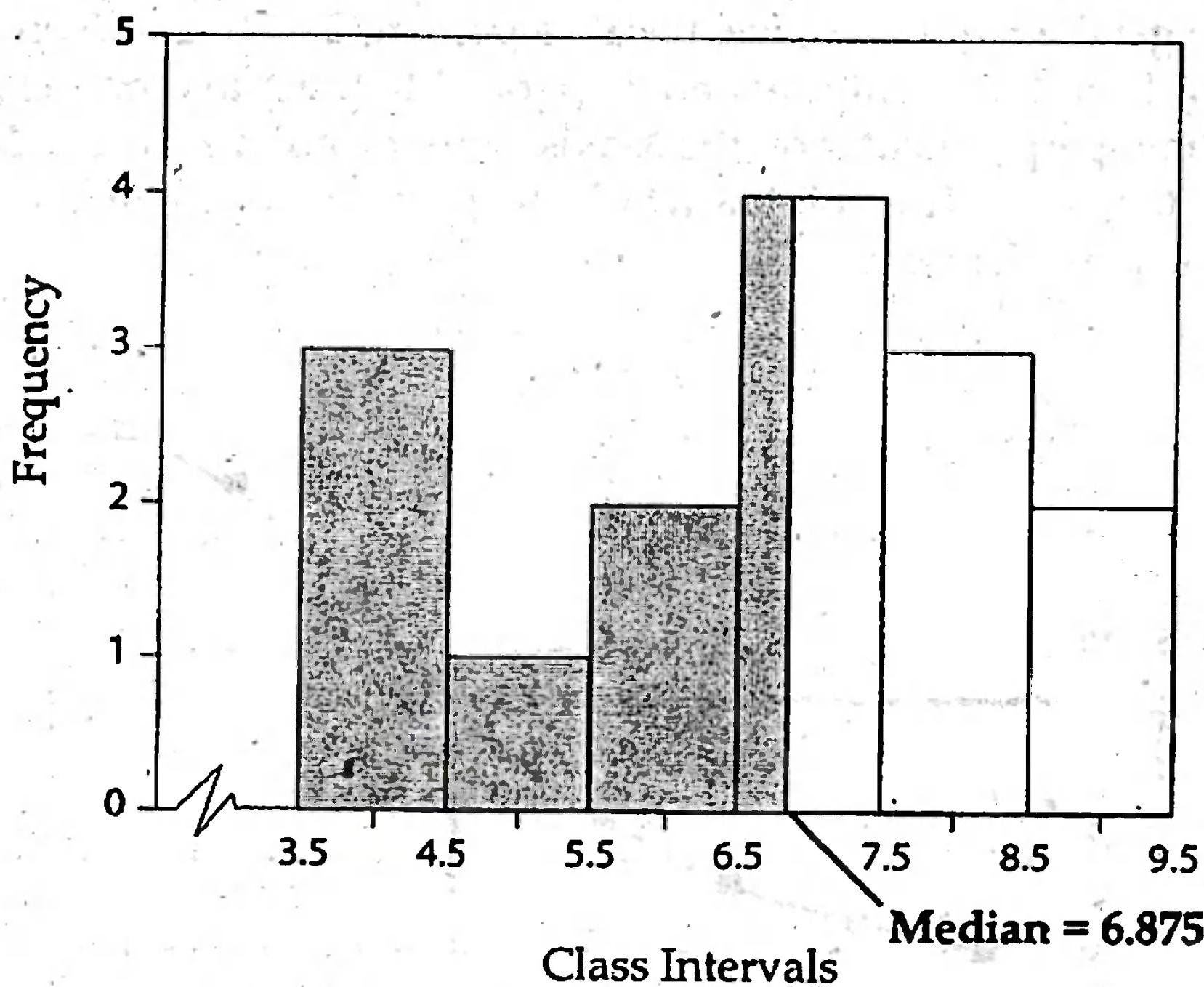


Fig. 5.3.

Looking at the histogram of figure (5.3), we see that, the median M lies between 6.5 and 7.5. Thus, the area to the left of M , which is the sum of the shaded area in Figure 5.3 must be 7.5, That is :

$$\begin{aligned}
 (1)(3) + (1)(1) + (1)(2) + (M-6.5)(4) &= 7.5 \\
 6 + 4M - 26 &= 7.5 \\
 4M &= 27.5 \\
 M &= 27.5/4 = 6.875
 \end{aligned}$$

Hence the median for the frequency distribution is 6.875.

Remarks. Usually, we locate median from ogive curve but it can also be locate from the histogram.

A matched problem to solve. Find the median for the grouped data in the following table:

Class interval	Frequency
3.5 – 4.5	4
4.5 – 5.5	2
5.5 – 6.5	3
6.5 – 7.5	5
7.5 – 8.5	4
8.5 – 9.5	3

5.4.5. Merits and demerits of median.

Merits

Median is a positional measure. It has the following merits:

1. It is easy to understand.
2. It is easy to calculate.
3. It is rigidly defined.
4. It is not affected by extreme values.
5. It can be computed in open-end frequency distribution.
6. It can be obtained from ogive. That means it can be found graphically.
7. It is a suitable measure of location in case of very skewed distribution.
8. The position of median can be easily located when a qualitative variable is measured in ordinal scale.

Demerits or limitations

1. It is not based on all the observations.
2. It is not capable of algebraic treatment.
3. It is more affected by sampling fluctuations.

5.4.6. Advantage of median over arithmetic mean. Arithmetic mean is the best measure of central tendency. But there some situations where median is superior to arithmetic mean.

1. In presence of outliers of a set of data, median is better than arithmetic mean
2. For highly skewed distribution, median is superior to arithmetic mean.
3. For open-ended distribution, arithmetic mean is not possible to calculate unless some assumption is made but median can be easily computed.
4. The position of median can be easily located when a qualitative variable measured in ordinal scale.

5.4.7. Median from ordinal data. Median can be obtained from the from qualitative data when measured in ordinal scale. Now we shall cite an example.

Example 5.4.10. The frequency distribution of letter grades obtained by 1180 students of a school in S.S.C. examination is as follows:

Letter grade	No. of student
A ⁺	30
A	65
A ⁻	160
B ⁺	200
B	350
C ⁺	250
C	75
D	50

Find an appropriate measure of central tendency.

Solution. This is a frequency distribution of qualitative data and the variable letter grade is measured in ordinal scale. Arithmetic mean is not possible with this data set. One of the appropriate measures of central tendency is the median. The cumulative frequency distribution table is given below:

Letter grade	No. of student f	Cumulative frequency
A ⁺	30	30
A	65	95
A ⁻	160	255
B ⁺	200	455
B	350	805
C ⁺	250	1055
C	75	1130
D	50	1180

Median = Size of $(n/2)$ th item = B

Hence, B is the median letter grade of the students. That means, 50% of the students obtained B or lower grade.

Remarks. Here, it is mentionable that median cannot be calculated for categorical data measured in nominal scale since data cannot be arranged in ascending or descending order.

A matched problem to solve. The health conditions of 375 workers of a factory are summarized in a frequency table given below:

Health condition	Number of workers f
Excellent	20
Good	72
Average	158
Poor	125

Find the median health status of the workers.

Outliers. Extremely large or small observations of a set of data are called outliers or wild observations or extreme values. Actually unusual observations are called outliers, wild or extreme values. These types of observations are quite dissimilar to other observations.

Now we cite an example and show how arithmetic mean is misleading as a measure of central tendency in presence of outlier.

Example 5.4.11. Suppose the annual salaries in taka of 7 employees in a small company are 17,000, 20,000, 28,000, 18,000, 1,000, 120,000, and 24,000.

Find arithmetic mean and median salary of the employees and comment.

Solution. The arithmetic mean of the salary is :

$$\bar{x} = \frac{17,000 + 20,000 + 28,000 + 18,000 + 18,000 + 120,000 + 24,000}{7}$$

$$= \frac{245,00}{7} = \text{Tk.} 35,000$$

Six of the salaries are below the average. The one large salary (extreme large value) Tk.120,000 distorts the results.

Computation of Median: We arrange the salaries in ascending order of magnitude. The ordered salaries are:

17,000	
18,000	
18,000	
20,000	Median = Tk. 20,000
24,000	
28,000	
120,000	Arithmetic Mean = Tk. 35,000

Median is the $\frac{7+1}{2} = 4$ th ordered observation of the data set which is

Tk.20,000. It is seen that median is the most representative salary of the employees. Here median is a better measure of central tendency than the arithmetic mean.

Example 5.4.12. The following table gives the frequency distribution of the yearly income in thousand taka of 603 persons in a community:

Income (in thousand Taka)	No. of persons f
Below 30	69
30 - 40	167
40 - 50	207
50 - 60	65
60 - 70	58
70 - 80	27
80 and above	10

- Calculate an appropriate measure of central tendency.

Solution. Here it is an open-ended frequency distribution. We cannot find the mid-points of the first and last class intervals of the frequency distribution without any assumption. So it is not possible to compute arithmetic mean of the grouped data. But we can easily find the median of the grouped data. For this, first we construct a cumulative frequency distribution table.

Cumulative frequency distribution table for computing median

Income (in thousand Taka)	No. of persons: f	Cumulative frequency: F
Below 30	69	69
30 - 40	167	236
40 - 50	207	443
50 - 60	65	508
60 - 70	58	566
70 - 80	27	593
80 and above	10	603

The formula for computing median for a grouped data is

$$M_e = L + \frac{n/2 - F}{f} \times c$$

$$\text{Median} = \text{size of } \frac{n}{2} \text{ th item} = \frac{603}{2} = 301.5 \text{ th item}$$

Median class corresponding to this item is 40 - 50.

Here $L = 40$, $n/2 = 301.5$, $F = 236$, $f = 207$, $c = 10$

$$\begin{aligned} \text{Median} &= M_e = 40 + \frac{301.5 - 236}{207} \times 10 \\ &= 40 + 3.16 = 43.16 \text{ thousand Tk. per annum.} \end{aligned}$$

Remarks. We can also compute mode and other positional measures in case of open-ended frequency distribution, which will be discussed in the subsequent sections of this chapter.

5.5. Mode

Mode is another important measure of central tendency. It is a value of the variable which occurs the maximum number of times, i.e., having highest frequency.

Definition. Mode is that value of a variable, which has highest frequency. According to Zizek, "Mode is the value occurring most frequently in a series and around which the other items are distributed most densely". There may be a unique mode, several modes or essentially no mode. The distribution for which there exists only one mode, is called unimodal distribution. Similarly, for two or more modes, the distributions are known as bimodal or multimodal distribution respectively. Like median mode is not influenced by extreme values.

5.5.1. Mode from ungrouped small set of data. First arrange the observations in ascending order of magnitude. Then count the number of times of repetition of each observation. The observation that has the highest frequency is called the mode.

Example 5.5.1. Find Mode, median and mean of the data sets

- (a) 4, 5, 5, 5, 6, 6, 7, 8, 12
- (b) 1, 2, 3, 3, 3, 5, 6, 7, 7, 7, 23
- (c) 1, 3, 5, 6, 7, 9, 11, 15, 16

Solution. The mode, median and mean of the three data sets are calculated and shown in the table given below.

Data Set	Mode	Median	Mean
(a) 4, 5, 5, 5, 6, 6, 7, 8, 12	5	6	6.44
(b) 1, 2, 3, 3, 3, 5, 6, 7, 7, 7, 23	3, 7	5	6.09
(c) 1, 3, 5, 6, 7, 9, 11, 15, 16	None	7	8.11

The data set (a) in example 5.5.1 is called unimodal, since there is one mode. Data set (b) is called bimodal, since there are two modes. On the other hand the data set (c) has no mode.

Remarks. For many sets of data median lies between mode and mean. But this is not always so.

Example 5.5.2. The marks obtained by 7 students in an examination were 52, 89, 96, 93, 89, 92, and 99. Find mean, median and mode of the marks.

Solution. The mean mark of the student is

$$\bar{x} = \frac{52+89+96+93+89+92+99}{7} = \frac{610}{7} = 87.14$$

The values of the data set in ascending order is 52, 89, 89, 92, 93, 96, 99

Mode = 89, since it occurs two times in the data set.

Median = 92, since it is the middle most value of the ordered observations.

Here median does not lie between mode and mean.

Remarks. Median does not always lie between mode and mean. But for a moderately skewed distribution median lies between mean and mode.

Matched problem to solve. Find the mean, median and mode for the sets of ungrouped data given below :

- (a) 1, 2, 2, 3, 3, 3, 3, 4, 4, 6
- (b) 2, 2, 2, 2, 2, 3, 4, 5, 6, 6,
- (c) 3, 4, 4, 4, 6, 6, 5, 5, 5, 7, 8
- (d) 2, 8, 5, 7, 3, 4, 9, 1, 6

5.5.2. Computation of mode from qualitative data. Mode is the only measure of central tendency, which can be computed for all kinds of qualitative data. We can easily construct frequency distribution of qualitative data. The category having the highest frequency is called the modal category.

Example 5.5.3. The frequency distribution of letter grades obtained by 1180 students of a school is as follows:

Letter grade	No. of student
A ⁺	30
A	65
A ⁻	160
B ⁺	200
B	350
C ⁺	250
C	75
D	50

Find an appropriate measure of central tendency.

Solution. We have already computed median from this data in example 5.4.10 We can easily find mode of this data set. The letter grade B is mode, since it has the highest frequency. The frequency of this letter grade is 350.

Example 5.5.4. A survey was conducted on 1454 people of Chittagong city having car. The frequency distribution of different colours of the cars is as follows:

Colour of the car	No. of persons f
Red	130
White	554
Black	141
Blue	213
Brown	257
Gold	159

Find median and modal colour of the car.

Solution. We cannot find the median colour of the car, since the data cannot be arranged orderly. The appropriate measure of central tendency is the mode. Modal colour of the car is white since it has the frequency 554, which is the highest number among the all colours.

A matched problem to solve. A survey was conducted on 937 school students to study their favorite flavor of ice cream. The results of the survey are summarized in the following frequency table.

Flavor of ice cream	No. of students
Vanilla	129
Chocolate	360
Strawberry	79
Pistachio	95
Cherry	53
Almond mocha	221

Find an appropriate measure of central tendency.

Example 5.5.5. Find an appropriate measure of location from the following set of data

8.6 7.9 8.7 6.9 7.8 8.4 7.7 8.1 6.2 81.3 8.8

Solution: There exists an extreme value in the data set which is 81.3. Moreover, mode does not exist in this data set. Hence the appropriate measure of location is median. The orderly arranged data set is :

6.2, 6.9, 7.7, 7.8, 7.9, 8.1, 8.4, 8.6, 8.7, 8.8, 81.3

Here $n = 11$, then median is 6th observation of the arranged data set. Hence Median = 8.1.

5.5.3. Mode from grouped data. First we shall discuss how to find mode from grouped data of discrete variable. In this case mode is that value of the variable, which has the highest frequency.

Now we shall cite one example.

Example 5.5.6. The data given below are size of shoes sold by a shop in a week:

Size of shoes:	5	6	7	8	9	10	11
Number of shoes sold:	10	23	27	45	24	17	9

Find the modal size of the shoes sold by the shop per week.

Solution. It is easily seen that the modal size of the shoe is 8 and its frequency is 45. That is 8 is the most popular size of the shoe.

Calculation of Mode from grouped data in case of continuous variable. When we have a grouped data of continuous variable, the formula for computing mode is :

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Here M_o = Mode, L = Lower limit of the modal class, Δ_1 = Frequency difference between the modal class and pre-modal class, Δ_2 = Frequency difference between the modal class and post-modal class, i = the size of the modal class.

It is seen from the formula that the modal group or class is the important class for finding mode. Modal class is the class, which has the highest frequency.

Example 5.5.7. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Compute mode.

Solution. It is obvious from the frequency table that the class 130-155 contains the highest frequency. Hence the modal class is 130 - 155. The formula for finding mode is

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Here $L = 130$, $\Delta_1 = 12 - 9 = 3$, $\Delta_2 = 12 - 11 = 1$, $i = 25$

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 130 + \frac{3}{3+1} \times 25$$

$$= 130 + 18.75 = 148.75 \text{ hours per month.}$$

Hence the modal working time is 148.75 hours per month. That means, most of the workers worked for 148.75 hours.

Sometimes, mode of a frequency distribution is difficult to compute. For example, when the highest frequency of a frequency distribution lies in the first class or in the last class, we cannot compute mode with the above formula. In that case mode can be obtained by the following approximate formula given by Karl Pearson

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Usually this formula is applicable in case of moderately skewed distribution.

Example 5.5.8. In a moderately skewed distribution arithmetic mean and mode are 24.6 and 26.1 respectively. Find the value of the median and explain the reason for the method employed.

Solution. The relationship among mean, median and mode in a moderately skewed distribution is :

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Median can be easily found from this relationship. Here

$$26.1 = 3 \text{ Median} - 2(24.6)$$

$$3 \text{ Median} = 26.1 + 49.2 = 75.3$$

$$\text{Median} = 75.3/3 = 25.1$$

Example 5.5.9. In a moderately asymmetrical distribution the value of mode and median are 40 and 42.8 respectively. Find the value of mean of the distribution.

Solution. Here median = 42.8 and mode = 40. Hence the mean of the distribution is found from the relationship given below:

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

$$2 \text{ Mean} = 3 \times 42.8 - 40 = 128.4 - 40 = 88.4$$

$$\text{Mean} = 88.4/2 = 44.2$$

Example 5.5.10. The values of the arithmetic mean and median in a moderately symmetrical distribution are 20.2 and 18.8 inches respectively. Find the mode of the distribution.

Solution. Here Mean = 20.2 and Median = 18.8. Hence the mode of the distribution is

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

$$= 3(18.8) - 2(20.2)$$

$$= 56.4 - 40.4 = 16.$$

5.5.4. Locating Mode graphically. Mode of a frequency distribution can be located graphically from a histogram. The steps in finding mode are:

1. First draw a histogram of the frequency distribution.
2. Then locate modal class and the rectangle over this class by inspecting highest frequency.
3. Then draw two lines diagonally on the inside of the modal class rectangle, starting from each upper corner of the rectangle to the upper corner of the adjacent rectangle.
4. Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis, which gives us modal value.

Example 5.5.11. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Draw the histogram and locate mode from it.

Solution. We know from the Example 5.5.7 that the mode of the distribution is 148.75 hours per month.

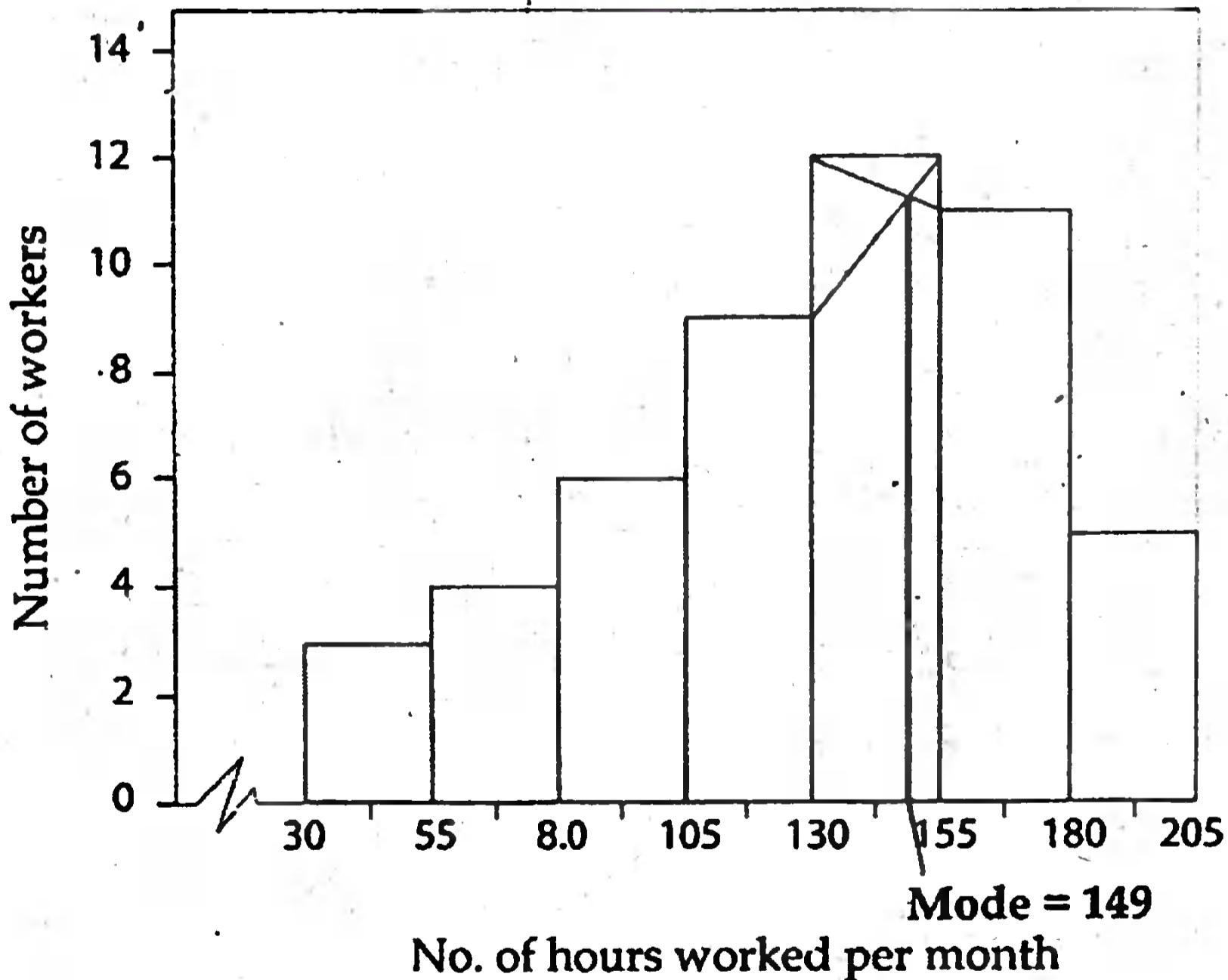


Fig. 5.4. Mode from Histogram.

From the histogram it is seen that the mode is 149.00 hours. Hence by both the methods we get approximately the same value of mode.

Remarks. Mode can also be obtained graphically from frequency polygon and frequency curve.

However, graphic method for determining mode can be used only when there is one class containing the highest frequency.

5.5.5. Merits and demerits of mode.

Merits

1. It is easy to understand.
2. It is easy to calculate.
3. It is not affected by extreme values.
4. It can be calculated for open-ended class interval.
5. It can be calculated graphically.
6. It can be calculated both for qualitative and quantitative data.

Demerits

1. It is not based on all the observations.
2. It is not clearly defined in case of bimodal or multimodal distribution.
3. Mode cannot be defined if each value of the variable occurs only once in a set of data.
4. It is affected by sampling fluctuation.
5. It is not suitable for further algebraic treatment.
6. Mode cannot be calculated if the highest frequency lies in the first or last class in a frequency distribution.

There are many situations in which arithmetic mean and median fail to reveal the true characteristics of data. For example, when we talk of most common wage, most common income, and most common sales size of shop or size of a ready-made garment, most common height etc. we have in mind mode and not the arithmetic mean or median discussed earlier.

5.6. Some Other Positional Measures

Median is the most important positional measure. It divides the whole distribution into two equal parts. That is 50% observations are equal to or smaller than median and 50% observations are equal to or larger than the median. Other important positional measures are (i) Quartiles, (ii) Deciles and (iii) Percentiles. In general, quartiles, deciles and percentiles are all known as quantiles.

5.6.1. Quartiles: Quartiles divide the ordered data into 4 equal parts. So there are three quartiles. They are Q_1 , Q_2 and Q_3 . Q_2 is the median. Q_1 and Q_3 are called the first and third quartiles.

First Quartile. The first quartile, Q_1 , is a value for which 25% of the observations are equal to or smaller than Q_1 and 75% are equal to or larger than Q_1 .

Second quartile. Second quartile, Q_2 , is a value for which 50% of the observations are equal to or smaller than Q_2 and 50% are equal to or larger than Q_2 . That is it divides the ordered observations into two equal parts.

Third Quartile. The third quartile, Q_3 , is a value for which 75% of the observations are smaller than or equal to it and 25% are equal to or larger than it.

Quartiles from ungrouped data :

Step 1. First we arrange the observations in ascending order of magnitude (smallest value to largest value).

Step 2. For finding i th quartile, Q_i ($i = 1, 2, 3$), we compute an index

$$j = \frac{in}{4}; \quad i = 1, 2, 3$$

Step 3. If j is an integer, Q_i is the mean of the j th and $(j+1)$ th ordered observations.

Step 4. If j is not an integer, Q_i is the value of the ordered observation corresponding to the next integer greater than j .

Different steps for finding Q_1 :

Step 1. Arrange the observations in ascending order of magnitude.

Step 2. Compute the index $j = \frac{n}{4}$.

Step 3. If $j = \frac{n}{4}$ is an integer, Q_1 is the mean of the j th and $(j + 1)$ th observations of the ordered array.

Step 4. If $j = \frac{n}{4}$ is not an integer, Q_1 is the value corresponding to the next higher integer than j of the ordered array.

Example 5.6.1. The following data refers to the monthly starting salaries in Taka for a sample of 12 business school graduates: 7850, 7950, 8050, 7880, 7755, 7710, 7890, 8130, 7940, 8325, 7920, and 7880. Find Q_1 , Q_2 and Q_3 .

Solution. Stepwise procedure for finding quartiles is followed here.

Step 1. First we arrange the data in ascending order,

7710, 7755, 7850, 7880, 7880, 7890, 7920, 7940, 7950, 8050, 8130, 9325,

Step 2. For finding Q_1 , we compute $j = \frac{n}{4}$

Step 3. Here $j = \frac{n}{4} = \frac{12}{4} = 3$ is an integer.

Step 4. Q_1 is the mean of the 3rd and 4th observations of the ordered array.

That is, $Q_1 = \frac{7850 + 7880}{2} = 7865$ Tk. per month.

Similarly,

For Q_2 or median, $\frac{2n}{4} = \frac{2 \times 12}{4} = 6$ is an integer. Hence median is the average of the 6th and 7th observations of the ordered array. That is

Median = $Q_2 = \frac{7890 + 7920}{2} = 7905$ Tk. per month.

For third quartile, $\frac{3n}{4} = \frac{3 \times 12}{4} = 9$ is an integer. Hence Q_3 is the mean of the 9th and 10th ordered observations. Here 9th and 10th observations are 7950 and 8050 respectively. Hence,

$Q_3 = \frac{7950 + 8050}{2} = 8,000$ Tk. per month.

5.6.2. Deciles. Deciles divide the total ordered data into 10 equal parts. So there are nine deciles. They are denoted by D_1, D_2, \dots, D_9 . D_5 is the median.

Deciles from ungrouped data: Different steps for finding deciles are:

Step 1. As in case of median and quartiles, we arrange the data in ascending order of magnitude.

Step 2. For i th decile, D_i ($i = 1, 2, \dots, 9$), we compute an index

$$j = \frac{in}{10}; i=1, 2, 3, \dots, 9$$

Step 3. If j is an integer, D_i is the mean of the j th and $(j+1)$ th ordered observations.

Step 4. If j is not an integer, D_i is the value of the ordered observation corresponding to the next integer greater than j .

Example 5.6.2. The following data refer to the monthly starting salaries in taka for a sample of 12 business school graduates: 7850, 7950, 8050, 7880, 7755, 7710, 7890, 8130, 7940, 8325, 7920, 7880. Compute D_1, D_5, D_8 .

Solution.

Step 1. First we arrange the data in ascending order of magnitude,

7710, 7755, 7850, 7880, 7880, 7890, 7920, 7940, 7950, 8050, 8130, 9325,

Step 2. For finding D_1 , we compute

$$j = \frac{n}{10} = \frac{12}{10} = 1.2$$

Step 3. Here $j = 1.2$ is not an integer, the next integer is 2. The second ordered observation is 7755. Hence D_1 is 7755. For finding D_5 , we compute the index

$$j = \frac{5n}{10} = 6$$

Here j is an integer. The mean of the 6th and 7th ordered observations is D_5 . Hence

$$D_5 = \frac{7890 + 7920}{2} = 7905 \text{ Tk. per month.}$$

D_5 and Q_2 are the median which we got before.

For finding D_8 we compute the index

$$j = \frac{8n}{10} = \frac{8 \times 12}{10} = 9.6$$

Here $j = 9.6$ is not an integer. The next integer is 10. The 10th ordered observation is 8050. Hence the 8th decile is 8050 Tk. $D_8 = \text{Tk. } 8050$ suggests that salaries of 80% of the business graduates are less than or equal to Tk. 8050.

5.6.3. Percentiles. Percentiles divide the total ordered data into 100 equal parts. So there are 99 percentiles. They are denoted by P_1, P_2, \dots, P_{99} . P_{50} is the median or 5th decile or 2nd quartile of the distribution, P_{25} is the first quartile value, P_{75} is the third quartile value, P_{20} is the second decile and so on.

Percentiles from ungrouped data : Different steps for finding percentiles are

Step 1. As before we arrange the data in ascending order of magnitude.

Step 2. For i th percentile, P_i ($i = 1, 2, \dots, 99$), we compute an index

$$j = \frac{in}{100}; \quad i = 1, 2, 3, \dots, 99$$

Step 3. If j is an integer, P_i is the mean of the j th and $(j+1)$ th ordered observations.

Step 4. If j is not an integer, P_i is the value of the ordered observation corresponding to the next integer greater than j .

Example 5.6.3. The following data refer to the monthly starting salaries in Taka for a sample of 12 business school graduates: 7850, 7950, 8050, 7880, 7755, 7710, 7890, 8130, 7940, 8325, 7920, 7880. Compute P_{11}, P_{50}, P_{80} .

Solution.

Step 1. First we arrange the data in ascending order of magnitude,

7710, 7755, 7850, 7880, 7880, 7890, 7920, 7940, 7950, 8050, 8130, 9325,

Step 2. For finding P_{11} , we compute

$$j = \frac{11n}{100} = \frac{12 \times 11}{100} = 1.32$$

Step 3. Here $j = 1.32$ is not an integer. The next integer is 2. The second ordered observation is the P_{11} which is 7755. Hence eleventh percentile, P_{11} is 7755.

For finding P_{50} , we compute the index

$$j = \frac{50n}{100} = \frac{50 \times 12}{100} = 6$$

Here j is an integer. The mean of the 6th and 7th ordered observations is the P_{50} . Here it is

$$P_{50} = \frac{7890 + 7920}{2} = 7905 \text{ Tk. per month.}$$

P_{50} , D_5 and Q_2 are the median which are same.

For finding P_{80} we compute the index

$$j = \frac{80n}{100} = \frac{80 \times 12}{100} = 9.6$$

Here j is not an integer. The next integer greater than 9.6 is 10. The 10th ordered observation is 8050. Hence the 80th percentile is 8050 Tk which is the same as 8th decile.

Steps for computing quartiles, deciles and percentiles from ungrouped data:

Step 1. Arrange the observations in ascending order of magnitude.

Step 2. Compute the index j defined for quartiles, deciles and percentiles.

Step 3. If the index j is an integer for finding any quartile or decile or percentile, the mean of the j th and $(j+1)$ th observations of the ordered array is the required quartile or decile or percentile.

Step 4. If the index j is not an integer, the ordered observation corresponding to the next greater integer than j is the required result.

Example 5.6.4. A bank branch located in a commercial district of a city has developed an improved process for serving customers during the 12:00 P.M. to 1:00 P.M. peak lunch period. The waiting time in minutes of all customers during this hour is recorded over a period of one week. For this purpose a random sample of 20 customers is selected and the results are as follows:

4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.52, 3.20, 4.50, 6.10, 0.38, 5.12, 65.00, 6.19, 3.79, 4.51, 4.21, 2.31, 3.42, 3.45. Compute 1st quartile, 3rd quartile, 7th decile and 65th percentile and comment.

Solution.

Step 1. First, we arrange the observations in ascending order of magnitude. The ordered array is

0.38, 2.31, 2.34, 3.02, 3.20, 3.42, 3.45, 3.52, 3.79, 4.21, 4.21, 4.50, 4.51, 4.77, 5.00, 5.12, 5.13, 5.55, 6.10, 6.19,

Step 2. For finding Q_1 , compute $j = \frac{n}{4}$

Step 3. Here $j = \frac{n}{4} = \frac{20}{4} = 5$

Step 4. Here $j = 5$ is an integer. The mean of the 5th and 6th ordered observations is the first quartile. Hence,

$$Q_1 = \frac{3.20 + 3.42}{2} = 3.31 \text{ minutes.}$$

Comment. $Q_1 = 3.31$ means waiting time for 25% of the customers are less than or equal to 3.31 minutes and 75% of the customers are more than 3.31 minutes. It is to be noted that first quartile is also 25th percentile.

For finding Q_3 , we compute j as $j = \frac{3n}{4}$

$$\text{Here } j = \frac{3n}{4} = \frac{3 \times 20}{4} = 15$$

Here $j = 15$ is an integer. The mean of the 15th and 16th ordered observations is the third quartile. Hence,

$$Q_3 = \frac{5.00 + 5.12}{2} = 5.06 \text{ minutes.}$$

Comment. $Q_3 = 5.06$ means waiting time of 75% of the customers are 5.06 minutes or less and 25% customers are more than 5.06 minutes. Third quartile is also 75th percentile.

Seventh Decile

For finding D_7 , we compute the index j as $j = \frac{7n}{10}$

$$\text{Hence, } j = \frac{7n}{10} = \frac{7 \times 20}{10} = 14$$

Here $j = 14$ is an integer. The mean of the 14th and 15th ordered observations is the seventh decile. Hence,

$$D_7 = \frac{4.77 + 5.00}{2} = 4.87 \text{ minutes.}$$

Comment. $D_7 = 4.87$ means waiting time of 70% of the customers are 4.87 minutes or less and 30% are more than 4.87 minutes. It is the 70th percentile.

65th percentile :

For finding P_{65} , we compute the index j as $j = \frac{65n}{100}$

$$\text{Here, } j = \frac{65n}{100} = \frac{65 \times 20}{100} = 13$$

Here $j = 13$ is an integer. Then the mean of the 13th and 14th ordered observations is the 65th percentile. Hence,

$$P_{65} = \frac{4.51 + 4.77}{2} = 4.64 \text{ minutes.}$$

Comment. $P_{65} = 4.64$ means waiting time of 65% of the customers is 4.64 minutes or less and 35% are more than 4.64 minutes.

Computation of quartiles, deciles and percentiles from grouped data

The formulae for computing quartiles, deciles and percentiles are very much similar with the formula of median. For grouped data, the following formulae are used for finding quartiles, deciles and percentiles;

$$Q_i = L_i + \frac{in / 4 - F_i}{f_i} \times c_i \quad \text{For } i = 1, 2, 3$$

Here Q_i is the i th quartile.

$$D_j = L_j + \frac{jn / 10 - F_j}{f_j} \times c_j \quad \text{For } j = 1, 2, \dots, 9$$

Here D_j is the j th decile

$$P_i = L_i + \frac{kn / 100 - F_i}{f_i} \times c_i \quad \text{For } i = 1, 2, \dots, 9$$

Here P_i is the i th percentile.

The other symbols have their usual meanings and interpretation.

Example 5.6.5. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate Q_1 , Q_3 , D_3 , P_{65} and interpret the values you obtained.

Solution. Table for computation

Class interval	Frequency f	Cumulative frequency F
30 - 55	3	3
55 - 80	4	7
80 - 105	6	13
105 - 130	9	22
130 - 155	12	34
155 - 180	11	45
180 - 205	5	50

First quartile, Q_1 , is the $n/4$ ordered observation = 12.5th ordered observation.
It lies in the class interval 80 - 105. The formula for Q_1 is

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1$$

Here $L_1 = 80$, $n/4 = 12.5$, $F_1 = 7$, $f_1 = 6$ and $c_1 = 25$

$$\begin{aligned} Q_1 &= L_1 + \frac{n/4 - F_1}{f_1} \times c_1 = 80 + \frac{12.5 - 7}{6} \times 25 \\ &= 80 + \frac{5.5 \times 25}{6} = 80 + 22.92 = 102.92 \text{ hours per month} \end{aligned}$$

Comment. 25% of the workers worked for 102.72 hrs or less per month whereas 75% worked for more than 102.72 hrs.

Third quartile, Q_3 is the $3n/4$ th ordered observation = 37.5th ordered observation. It lies in the class 155 - 180. Third quartile is

$$Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3$$

Here $L_3 = 155$, $3n/4 = 37.5$, $F_3 = 34$, $f_3 = 11$, $c_3 = 25$

$$\begin{aligned} Q_3 &= L_3 + \frac{3n/4 - F_3}{f_3} \times c_3 = 155 + \frac{37.5 - 34}{11} \times 25 \\ &= 155 + \frac{3.5 \times 25}{11} = 155 + 7.95 = 162.95 \end{aligned}$$

Comment. 75% of the workers worked 162.95 hrs or less per month, whereas 25% worked more than 162.95 hrs per month.

Third decile, D_3 is the $(3n/10)$ th ordered observation = $\frac{3 \times 50}{10} = 15$ th ordered observation. It lies in the class 105-130. The third decile is

$$D_3 = L_3 + \frac{3n/10 - F_3}{f_3} \times c_3$$

Here $L_3 = 105$, $3n/10 = 50$, $F_3 = 13$, $f_3 = 9$, $c_3 = 25$.

Hence

$$\begin{aligned} D_3 &= L_3 + \frac{3n/10 - F_3}{f_3} \times c_3 = 105 + \frac{3 \times 50/10 - 13}{9} \times 25 \\ &= 105 + \frac{2 \times 25}{9} = 110.56 \text{ hrs.} \end{aligned}$$

Comment. 30% of the workers worked 105.56 hrs or less per month, whereas 70% worked more than 105.56 hours per month.

Sixty fifth percentile, P_{65} is the $(65n/100)$ th ordered observation $= (65 \times 50)/100 = 32.5$ th \approx 33rd observation. 33rd observation lies in the class 130-155. Hence the 65th percentile is

$$P_{65} = L_{65} + \frac{65n/100 - F_{65}}{f_{65}} \times c_{65}$$

Here $L = 130$, $65n/100 = 32.5$, $F_{65} = 22$, $f_{65} = 12$ and $c_{65} = 25$

$$\begin{aligned} P_{65} &= L_{65} + \frac{65n/100 - F_{65}}{f_{65}} \times c_{65} = 130 + \frac{65 \times 50/100 - 22}{12} \times 25 \\ &= 130 + \frac{32.5 - 22}{12} \times 25 = 130 + 21.875 = 151.875 \text{ hours} \end{aligned}$$

Comment. The value of sixty-fifth percentile is 151.875 hours. This means 65% of the workers worked 151.875 hrs or less per month and 35% worked more than 151.875 hrs per month.

Example 5.6.6. Following frequency distribution gives the pattern of overtime work per week by 100 employees of a company. Calculate the median, first quartile, and 7th decile and comment.

Overtime (in hour):	10-15	15-20	20-25	25-30	30-35	35-40
No. of employees:	11	20	35	20	8	6

Solution. Table for calculation of median, Q_1 and D_7 .

Overtime (in hours)	Frequency	Cumulative frequency
10-15	11	11
15-20	20	31
20-25	35	66
25-30	20	86
30-35	8	94
35-40	6	100
	100	

Median = $(n/2)$ th observation = $(100/2)$ th observation = 50th observation
which lies in the class 20-25

The formula for computing median is

$$M_e = L + \frac{n/2 - F}{f} \times c$$

Where M_e = median, L = lower limit of the median class, n = Total number of observations, F = cumulative frequency pre-median class, f = Frequency of the median class, c = Width of the median class.

$$\text{So, } M_e = L + \frac{n/2 - F}{f} \times c = 20 + \frac{50 - 31}{35} \times 5 = 20 + 2.714 = 22.714 \text{ hrs.}$$

First quartile Q_1 :

$Q_1 = (n/4)$ th observation = $(100/4)$ th observation = 25th Observation,

So, Q_1 lies in the class 15-20

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1$$

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1 = 15 + \frac{25 - 11}{20} \times 5 = 15 + 3.5 = 18.5 \text{ hrs.}$$

D_7 is the 7th decile

$$D_7 = (7n/10) \text{th observation} = \left(\frac{7 \times 100}{10}\right) \text{th observation} = 70 \text{th Observation}$$

D_7 lies in the class 25 - 30

$$D_7 = L_7 + \frac{7n/10 - F_7}{f_7} \times c_7$$

$$D_7 = L_7 + \frac{7n/10 - F_7}{f_7} \times c_7 = 25 + \frac{70 - 66}{20} \times 5 = 25 + 1 = 26 \text{ hrs.}$$

Example 5.6.7. The number of days of absenteeism of 80 workers of a factory over a particular year are recorded as follows:

Days	No. of workers	Days	No. of workers
0-3	5	16-19	10
4-7	14	20-23	6
8-11	17	24-27	3
12-15	25		

- Calculate the average days of absenteeism
- Calculate median, mode and Q_3 of days of absenteeism and comment
- If the authority decides to terminate 30% of the most irregular workers, how can you help the authority to take decision in this case?
- On the other hand, if the authority decides to reward 30% of the most regular workers, how can you help the authority to take decision?
- Locate D_3 , median, P_{70} and Q_3 graphically.

Solution. The following table is constructed for necessary calculations:

Class interval	Mid-values (x)	Class boundary	Frequency (f)	fx	Cumulative frequency
0 - 3	1.5	'-0.5 - 3.5	5	7.5	5
4 - 7	5.5	3.5 - 7.5	14	77	19
8 - 11	9.5	7.5 - 11.5	17	161.5	36
12 - 15	13.5	11.5 - 15.5	25	337.5	61
16 - 19	17.5	15.5 - 19.5	10	175	71
20 - 23	21.5	19.5 - 23.5	6	129	77
24 - 27	25.5	23.5 - 27.5	3	76.5	80
Total				964	

(a) We know, average $\bar{x} = \frac{\sum f_i x_i}{n} = \frac{964}{80} = 12.05$ days

(b) We know, Median = $(n/2)$ th observation = $(80/2)$ th observation = 40th observation lies in the class 12 – 15. Hence the median class = 12 – 15.

The formula for computing median is

$$M_e = L + \frac{n/2 - F}{f} \times c$$

Where $L = 11.5$, $n = 80$, $F = 36$, $f = 25$, $c = 4$.

$$\text{So, } M_e = L + \frac{n/2 - F}{f} \times c = 11.5 + \frac{40 - 36}{25} \times 4 = 11.5 + 0.44 = 11.94 \text{ days.}$$

That means, 50% of the workers were absent for 11.94 days or less.

Again, we know, mode is the values which occurs most frequently, here the most frequent class is 12-15, so mode lies in this class interval. The formula for mode is

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c = 130 + \frac{3}{3+1} \times 25$$

Here, $L = 11.5$, $\Delta_1 = 25 - 17 = 8$, $\Delta_2 = 25 - 10 = 15$, $c = 4$,

$$\text{So, mode} = M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c = 11.5 + \frac{8}{8+10} \times 4 = 13.28$$

That means, most of the workers were absent approximately for 13 days.

Again, $Q_3 = (3n/4)\text{th observation} = (3 \times 80/4)\text{th observation}$
 $= 60\text{th Observation,}$

So, Q_3 also lies in the class 12 – 15. We know, the formula for Q_3 is

$$Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3$$

Here, $L_3 = 11.5$, $n = 80$, $F_3 = 36$, $f_3 = 25$ and $c_3 = 4$

$$\text{So, } Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3 = 11.5 + \frac{60 - 36}{25} \times 4 = 15.34 \text{ days.}$$

Which means, 75% of the workers were absent for 15.34 days or less.

(c) According to the question, the authority decides to terminate 30% of the most irregular workers, here, the workers whose days of absenteeism are more, are more irregular. So, for taking decision, at first we have to compute the value of P_{70} or D_7 .

We know, $D_7 = (7n/10)\text{th observation} = (3 \times 80/10)\text{th observation} = 56\text{th Observation,}$

So, D_7 lies in the class 12 – 15.

We know, the formula for D_7 or P_{70} is

$$P_{70} = D_7 = L_7 + \frac{7n/10 - F_7}{f_7} \times c_7 =$$

Here, $L_7 = 11.5$, $F_7 = 36$, $f_7 = 25$, $c_7 = 4$,

$$D_7 = L_7 + \frac{7n/10 - F_7}{f_7} \times c_7 = 11.5 + \frac{56 - 36}{25} \times 4 = 14.70 \text{ days}$$

Thus, to take decision in this case, the authority has to serve a notice that the workers who were absent for more than 14 days in that year will be terminated from company.

(d) According to the question, the authority decides to reward 30% of the most regular workers, here, the workers whose days of absenteeism are fewer, are more regular. So, for taking decision, at first we have to compute the value of P_{30} or D_3 .

We know, $D_3 = (3n/10)$ th observation = $(3 \times 80/10)$ th observation = 24th observation which lies in the class 8 - 11. We know, the formula for which is given by

$$P_{30} = D_3 = L_3 + \frac{3n/10 - F_3}{f_3} \times c_3$$

Here, $L_3 = 7.5$, $F_3 = 19$, $f_3 = 17$, $c_3 = 4$,

$$D_3 = L_3 + \frac{3n/10 - F_3}{f_3} \times c_3 = 7.5 + \frac{24 - 19}{17} \times 4 = 8.67 \text{ days}$$

Thus, to take decision in this case, the authority will just issue a notice that the workers who were absent for less than 9 days will be rewarded by the company.

(e) Like median quartiles, deciles and percentiles are obtained graphically from the ogive curve. First an ogive curve is drawn with the cumulative frequencies. Then position of D_3 , median, D_7 and Q_3 are obtained by drawing lines parallel to the X-axis corresponding to the points $\frac{3n}{10}, \frac{n}{2}, \frac{7n}{10}$ and $\frac{3n}{4}$ on the ogive. Then draw four perpendiculars on the X-axis parallel to the Y-axis. The points $D_3 = 8.70$, median = 12.10, $D_7 = 14.90$ and $Q_3 = 14.45$ on the X-axis give the required values graphically.

The values $P_{30} = D_3$, median, Q_3 , $P_{70} = D_7$, are shown in figure 5.6.1.

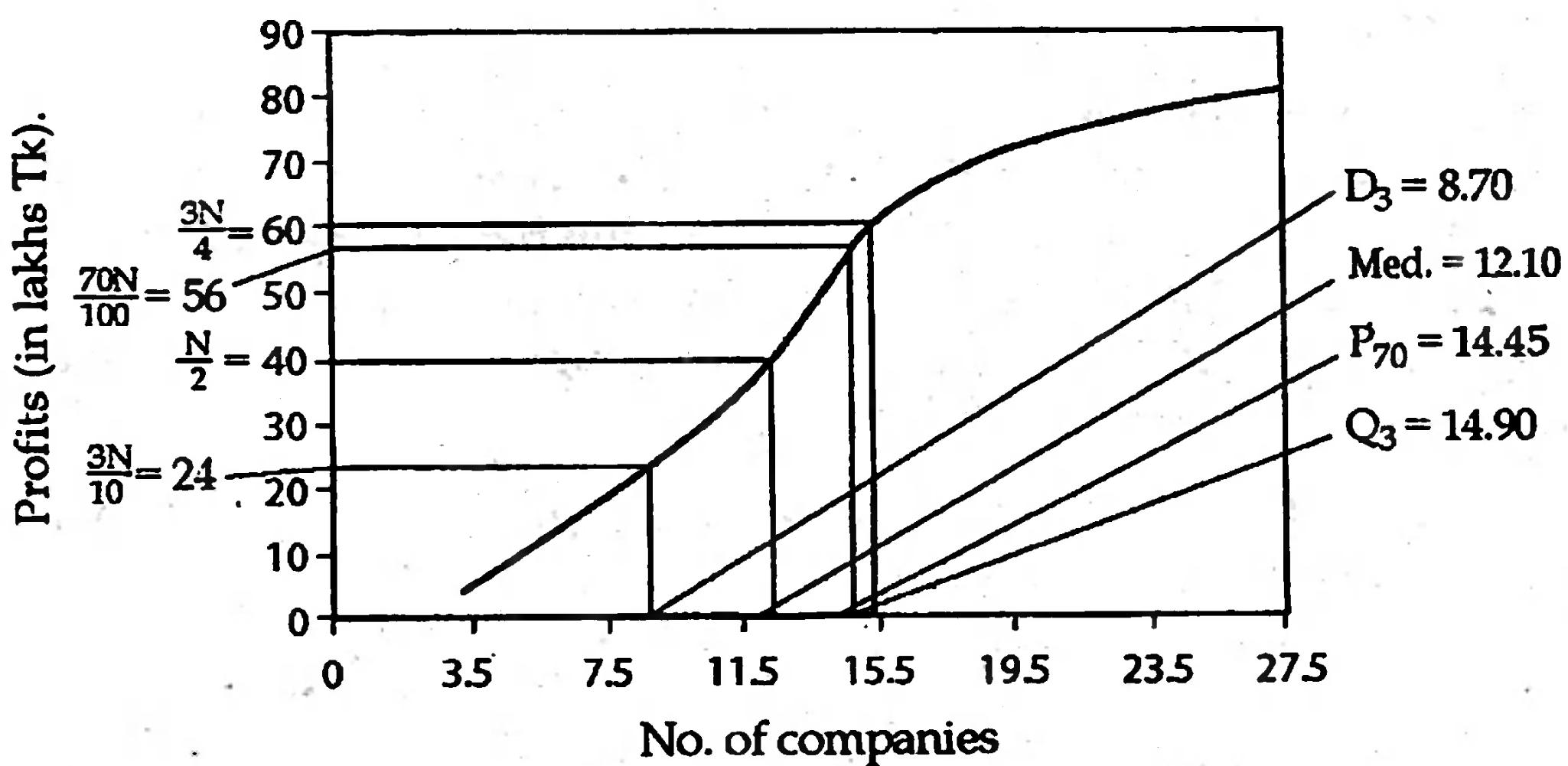


Figure 5.6.1.

A matched problem to solve. The profits earned by 100 companies during 2001-2002 are given below;

Profits (in lakhs Tk.)	No. of companies	Profits (in lakhs Tk.)	No. of companies
20-30	3	60-70	15
30-40	8	70-80	10
40-50	18	80-90	8
50-60	30	90-100	7

Compute Q_1 , median, Q_3 , D_4 and P_{80} and interpret the values.

Quartiles, Deciles and Percentiles Graphically. Like median quartiles, deciles and percentiles are obtained graphically from the ogive curve.

5.7. Geometric Mean

In business and economic problems, very often we face the questions pertaining to percentage, rates of change over time. In that case, neither the mean, the median nor mode is the appropriate measure of location. The geometric mean is a useful measure of the average rate of change of a variable over time.

Definition. Geometric mean denoted by G.M, of n positive and non-zero observations x_1, x_2, \dots, x_n is the n th root of their product and is defined by

$$G.M. = \sqrt[n]{x_1 \times x_2 \times \dots \times x_n}$$

That is, for two positive observations, we take the square root of their product as a geometric mean. If there are three observations, we take the cube root, and so on.

The geometric mean is necessarily zero if any value is zero, and may become imaginary if odd numbers of negative values occur. Mathematically, geometric mean can be calculated if even number of observations carries negative values. But it is meaningless. Otherwise, geometric mean is always determined and is rigidly defined. Following examples will make the mentioned points clear..

Example 5.7.1. Find arithmetic and means geometric means from the following sets of observations;

- (i) 2, 8
- (ii) 2, 4, 8,
- (iii) -4, -16

Solution. (i) The arithmetic mean of 2 and 8

$$\text{A.M.} = \frac{2+8}{2} = \frac{10}{2} = 5.$$

The geometric mean of 2 and 8 is

$$\text{G.M.} = \sqrt{2 \times 8} = \sqrt{16} = 4.$$

(ii) The arithmetic mean of 2, 4 and 8

$$\text{A.M.} = \frac{2+4+8}{3} = \frac{14}{3} = 4.67.$$

$$\text{G.M.} = \sqrt[3]{2 \times 4 \times 8} = \sqrt[3]{64} = 4.$$

Comment. Here it is seen that for both the cases A.M is greater than G.M. These results are true for any non-zero, positive and unequal set of observations.

(iii) The arithmetic mean of -4 and -16 is

$$\text{A.M.} = \frac{(-4) + (-16)}{2} = \frac{-20}{2} = -10.$$

$$\text{G.M.} = \sqrt{(-4) \times (-16)} = \sqrt{64} = 8.$$

Here it is seen that arithmetic mean gives a reasonable measure of central tendency. But geometric mean gives a positive value 8 which is not a reasonable value of a measure of central tendency. Moreover, it is greater than arithmetic mean. We can explain the result as you have borrowed 4 and 16 taka from your two friends. That is you have borrowed Tk.10 on average. But geometric mean tells that you are getting on average Tk. 8 from your friends.

When the number of observations is three or more the task of multiplying the number and of extracting the root becomes quite difficult. To simplify calculation, logarithms are used. Geometric mean is then computed as follows:

$$\log \text{G.M.} = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{n} = \frac{\sum \log x}{n}$$

That means, the logarithm of geometric mean of a set of non-zero positive observations is the arithmetic mean of logarithm of observations.

Example 5.7.2. Find arithmetic mean and geometric mean from the following set of observations

$$2, 4, 8, 12, 16, 24$$

Solution. Table for finding arithmetic mean and geometric mean.

Values of x	Log x
2	0.3010
4	0.6021
8	0.9031
12	1.0792
16	1.2041
24	1.3802
$\Sigma x = 66$	$\Sigma \log x = 5.4697$

Here $n = 6$.

$$A.M. = \bar{x} = \frac{\Sigma x}{6} = 11.$$

$$G.M. = \text{Antilog}\left(\frac{\Sigma \log x}{n}\right) = \text{Antilog}\left(\frac{5.4697}{6}\right) = \text{Antilog}(0.9116) = 8.159.$$

Comment. Again it is proved that A.M. is greater than G.M.

Like geometric mean there is another summary measure called **geometric mean rate of return**. It is computed by the following formula:

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_n)]^{\frac{1}{n}} - 1$$

where R_i is the rate of return in time period i .

Now we shall cite an example to show the appropriate use of geometric mean.

The geometric mean is used primarily to average the data for which the ratio of consecutive terms remains approximately constant. This occurs, for example, with such data as rates of change, ratios, economic index numbers, and population sizes over consecutive time periods, etc.

Example 5.7.3. The populations of a country for seven censuses are as follows.

Census date	1890	1900	1910	1920	1930	1940	1950
Population: P	205,876	285,704	465,766	993,678	1,568,662	1,623,452	1,849,568

Compute the average percentage increase of population of the census.

Solution. First we compute the percentage of population of a year based on previous census data. For example, percentage of population for the year 1900 over 1890 is

$$\frac{P_{1900}}{P_{1890}} \times 100 = \frac{285,704}{205,876} \times 100 = 138.8$$

Percentages of population for other years are computed in a similar way and presented in the following table:

Census data	Population	Percentage of population on previous census data : x	$\log x$
1890	205,876		
1900	285,704	138.8	2.14239
1910	465,766	163.0	2.21219
1920	993,678	213.3	2.32899
1930	1,568,662	157.9	2.19838,
1940	1,623,452	103.5	2.01494
1950	1,849,568	113.9	2.05652
			12.95341

$$\log G.M. = \frac{\sum \log x}{n} = \frac{12.95341}{6} = 2.15890$$

$$G.M. = \text{antilog}(2.15890) = 144.2$$

$$\text{Average rate of increase} = 144.2 - 100 = 44.2 \text{ percent.}$$

The population increased from 205,876 in 1890 to 1,849,568 in 1950. In other words, the 1950 population was 8.9839 times of the 1890 population. This means the population has increased by 8.9839 times over the last six decades. If we want to know the average increase per decade we may find it directly by taking the sixth root of 8.9839. The average increase per decade is the geometric mean of the different rates over the period. That is

$$G.M. = \sqrt[6]{8.983}$$

$$\log G.M. = \frac{\log 8.9839}{6} = \frac{0.95347}{6} = 0.15891$$

$$G.M. = \text{antilog } 0.15891 = 1.442$$

$$\text{Hence, the average rate of increase} = 1.442 - 1.00 = 0.442$$

This is the same as that we got before.

But it is easily seen that the arithmetic mean of the six percentages is

$$A.M. = \frac{138.8 + 163.0 + 213.3 + 157.9 + 103.5 + 113.9}{6} = \frac{890.4}{6} = 148.4$$

$$\text{Hence the arithmetic mean of the rate of increase} = 148.4 - 100.0 = 48.4.$$

So, the appropriate average in this case is the geometric mean.

Example 5.7.4. Consider an investment of Tk.100,000 that declined to a value of Tk. 50,000 at the end of one year and then bounded back to its original Tk.100,000 value at the end of year two.

It is clear that the rate of this investment for the two-year period is 0, because the starting and ending value of the investment is unchanged. However, the arithmetic mean of the yearly rates of return is

$$\bar{X} = \frac{(-0.5) + (1.00)}{2} = 0.25 \text{ or } 25\%$$

Since the rate of return for year 1 is

$$R_1 = \left(\frac{50,000 - 100,000}{100,000} \right) = 0.50 \text{ or } 50\%$$

and the rate of return for year 2 is

$$R_2 = \left(\frac{100,000 - 50,000}{50,000} \right) = 1.00 \text{ or } 100\%$$

However, the geometric rate of return for the two years is

$$\begin{aligned} \bar{R}_G &= [(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_n)]^{\frac{1}{n}} - 1 \\ &= [(1 + (-0.50)) \times (1 + (1.00))]^{\frac{1}{2}} - 1 \\ &= [(0.50) \times (2.0)]^{\frac{1}{2}} - 1 = 1 - 1 = 0 \end{aligned}$$

Thus, the geometric mean more accurately reflects the change in the value of the investment for the two year period than does the arithmetic mean.

Application of Geometric Mean. Geometric mean is especially useful in the following cases:

1. It is used to find the average percent increase in sales, production, population or other economic or business data.
2. It is theoretically considered to be the best average in the construction of index number.
3. It is an average, which is most suitable when large weights have to be given to small values of observations and small weights to large values of observations, situation which we usually come across in social and economic fields.

5.8. Harmonic Mean

It is rarely used to analysis business and economic data. It is useful for computing the rate of increase of profits or average speed at which a journey has been performed or the average price in which an article has been sold.

Definition. Harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observations. It is computed by the formula

$$H.M. = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum \left(\frac{1}{x} \right)} ; \text{ for ungrouped data.}$$

$$\text{For grouped data, } H.M. = \frac{n}{\sum \left(f \times \frac{1}{x} \right)}$$

$$\text{Or, } \frac{1}{H.M.} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n} = \frac{\sum \left(\frac{1}{x} \right)}{n}$$

That means, the reciprocal of harmonic mean is the arithmetic mean of the reciprocal of observations.

In actual practice, the harmonic mean is most frequently used in averaging speeds for various distances covered where the distances remain constant, and also in finding the average cost of some commodity, such as mutual funds, when several different purchases are made by investing the same amount of money each time.

Example 5.8.1. A toy factory has assigned a group of 4 workers to complete an order of 1,400 toys of a certain type. The productive rates of the four workers are given below:

Workers	Productive rates
A	4 minutes per toy
B	6 minutes per toy
C	10 minutes per toy
D	15 minutes per toy

Find the average productive rate of workers per toy if the same amount of times are assigned to each worker.

Solution. Here harmonic mean is the appropriate average.

$$H.M. = \frac{4}{\frac{1}{4} + \frac{1}{6} + \frac{1}{10} + \frac{1}{15}} = \frac{4 \times 60}{35} = \frac{48}{7} = 6.\overline{6} \text{ minutes per toy}$$

$$\text{Total time required to complete 1,400 toys} = \frac{1400 \times 48}{7} = 9,600 \text{ minutes}$$

$$\text{Here, the arithmetic mean is, } \bar{x} = \frac{4+6+10+15}{4} = 8\frac{3}{4} \text{ minutes per toy}$$

Verification.

Each worker works for $\frac{9600}{4} = 2400$ minutes

Toys produced by A in 2400 minutes $= 2400/4 = 600$

Toys produced by B in 2400 minutes $= 2400/6 = 400$

Toys produced by C in 2400 minutes $= 2400/10 = 240$

Toys produced by D in 2400 minutes $= 2400/15 = 160$

Total = 1400.

According to the arithmetic Mean, total time required to complete 1400 toys is $= 1400 \times 35/4 = 12250$ minutes which is higher than the time obtained by harmonic mean.

A matched problem

A college professor invest Taka 100 a month in a mutual growth fund. For the last 4 months the prices per share in the fund have been Tk. 5.45, Tk. 5.76, Tk. 6.10 and Tk. 5.90. Calculate the average price per share paid by the professor over this period of time.

Ans. Tk. 5.79

Remarks. For a set of non-zero positive observations

$$A.M. \geq G.M. \geq H.M.$$

They are equal when all the observations are non-zero, positive and equal.

Example 5.8.2. Find A.M., G.M., and H.M. from the following set of observation

$$5, 5, 5$$

$$A.M. = \frac{5+5+5}{3} = \frac{15}{3} = 5$$

$$G.M. = \sqrt[3]{5 \times 5 \times 5} = \sqrt[3]{125} = 5$$

$$H.M. = \frac{3}{\frac{1}{5} + \frac{1}{5} + \frac{1}{5}} = \frac{3}{\frac{3}{5}} = 5$$

Hence A.M., G.M. and H.M. are all equal when all the observations are equal.

General comment. Among all the measures of location arithmetic mean is the best since it satisfies most of the criteria of a good measure of central tendency. But it is highly affected by extreme values. Mode or median is the best measure of location in presence of extreme values. But mode and median may not be rigidly defined. Mode is the only measure of location for categorical data measured in nominal scale.

5.9. Some more Measures of Central Tendency

Midrange. The midrange of a data set is defined as the mean of the largest and smallest values.

Example 5.9.1. The number of building permits issued last month to 12 construction firms in a small city were 4, 7, 0, 7, 11, 4, 1, 15, 3, 5, 8 and 7. Find the midrange.

Solution. The largest value of the dataset is 15 and the smallest value is 0.

Hence the midrange is $\frac{15+0}{2} = 7.5$.

5.10. Some Simple Theorems and Problems on Measures of Location

Theorem 5.10.1. For two positive non zero quantities

$$A.M. \geq G.M. \geq H.M.$$

Here A.M. = Arithmetic mean, G.M. = Geometric mean and H.M. = Harmonic mean.

Proof. Suppose x_1 and x_2 are two positive and non-zero quantities. Then

$$A.M. = \frac{x_1 + x_2}{2}, \quad G.M. = \sqrt{x_1 \times x_2} \quad \text{and} \quad H.M. = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}.$$

Here $(\sqrt{x_1} - \sqrt{x_2})^2 \geq 0$ since x_1 and x_2 are positive.

$$\text{or, } x_1 + x_2 - 2\sqrt{x_1 x_2} \geq 0$$

$$\text{or, } x_1 + x_2 \geq 2\sqrt{x_1 x_2}$$

$$\text{or, } \frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$$

Hence $A.M. \geq G.M.$ (i)

Again, $\left(\frac{1}{\sqrt{x_1}} - \frac{1}{\sqrt{x_2}} \right)^2 \geq 0$

$$\text{or, } \frac{1}{x_1} + \frac{1}{x_2} - 2\frac{1}{\sqrt{x_1 x_2}} \geq 0$$

$$\text{or, } \frac{1}{x_1} + \frac{1}{x_2} \geq 2\frac{1}{\sqrt{x_1 x_2}}$$

$$\text{or, } \sqrt{x_1 \times x_2} \geq \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$$

Hence, G.M. \geq H.M. (ii)

From (i) and (ii), we have A.M. \geq G.M. \geq H.M.

Theorem 5.10.2. For two non-zero and positive quantities

$$\text{G.M.} = \sqrt{\text{A.M.} \times \text{H.M.}}$$

Proof. Suppose x_1 and x_2 are two positive and non-zero quantities. Then

$$\text{A.M.} = \frac{x_1 + x_2}{2} \quad \text{G.M.} = \sqrt{x_1 \times x_2} \quad \text{and H.M.} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$$

$$\begin{aligned} \text{Now, } \text{A.M.} \times \text{H.M.} &= \frac{x_1 + x_2}{2} \times \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} = \frac{x_1 + x_2}{2} \times \frac{2}{\frac{x_1 + x_2}{x_1 \times x_2}} \\ &= \frac{x_1 + x_2}{2} \times \frac{2(x_1 \times x_2)}{x_1 + x_2} = x_1 \times x_2 = (\sqrt{x_1 \times x_2})^2 = (\text{G.M.})^2 \end{aligned}$$

$$\text{Hence, G.M.} = \sqrt{\text{A.M.} \times \text{H.M.}}$$

Example 5.10.1. Suppose the arithmetic mean and harmonic of two positive quantities are 8 and 2. Find their geometric mean.

Solution. Here A.M. = 8 and H.M. = 2.

We know, A.M. \times H.M. = (G.M.)²

$$8 \times 2 = (\text{G.M.})^2$$

$$\text{Hence, G.M.} = \sqrt{16} = 4$$

Example 5.10.2. Suppose the geometric mean and harmonic of two positive quantities are $4\sqrt{3}$ and 6. Find their arithmetic mean.

Solution. Here, G.M. = $4\sqrt{3}$ and H.M. = 6.

We know, A.M. \times H.M. = (G.M.)²

$$\text{A.M.} \times 6 = (4\sqrt{3})^2 = 48$$

$$\text{Hence, A.M.} = 48/6 = 8.$$

Theorem 5.10.3. For n natural numbers, arithmetic mean is equal to $(n+1)/2$.

Proof. The first n natural numbers are $1, 2, 3, \dots, n$.

$$\text{Then A.M.} = \frac{1+2+3+\dots+n}{n} = \frac{n(n+1)/2}{n} = \frac{n+1}{2}$$

Example 5.10.3. Find the arithmetic mean of $1, 2, 3, \dots, 20$.

$$\text{Here } n = 20. \text{ Hence A.M.} = \frac{20+1}{2} = 10.5.$$

Theorem 5.10.3. For n natural numbers, median is equal to $(n+1)/2$.

Proof. Let $1, 2, 3, \dots, n$ be n natural numbers. n may be odd or even number. First let us take n is odd number. Suppose $n = 2m+1$, then median will be $(m+1)$ th value. In this case $m = (n-1)/2$.

By definition, Median = $\{(n-1)/2+1\}$ th value = $(n+1)/2$

Now, let us take n is even. Then $n=2m$. In this case $m = n/2$.

By definition, median will be mean of m th value and $(m+1)$ th value.

$$\text{Then, Median} = \frac{\frac{n}{2} + \frac{n}{2} + 1}{2} = \frac{n+1}{2}$$

It is seen that in both the cases median is $\frac{n+1}{2}$.

Remarks : For first n natural numbers mean and median are same.

Example 5.10.4. Find median for the two sets of data

- (i) $1, 2, 3, \dots, 50$
- (ii) $1, 2, \dots, 51$

Solution. (i) Here $n = 50$ is even number.

Then median is $(n+1)/2 = (50+1)/2 = 25.5$.

(ii) Here $n = 51$ is odd number. The median is $(n+1)/2 = (51+1)/2 = 26$.

Theorem. Arithmetic mean depends on the shift of origin and change of scale.

Proof. Let x_1, x_2, \dots, x_n be n values of a variable x .

$$\text{By definition, } \bar{x} = \frac{\Sigma x}{n}$$

Let, $u = \frac{x-A}{k}$, this means we have shift origin of x to A and change by scale k .

In this case, $x = ku + A$

Then, $x_1 = ku_1 + A, x_2 = ku_2 + A, \dots, x_n = ku_n + A$

$$x_1 + x_2 + \dots + x_n = ku_1 + ku_2 + \dots + ku_n$$

$$\frac{\sum x}{n} = \frac{k\sum u}{n} + \frac{nA}{n}$$

Hence, $\bar{x} = k\bar{u} + A$

Here it is seen that arithmetic mean \bar{x} depends on both k and A . This means arithmetic mean depends on the shift of origin and change of scale.

Remarks : This formula is used for finding arithmetic mean by short cut method.

Exercise

1. What do you mean by central tendency? What are the important measures of central tendency? Find arithmetic mean, median and mode from the following data set: 2, 5, 7, 3, 4, 3, 3
2. Define arithmetic mean and median. Why arithmetic mean is the best measure of central tendency?
3. What are the desirable properties of a good measure of central tendency? Which one is the best? Why?
4. When arithmetic mean is not a good measure of location? When mode or median is better than arithmetic mean as a measure of location?
5. What are quartiles of a distribution? Define first and third quartiles. What do they mean?
6. Define mode. Is it unique for distribution? When it is an appropriate measure of location?
7. Find mean, median and mode from the following sets of data:
 - (i) 4, 5, 5, 12, 8, 6, 7, 6, 5 Ans. $\bar{x} = 6.44, M_e = 6, M_o = 5$
 - (ii) 23, 1, 3, 2, 3, 7, 6, 7, 7, 5, 3 Ans. $\bar{x} = 6.09, M_e = 5, M_o = 3.7$
 - (iii) 16, 1, 3, 15, 11, 5, 6, 7, Ans. $\bar{x} = 8.11, M_e = 7, M_o = \text{None } 9$

8. Find appropriate measure of location for the following data:

Favour	Vanilla	Chocolate	Strawberry	Cherry	Almond mocha
Number Preferring	149	370	90	60	220

Ans. Chocolate is the modal preference

9. Compute first quartile, third quartile, 4th decile and 36th percentile from the following 22, 18, 20, 23, 17, 22, 17, 19, 20, 22.

Ans. $Q_1 = 18$, $Q_3 = 22$, $D_4 = 19.5$, $P_{36} = 19$

10. Find mean from the following set of grouped data:

Class interval	Frequency
0.5 – 2.5	2
2.5 – 4.5	5
4.5 – 6.5	7
6.5 – 8.5	1

Ans. Mean = 4.4

11. Find an appropriate measure of location from the following data set:

18.6 17.9 18.7 16.9 17.8 18.4 17.7, 18.1 16.2 81.3 18.8

Ans. The median is the appropriate measure of location and it is 18.1, since there exist an extreme value in the data set.

Application

12. Following frequency distribution gives the pattern of overtime work per week done by 100 employees of a company. Calculate the median, third quartile, 8th decile, 55th percentile and comment.

Overtime (in hour)	10-15	15-20	20-25	25-30	30-35	35-40
No. of employees	12	22	33	20	8	5

13. Find the geometric mean of 1, 4, and 128.

Ans. 8.

14. Over a period of 4 consecutive years an employee has received 7.2, 8.6, 6.9 and 9.8 percent annual pay increases. The ratios, there of each new salary to the previous year's salary are 1.072, 1.086, 1.069, 1.098. Using logarithms, find the geometric mean for these four ratios and then determine the average percent increase for this employee over the 4-year period.

Ans. 8.119%

15. On a vacation trip to Chittagong, a family travels 500 kilometers each day. If the trip lasts 3 days and the family travels at the rate of 80 kilometers per hour the first day, 93 kilometers per hour the second day, and 87 kilometers per hour the third day, find the average speed for the entire trip.

Ans. 86.3 kilometers per hour.

16. The following table gives the distribution of number of defects found in a lot of 100 pieces of certain kind of garments products of a company:

Number of defects	Number of products	Number of defects	Number of products
4-5	11	10-11	32
6-7	17	12-13	15
8-9	20	14-15	5

Calculate and interpret each of the following characteristics of defects

- a. Average numbers of defects by indirect method.
 - b. Median, Q_3 and Mode of defects.
 - c. The management of the company wants to reject 35% garments with maximum defects and repair others for resale. State how you can help the management in taking decision in this regard.
 - d. If the average number of defects found in additional 125 products is 11, calculate the average number of defects for all the products together.
17. The profits (in thousand) in a particular month earned by a number of salesmen of a large company are given below.

Class Intervals	Frequency
0-9	10
10-19	20
20-29	35
30-39	40
40-49	25
50-59	25
60-69	15

- a. Calculate the average score of the players using A.M.
 - b. Calculate median and modal score and comment.
 - c. If the team manager wants to increase 10% payment of each of the top 15% scorers, determine the lowest score of the players who will receive this benefit.
18. The following table gives the distribution of hourly wages (in Taka) of workers in a certain commercial organization:

Hourly wages	No. of workers	Hourly wages	No. of workers
5-9	2	25-29	62
10-14	9	30-34	39
15-19	25	35-39	20
20-24	30	40-44	3

- a. Calculate average wage of employees
- b. Calculate Median wage and comment
- c. Suppose the organization intends to hike the wages of lower 35% workers by 25%, Calculate the maximum wage of the workers who will belong to this scheme.

- d. On the other hand, if the company plans to hike the wage of upper 12% workers by 20%, calculate the minimum wage of the workers who will be under this plan.
19. The following table gives the distribution of premiums (in thousand taka) earned by 100 agents of certain a company in a particular quarter :

Premium	No. of Agents	Premium	No. of Agents
6 - 8	11	15 - 17	32
9 - 11	17	18 - 20	15
12 - 14	20	21 - 23	5

- a. Calculate average premium by indirect method
 b. Calculate Median, Q_3 , P_{35} , D_8 and Mode of premium and comment
 c. If the company wants to reward 45% of the agents who earned maximum premium, calculate the minimum amount of premium which would be rewarded
 d. If the average premium of 125 agents of another company is Taka 25 thousand, calculate the average premium of agents of both companies.
20. The frequency distribution given below refers to the blood cholesterol levels (given in milligrams per deciliter) of 65 patients admitted into a clinic:
- Frequency distribution of Blood Cholesterol Levels of 65 persons

Blood cholesterol Level	Number of patients
149.5 - 169.5	4
169.5 - 189.5	11
189.5 - 209.5	15
209.5 - 229.5	25
229.5 - 249.5	13
249.5 - 269.5	7
269.5 - 289.5	3
289.5 - 309.5	2

Compute mean, median, mode, Q_1 , Q_3 , D_4 , P_{65} of the frequency distribution and comment.

21. Suppose the daily profits in taka of 100 shops are distributed as follows:

Profit in taka	Number of shops
0-1000	12
1000-2000	18
2000-3000	27
3000-4000	20
4000-5000	17
5000-6000	6

- (i) Compute mode and median.
 (ii) Locate mode and median graphically.
 (iii) Compute Q_1 and Q_3 .
 (iv) Locate Q_1 and Q_3 graphically.

CHAPTER - 6

MEASURES OF DISPERSION

6.1. Introduction

Dispersion is the second important characteristic of a frequency distribution. Two distributions may have the same mean, median and mode, but the variation of the individual observations of the two distributions may differ. For example, the daily wages in Taka of seven workers of two factories are as follows:

Wages of factory : I	142	143	150	150	153	155	157
Wages of factory : II	122	140	150	150	154	159	175

Here it is seen that the mean, median and modal wages of both the factories are same. It is Tk. 150. That is both the distributions have the same measures of location. But the individual observations are not same. The range of the wage structure of the first factory is Tk. 15 = 157 - 142 only whereas the range of the second factory is Tk. 53 = 175 - 122. That is the wage structure of the factory I is more compact than the wage structure of the factory II. The variability or dispersion of the individual values of factory 1 is less than the factory 2. Suppose the depths of a river at 5 different points are 4, 3, 2, 3, 8 feet. The average depth of a river at different points is 4ft; it would not give you the guarantee to cross the river safely. To cross the river safely, you have to know the maximum depth of the river. This means measures of location have failed to measure the variability of a set of data. The need for a measure of dispersion in addition to a measure of location is necessary. Small dispersion indicates high uniformity of the observations in the distribution.

Dispersion tells us how compactly the individual values are distributed around the central values. Dispersion of a single variable might not bear that much meaning, while comparison of dispersion of two sets of variables is more useful for taking decision.

According to Books and Dicks "Dispersion or spread is the degree of the scatter or variation of the variables about a central value."

Definition.6.1. Dispersion measures the variability of a set of observations among themselves or about some central values.

6.2. Purposes of Dispersion

Measure of dispersion is needed for four basic purposes :

- i) To determine the reliability of an average.
- ii) To serve as a basis for the control of the variability.
- iii), To compare two or more series with regard to their variability.
- iv) To facilitate the computation of other statistical measures.

6.3. Properties of a Good Measures of Dispersion

The properties of a good measure of dispersion are the same as those of a good measure of central tendency. They are as follows:

- i) It should be simple to understand,
- ii) It should be easy to compute,
- iii) It should be rigidly defined,
- iv) It should be based on all the observations,
- v) It should have sampling stability,
- vi) It should be suitable for further algebraic treatment, and
- vii) It should not be affected by extreme observations.

6.4. Measures of Dispersion

The numerical values by which we measure the dispersion or variability of a data set or a frequency distribution are called measures of dispersion.

There are two kinds of measures of dispersion:

- i) Absolute measures of dispersion and
- ii) Relative measures of dispersion.

Important absolute measures of dispersion are:

- i) Range,
- ii) Quartile deviation,
- iii) Mean deviation,
- iv) Variance and Standard deviation.

The relative measures of dispersion are:

- i) Coefficient of range,
- ii) Coefficient of quartile deviation,
- iii) Coefficient of mean deviation,
- iv) Coefficient of variation.

It is seen that for each absolute measure of dispersion, there is a relative measure of dispersion. Absolute measures are expressed in the same unit in which the original variables are measured. For example, any absolute measure of weight may be measured as kilogram or pound, height as meter or inch, price as taka or dollar etc. On the other hand, relative measures are pure number and independent of the unit of measurement and express in percentage.

Hence relative measures are better than absolute measures to compare the variability of two sets of observations or distributions measured in different units. Now we shall discuss them.

6.5. Range

Range is the difference between the largest and smallest observations in a set of data. Symbolically,

$$\text{Range} = X_L - X_S$$

Here X_L = largest observation and X_S = smallest observation.

6.5.1. Coefficient of range. The relative measure corresponding to range, called the coefficient of range is computed by the following formula:

$$\text{Coefficient of range} = \frac{X_L - X_S}{X_L + X_S} \times 100$$

Example 6.5.1. The monthly incomes in taka of seven employees of a firm are 5,500, 5,750, 6,500, 6,750, 7,000, 7,500 and 8,500. Compute range and coefficient of range.

Solution. The range of the incomes of the employees is

$$\text{Range} = 8500 - 5500 = \text{Tk. } 3000$$

$$\begin{aligned}\text{Coefficient of range} &= \frac{X_L - X_S}{X_L + X_S} \times 100 \\ &= \frac{8500 - 5500}{8500 + 5500} \times 100 = \frac{3000}{14000} \times 100 = 21.43\%\end{aligned}$$

In a frequency distribution, range is calculated by taking the difference between the lower limit of the lowest class and the upper limit of the highest class.

Example 6.5.2. Compute range and coefficient of range from the following frequency distribution:

Profits (Tk. lakhs)	10-20	20-30	30-40	40-50	50-60	60-70
No. of companies	8	18	30	15	10	7

Solution. Range = $X_L - X_S$.

Here $X_L = 70$ and $X_S = 10$. Range = $70 - 10 = 60$

$$\text{Coefficient of range} = \frac{X_L - X_S}{X_L + X_S} \times 100 = \frac{70 - 10}{70 + 10} \times 100 = \frac{60}{80} \times 100 = 75\%.$$

6.5.2. Merits and limitations. The following are the merits and limitations of range:

Merits

- i) The range measures the total spread in the set of data.
- ii) It is rigidly defined.
- iii) It is the simplest measure of dispersion and easiest to compute.
- iv) It takes minimum time to compute.
- v) It is based on only maximum and minimum values of data set.

Limitations

- i) It is not based on all the values of a set of data.
- ii) It is affected by sampling fluctuation.
- iii) It cannot be computed in case of open-end distribution.
- iv) It is highly affected by extreme values.

6.5.3. Uses of Range. Although range measures the total spread of the set of observations, its uses are prominent in following fields:

- i) **Quality Control.** It is widely used in production process to control the quality of the products.
- ii) **Share Market.** It is widely used to study the variations in the prices of stocks and shares and other commodities.
- iii) **Weather forecasts.** The meteorological department uses the range to determine the difference between the minimum and maximum temperature, which is a very useful index for people to know the limits of temperature in a particular season. Also maximum and minimum values of other climatic factors such as rainfall, humidity, wind velocity etc. are very important from the metrological points of view.

6.6. Inter - Quartile Range and Quartile Deviation

Another measure similar to range is the inter-quartile range.

Definition. The range which includes the middle 50% of the observations is called inter-quartile range. In other words, it is the difference between the third quartile and the first quartile. Symbolically, if Q_3 and Q_1 are third and first quartiles of a data set respectively then inter-quartile range is

$$\text{Inter - quartile Range (IQR)} = Q_3 - Q_1.$$

6.6.1. Quartile deviation. Half of the inter-quartile range is called quartile deviation. It is also sometimes known as semi-inter-quartile range (SIQR). Symbolically, if Q_3 and Q_1 are third and first quartiles of a data set, quartile deviation, denoted by Q.D. is given by

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

Quartile deviation is better than range as it measures the variation of the middle 50% of the observations. Small quartile deviation means high uniformity or small variation of the central 50% observations, whereas a large quartile deviation means large variation among the central observations.

6.6.2. Coefficient of quartile deviation. Quartile deviation is an absolute measure of variation. The relative measure corresponding to this measure, called the coefficient of quartile deviation, is defined by

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

It is used to compare the variation of two distributions measured in different units of measurements.

Example 6.6.1. The monthly salary in taka of seven employees of a firm are 5,500, 6,750, 750, 7500, 7,000, 6,500 and 8,500. Calculate the quartile deviation and coefficient of quartile deviation.

Solution. To find the first and third quartiles, we first arrange the observations in ascending order of magnitude. The ordered observations are 5500, 5750, 6500, 6750, 7000, 7500, 8500.

For finding Q_1 , we compute $j = n/4$ which is $7/4 = 1.75$. Hence second term of the ordered array is the first quartile. That is $Q_1 = 5750$.

For finding third quartile, we compute $j = 3n/4$ which is $(3 \times 7)/4 = 5.25$. Hence sixth term of the ordered array is the third quartile. That is $Q_3 = 7500$.

$$\text{Hence, Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{7500 - 5750}{2} = \frac{1750}{2} = \text{Tk. } 875.00$$

$$\begin{aligned} \text{Coefficient of quartile deviation} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 = \frac{7500 - 5750}{7500 + 5750} \times 100 \\ &= \frac{1750}{13250} \times 100 = 13.21\%. \end{aligned}$$

Example 6.6.2. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate quartile deviation and coefficient of quartile deviation.

Solution. Quartile deviation and coefficient of quartile deviations are computed by the formulas:

$$Q.D. = \frac{Q_3 - Q_1}{2} \text{ and Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

First quartile, Q_1 , is the $n/4$ ordered observation = 12.5th ordered observation. It lies in the class interval 80 - 105.

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1$$

Here $L = 80$, $n/4 = 12.5$, $F_1 = 7$, $f_1 = 6$ and $c_1 = 25$

$$\begin{aligned} Q_1 &= L_1 + \frac{n/4 - F_1}{f_1} \times c_1 = 80 + \frac{12.5 - 7}{6} \times 25 \\ &= 80 + \frac{5.5 \times 25}{6} = 80 + 22.92 = 102.92 \text{ hours per month.} \end{aligned}$$

Third quartile, Q_3 is the $3n/4$ th ordered observation = 37.5th ordered observation. It lies in the class 155 - 180. Third quartile is

$$Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3$$

Here $L = 155$, $3n/4 = 37.5$, $F_3 = 34$, $f_3 = 11$, $c_3 = 25$

$$\begin{aligned} Q_3 &= L_3 + \frac{3n/4 - F_3}{f_3} \times c_3 = 155 + \frac{37.5 - 34}{11} \times 25 = 155 + \frac{3.5 \times 25}{11} \\ &= 155 + 7.95 = 162.95 \text{ hours per month.} \end{aligned}$$

$$\text{Hence Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{162.95 - 102.92}{2} = \frac{60.03}{2} = 30.015 \text{ hours per month.}$$

$$\begin{aligned} \text{Coefficient of Q.D.} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 = \frac{162.95 - 102.92}{162.95 + 102.92} \times 100 \\ &= \frac{60.03}{265.87} \times 100 = 22.58\%. \end{aligned}$$

6.6.3. Merits and demerits of quartile deviation. The range of a set of variable possesses following merits

- i) It is superior to range as a measure of variation.
- ii) It is useful in case of open-end distribution.
- iii) It is not affected by the presence of extreme values.
- iv) It is useful in case of highly skewed distribution.

While the QD is liable for following limitations:

- i) It is not based on all the observations.
- ii) Quartile deviation ignores the first 25% and the last 25% observations.
- iii) It is not capable of mathematical manipulation.
- iv) Its value is very much affected by sampling fluctuations.
- v) It is not a good measure of dispersion since it depends on two positional measures.

6.7. Mean Deviation

Range and quartile deviation do not measure the scatterness of the observations about any central values. However, to study the formation of a distribution we should take the deviation of the observations from an average. Mean deviation is based on such kind of deviation.

Mean Deviation is obtained by calculating the absolute deviations of each observation from mean or median or mode and then averaging these deviations by taking their arithmetic mean. So we can define mean deviation in three ways, namely,

- i) Mean deviation about mean,
- ii) Mean deviation about median and
- iii) Mean deviation about mode.

6.7.1. Mean deviation for ungrouped data.

Mean Deviation. Suppose x_1, x_2, \dots, x_n are n values of a variable, and \bar{x} is the mean then mean deviation (M.D.) about mean is defined by

$$M.D. (\bar{x}) = \frac{\sum |x - \bar{x}|}{n}$$

Similarly, mean deviation about median and mean deviation about mode are defined by

$$M.D. (M_e) = \frac{\sum |x - M_e|}{n} ; \quad M.D. (M_o) = \frac{\sum |x - M_o|}{n}$$

Where M_e and M_o are median and mode of the variable respectively.

The reason for taking absolute deviations is that it would be useless to take the sum of the deviations of the observations from their arithmetic mean as a measure of dispersion since their algebraic sum is zero.

It is to be noted that mean deviation about median is the least. In actual practice the arithmetic mean is more popularly used in calculating the value of mean deviation because of its wide usage as a measure of central tendency.

6.7.2. Coefficient of mean deviation. The relative measure corresponding to the mean derivation is called coefficient of mean deviation. It is obtained by dividing the mean deviation by the particular average used in computing mean deviation. The coefficient of mean deviation can be computed by the following three formulae:

$$\text{Coefficient of mean deviation about mean} = \frac{\text{M.D.}(\bar{x})}{\bar{x}} \times 100;$$

$$\text{Coefficient of mean deviation about median} = \frac{\text{M.D.}(M_e)}{M_e} \times 100;$$

$$\text{Coefficient of mean deviation about mode} = \frac{\text{M.D.}(M_o)}{M_o} \times 100.$$

Like mean deviation, coefficient of mean deviation is also least when measured from the median.

Example 6.7.1. The following data relate to the marks obtained by nine students in a class test:

3, 4, 4, 5, 5, 7, 7, 7, 12

Find mean deviation about mean, median, mode and comment.

Solution.

$$\text{Mean} = \bar{x} = \frac{3+4+4+5+5+7+7+7+12}{9} = \frac{54}{9} = 6$$

$$\text{Median} = M_e = \text{Size of } \frac{n+1}{2}^{\text{th item}} = \frac{9+1}{2} = 5^{\text{th item of the arranged data}}$$

3, 4, 4, 5, 5, 7, 7, 7, 12.

The size of the 5th item = 5. Hence median = 5

Mode = $M_o = 7$ since 7 occurs three times.

Computation of Mean Deviation

Number of marks	Deviation from mean $ D = x-6 $	Deviation from median $ D = x-5 $	Deviation from mode $ D = x-7 $
3	3	2	4
4	2	1	3
4	2	1	3
5	1	0	2
5	1	0	2
7	1	2	0
7	1	2	0
7	1	2	0
12	6	7	5
$\Sigma x = 80$	$\Sigma D = 18$	$\Sigma D = 17$	$\Sigma D = 19$

$$M.D.(\bar{x}) = \frac{\sum |x - \bar{x}|}{n} = \frac{18}{9} = 2$$

$$M.D.(M_e) = \frac{\sum |x - M_e|}{n} = \frac{17}{9} = 1.78$$

$$M.D.(M_o) = \frac{\sum |x - M_o|}{n} = \frac{19}{9} = 2.11$$

Comment. It is seen that mean deviation about median is the least.

Example 6.7.2. The following data refer to the marks obtained by 7 students in a class test:

1, 4, 5, 3, 9, 3, 10

Find mean deviation about

- i) Mean,
- ii) Median and
- iii) Mode and comment.

Solution.

$$\text{Mean} = \bar{x} = \frac{1+4+5+3+9+3+10}{7} = \frac{35}{7} = 5$$

$$\text{Median} = M_e = \text{Size of } \frac{n+1}{2} \text{ th item} = \frac{7+1}{2}$$

= 4th item of the arranged data 1, 3, 3, 4, 5, 9, 10

The size of the 4th item = 4, Hence, median = 4

Mode = $M_o = 3$ since 3 occurs twice.

Computation of Mean Deviations

Marks	Deviation from mean $ D = x - 5 $	Deviation from median $ D = x - 4 $	Deviation from mode $ D = x - 3 $
1	4	3	2
3	2	1	0
3	2	1	0
4	1	0	1
5	0	1	2
9	4	5	6
10	5	6	7
$\Sigma x = 80$	$\Sigma D = 18$	$\Sigma D = 17$	$\Sigma D = 18$

$$M.D.(\bar{x}) = \frac{\sum |x - \bar{x}|}{n} = \frac{18}{7} = 2.57$$

$$M.D.(M_e) = \frac{\sum |x - M_e|}{n} = \frac{17}{7} = 2.43$$

$$M.D.(M_o) = \frac{\sum |x - M_o|}{n} = \frac{18}{7} = 2.57$$

Comment. It is seen that mean deviation about median is least.

A matched Problem. Compute mean deviation about mean, median and mode from the following set of data: 10, 3, 2, 4, 5, 2, 9 Ans. 2.5, 2.4, 3

Example 6.7.3. The following data refer to number of years worked by 9 employees of a factory: 7, 4, 10, 9, 15, 12, 7, 9, 7.

Compute the mean deviation from

- i) Mean,
- ii) Median,
- iii) Mode and
- iv) Show that the mean deviation from median is minimum.
- v) Also compute coefficient of mean deviation from mean, median, mode and comment.

Solution.

$$\text{Mean} = \bar{x} = \frac{7+4+10+9+15+12+7+9+7}{9} = \frac{80}{9} = 8.89 \text{ years}$$

Median = M_e = Size of $\frac{n+1}{2}$ th item = $\frac{9+1}{2} = 5$ th item of the arranged data 4, 7, 7, 7, 9, 9, 10, 12, 15

The size of the 5th item = 9. Hence median = 9 years.

Mode = M_o = 7 years (since 7 has the highest frequency)

Computation of Mean Deviations

No. of years worked	Deviation from mean $ D = x - \bar{x} $	Deviation from median $ D = x - M_e $	Deviation from mode $ D = x - M_o $
7	1.9	2	0
4	4.9	5	3
10	1.1	1	3
9	0.1	0	2
15	6.1	6	8
12	3.1	3	5
7	1.9	2	0
9	0.1	0	2
7	1.9	2	0
$\Sigma x = 80$	$\Sigma D = 21.1$	$\Sigma D = 21$	$\Sigma D = 23$

$$M.D.(\bar{x}) = \frac{\sum |x - \bar{x}|}{n} = \frac{21.1}{9} = 2.34 \text{ years.}$$

$$M.D.(M_e) = M_e = \frac{\sum |x - M_e|}{n} = \frac{21}{9} = 2.33 \text{ years.}$$

$$M.D.(M_o) = \frac{\sum |x - M_o|}{n} = \frac{23}{9} = 2.56 \text{ years.}$$

Coefficient of mean deviation about mean

$$= \frac{M.D.(\bar{x})}{\bar{x}} \times 100 = \frac{2.34}{8.9} \times 100 = 26.29\%.$$

Coefficient of mean deviation about median

$$= \frac{M.D.(M_e)}{M_e} \times 100 = \frac{2.33}{9} \times 100 = 25.89\%.$$

Coefficient of mean deviation about mode

$$= \frac{M.D.(M_o)}{M_o} \times 100 = \frac{2.56}{7} \times 100 = 36.57\%.$$

From these calculations it is clear that both the mean deviation and the coefficient of mean deviations are the least when measured from the median.

A matched problem. The daily wages (in Taka) of 7 workers of a factory are as follows: 52, 40, 47, 51, 67, 60 and 47. Compute mean deviations and coefficient of mean deviations from mean, median. Mode and comment.

Ans. 6.57, 6.43, 7, 12.63, 12.6, 14.89.

6.7.3. Mean deviation for grouped data. Suppose x_1, x_2, \dots, x_k are k mid-points of k classes with f_1, f_2, \dots, f_k are their respective class frequencies. If \bar{x} , M_e and M_o are the mean, median and mode of the frequency distribution, then mean deviations for the grouped data are defined as follows:

$$\text{Mean deviation about mean} = M.D.(\bar{x}) = \frac{\sum f|x - \bar{x}|}{n}$$

$$\text{Mean deviation about median} = M.D.(M_e) = \frac{\sum f|x - M_e|}{n}$$

$$\text{Mean deviation about mode} = M.D.(M_o) = \frac{\sum f|x - M_o|}{n}$$

Example 6.7.4. Calculate mean deviation and coefficient of mean deviation from mean from the following frequency distribution:

Class interval:	1-3	3-5	5-7	7-9	9-11
Frequency:	1	4	6	4	1

Solution. Calculation of Mean Deviation :

Class interval	Midpoint : x	F	fx	$ D = x - \bar{x} $	$f D $
1-3	2	1	2	4	4
3-5	4	4	16	2	8
5-7	6	6	36	0	0
7-9	8	4	32	2	8
9-11	10	1	10	4	4
Total		$n = 16$	96		24

$$\text{Mean} = \frac{\sum fx}{n} = \frac{96}{16} = 6.$$

$$\text{M.D.}(\bar{x}) = \frac{\sum f|x - \bar{x}|}{n} = \frac{24}{16} = 1.5.$$

$$\text{Coefficient of M.D.}(\bar{x}) = \frac{\text{M.D.}(\bar{x})}{\bar{x}} \times 100 = \frac{1.5}{6} \times 100 = 25\%.$$

6.7.4. Merits and demerits of Mean deviation.

Merits

- (i) It is easy to understand and easy to compute.
- (ii) It is based on all the observations
- (iii) It is less affected by extreme values.
- (iv) Since deviations are taken from the central value, comparison about formation of different distributions can easily be made.

Demerits and limitations

- (i) The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the observations from the central values.
- (ii) This method may not give us very accurate results.
- (iii) It is not capable of further algebraic treatment.
- (iv) It is rarely used in sociological and business studies.

6.8. Variance and Standard Deviation

The variance and its square root called the standard deviation are the most powerful measures of dispersion, which take into account how all the observations in the data are distributed around the most important central

value the arithmetic mean of the data. Karl Pearson introduced the concept of standard deviation in 1893.

Definition. The arithmetic mean of the squares of the deviations of the observations from their arithmetic mean is known as variance.

Definition. The positive square root of variance is called standard deviation.

The variance and the standard deviation measure the average scatter of the observations around the mean. Now we shall define population and sample variance and standard deviation. Population variance is denoted by the Greek symbol σ^2 and the deviations of the observations are taken from the population mean μ . Whereas, sample variance is denoted by s^2 and the deviations of the observations are taken from the sample mean \bar{x} . Population standard deviation is denoted by σ and sample standard deviation is denoted by s .

Population Variance. Suppose X_1, X_2, \dots, X_N are N observations of a population and μ its mean then population variance denoted by σ^2 is defined as

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

Population standard deviation is the square root of the population variance. It is calculated by the following formula:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

6.8.1. Interpreting the value of standard deviation. If all the observations are very close to each other, the standard deviation is close to zero. It is zero when all observations have the same value. In this case the data set contains no dispersion. If the observations are well dispersed, the standard deviation will tend to be large. That is smaller standard deviation ensures the consistency of the data set.

In practical work, we always use the value of standard deviation rather than the value of the variance, for one reason that the variance is measured in squared unit such as squared Tk., squared inches, and squared pounds and so on, since we are squaring the deviations. Thus the important absolute measure of dispersion is the standard deviation, which value is the original units of the data such as Tk., inches or pounds.

Example 6.8.1. The number of employees at six different drugstores of a small city are 4, 5, 6, 6, 7, 8. Find the variance and standard deviation of the data set taking it as population.

Solution.

$$\text{Population mean} = \mu = \frac{4+5+6+6+7+8}{6} = \frac{36}{6} = 6$$

Population variance,

$$\begin{aligned}\sigma^2 &= \frac{\sum(X-\mu)^2}{N} = \frac{\sum(X-6)^2}{6} \\ &= \frac{(4-6)^2 + (5-6)^2 + (6-6)^2 + (6-6)^2 + (7-6)^2 + (8-6)^2}{6} \\ &= \frac{4+1+0+0+1+4}{6} = \frac{10}{6} = \frac{5}{3} = 1.67.\end{aligned}$$

The standard deviation is then given by $\sigma = \sqrt{1.67} = 1.29$.

In most statistical applications, the data being analyzed is a sample. When we compute a sample variance, we are often interested in using it to estimate the population variance σ^2 . To get an unbiased estimate of the population variance σ^2 , we replace n by $n-1$ in the divisor in computing the sample variance s^2 .

Sample Variance. Suppose x_1, x_2, \dots, x_n are n values of a sample data and \bar{x} its sample mean then the sample variance denoted by s^2 is defined as

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

The sample standard deviation s is the positive square root of the sample variance. Hence it is defined by

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Steps for computing sample variance and standard deviation:

1. First compute sample mean.
2. Obtain the difference between each observation and the mean.
3. Square each difference.
4. Add the squared differences.
5. Divide this total by $n - 1$ and obtain s^2

To compute s , the sample standard deviation, take the positive square root of the sample variance.

Example 6.8.2. A comparison of tea prices at 4 randomly selected grocery stores in an area of Chittagong city showed increase of 12, 15, 17 and 20 Tk.s per kilogram from the previous month. Find the standard deviation of the price increases.

Solution. The sample mean is

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = \frac{64}{4} = \text{Tk. } 16.$$

Therefore,

$$\begin{aligned}s^2 &= \frac{\sum (x - \bar{x})^2}{n-1} = \frac{\sum (x - 16)^2}{3} \\&= \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3} \\&= \frac{16+1+1+16}{3} = \frac{34}{3} = 11.33.\end{aligned}$$

The sample standard deviation is $s = \sqrt{11.33} = \text{Tk. } 3.37$.

If \bar{x} is a decimal number that has been rounded off, we may accumulate a large error using the sample variance or standard deviation formulae in the above forms. To avoid the error, we state an equivalent but more useful computation formula.

The sample variance s^2 may be written as

$$s^2 = \frac{n \sum x^2 - (\sum x)^2}{n(n-1)}$$

Another computation formula for s^2 is

$$s^2 = \frac{1}{n-1} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right].$$

These two formulae for computing s^2 are equivalent.

Similarly, the computation formula for population variance is

$$\sigma^2 = \frac{1}{N} \left[\sum X^2 - \frac{(\sum X)^2}{N} \right].$$

A matched problem. The following data refers to the marks obtained by 5 students randomly selected from a class: 9, 7, 8, 6, 5. Find variance of the data set.

Ans.2.5

Example 6.8.3. The grade point averages obtained by randomly selected 6 students in their H.S.C examination are as follows: 4.9, 4.1, 4.4, 3.3, 4.6 and 4.8. Compute the standard deviation and comment.

Solution. Now, we can make a table for values of x and x^2 as follows:

x	x^2
4.9	24.01
4.1	16.81
4.4	19.36
3.3	10.89
4.6	21.16
4.8	23.04
$\Sigma x = 26.1$	$\Sigma x^2 = 115.27$

Here $n = 6$

$$\text{Hence } s^2 = \frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}$$

$$= \frac{(6)(115.27) - (26.1)^2}{(6)(5)} = \frac{691.62 - 681.21}{30} = \frac{10.41}{30} = 0.347.$$

The standard deviation, $s = \sqrt{0.347} = 0.59$.

Comment: Since the grade-point averages of all the students are more than 4 and the standard deviation is only 0.59, the quality of the students are very good (grade points of the student are close to each other and around 4).

Sample variance by the second computational formula

$$s^2 = \frac{1}{n-1} \left[\Sigma x^2 - \frac{(\Sigma x)^2}{n} \right] = \frac{1}{5} \left[115.27 - \frac{(26.1)^2}{6} \right]$$

$$= \frac{1}{5} \left[115.27 - \frac{681.21}{6} \right] = \frac{1}{5} [115.27 - 113.535] = \frac{1.735}{5} = 0.347.$$

Notation	
Sample	Population
n : number of observations in the sample	N : number of observations in the population
s^2 : sample variance	σ^2 : population variance
$s = \sqrt{s^2}$: sample standard deviation	$\sigma = \sqrt{\sigma^2}$: population standard deviation

6.8.2. Sample variance for grouped data. Let x_1, x_2, \dots, x_k be k values of a variable or k mid points of k classes with corresponding frequencies f_1, f_2, \dots, f_k then the sample variance is defined by

$$s^2 = \frac{\sum f(x - \bar{x})^2}{n-1}; \text{ where } n = \sum f.$$

For convenience and simplicity most of the textbooks use the divisor n in place of $n-1$ and they use the following formula for sample variance:

$$s^2 = \frac{\sum f(x - \bar{x})^2}{n};$$

However, the computing formula for sample variance is

$$s^2 = \frac{1}{n-1} \left[\sum f x^2 - \frac{(\sum f x)^2}{n} \right] \quad \text{or, } s^2 = \frac{n \sum f x^2 - (\sum f x)^2}{n(n-1)}.$$

Example 6.8.4. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate variance and standard deviation of the frequency distribution.

Solution. Calculation of variance and standard deviation

Class interval	Mid-point : x	Frequency : f	fx	fx^2
30-55	42.5	3	127.5	5418.75
55-80	67.5	4	270.0	18225.00
80-105	92.5	6	555.0	51337.50
105-130	117.5	9	1057.5	124256.25
130-155	142.5	12	1710.0	243675.00
155-180	167.5	11	1842.5	308618.75
180-205	192.5	5	9672.5	185281.25
Total		50	$\sum f x = 6525$	$\sum f x^2 = 936812.5$

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left[\sum f x^2 - \frac{(\sum f x)^2}{n} \right] = \frac{1}{49} \left[936812.50 - \frac{(6525)^2}{50} \right] \\ &= \frac{1}{49} [936812.50 - 851512.50] = \frac{85300}{49} = 1740.82 = 1740.82 \end{aligned}$$

Hence, $s = \sqrt{1740.82} = 41.72$.

6.8.3. Calculation of variance and standard deviation by short cut method. Suppose x_1, x_2, \dots, x_k are k values of a variable or k mid points of k classes with corresponding frequencies f_1, f_2, \dots, f_k then the sample variance is

$$s^2 = \frac{\sum f(x - \bar{x})^2}{n-1}; \quad \text{where } n = \sum f$$

Let, $d = \frac{x - A}{i}$; then $x = A + id$ and $\bar{x} = A + i\bar{d}$

Here A is called assumed mean which is usually taken as middle value of x or the value of x which has the highest frequency to get the maximum benefit of calculation and i is the width of the class interval.

Substituting the value of $(x - \bar{x})$ in the formula of s^2 , we have

$$s^2 = \frac{\sum f(x - \bar{x})^2}{n-1} = \frac{\sum f[i(d - \bar{d})]^2}{n-1} = \frac{\sum f(d - \bar{d})^2}{n-1} \times i^2 = \frac{1}{n-1} \left[\sum fd^2 - \frac{(\sum fd)^2}{n} \right] \times i^2$$

Then standard deviation is computed by the following formula,

$$s = \sqrt{\frac{1}{n-1} \left[\sum fd^2 - \frac{(\sum fd)^2}{n} \right]} \times i$$

Example 6.8.5. The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory.

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate variance and standard deviation of the frequency distribution by the short cut method.

Solution. Let $A = 117.5$ and $i = 25$; then $d = \frac{x - A}{i} = \frac{x - 117.5}{25}$

Calculation of variance and standard deviation.

Class interval	Mid-point : x	Frequency : F	$d = \frac{x - 117.5}{25}$	fd	fd^2
30-55	42.5	3	-3	-9	27
55-80	67.5	4	-2	-8	16
80-105	92.5	6	-1	-6	6
105-130	117.5	9	0	0	0
130-155	142.5	12	1	12	12
155-180	167.5	11	2	22	44
180-205	192.5	5	3	15	45
Total		50		$\Sigma fd = 26$	$\Sigma fd^2 = 150$

$$\text{Variance} = s^2 = \frac{1}{n-1} \left[\sum fd^2 - \frac{(\sum fd)^2}{n} \right] \times i^2$$

$$\begin{aligned}
 &= \frac{1}{49} \left[150 - \frac{(26)^2}{50} \right] \times (25)^2 \\
 &= \frac{1}{49} [150 - 13.52] \times 625 \\
 &= \frac{136.48 \times 625}{49} = 1740.82.
 \end{aligned}$$

Hence the standard deviation, $s = \sqrt{1740.82} = 41.72$.

From the computation point of view, it is the easiest method of calculating variance as well as standard deviation.

6.8.4. Merits and demerits of standard deviation. Merits and demerits of standard deviation are as follows:

Merits

- (i) It is rigidly defined.
- (ii) It is based on all observations of the distribution.
- (iii) It is amenable to algebraic treatment.
- (iv) It is less affected by sampling fluctuation.
- (v) It is possible to calculate the combined standard deviation of two or more groups.

Demerits and Limitations

- (i) As compared to other measures it is difficult to compute.
- (ii) It is affected by the extreme values.
- (iii) It is not useful to compare two sets of data when the observations are measured in different units.

6.8.5. Some comments on standard deviation. It is the best absolute measure of dispersion. Some important characteristics of standard deviation are:

1. The value s is always greater than or equal to zero.
2. The larger the value of s , the greater the variability of the data set.
3. If s is equal to zero, all the observations must have the same value. That is there is no variability among the observations in the data set.
4. The original variable and the standard deviation have the same unit of measurement.

General comments on the absolute measures of dispersion

1. The more spread out or dispersed the data are, the larger will be the range, the quartile deviation, the variance and the standard deviation.

2. The more concentrated or homogenous the data are, the smaller will be the range, the quartile deviation, the variance and the standard deviation.
3. If the observations are all the same (so that there is no variation in the data), the range, the quartile deviation, the variance and the standard deviation will be all zero.
4. None of the measures of dispersion can be negative.

6.8.6. Empirical Relations among the measures of Dispersion. For a symmetrical and moderately skewed distribution there exist relationships among the three commonly used measures of dispersion. The quartile deviation (Q.D.) is the smallest, following the mean deviation (M.D.) and the standard deviation (S.D.) is the largest. Following are the relationships among themselves:

$$Q.D. = \frac{2}{3}\sigma ; \quad M.D. = \frac{4}{5}\sigma ; \quad \text{and} \quad Q.D. = \frac{5}{6}M.D.$$

They are useful in estimating one measure of dispersion when another is known or in verifying the consistency of the calculated values roughly. If the computed σ differs very widely from its value estimated from Q.D. or M.D. either an error has been made or the distribution differs considerably from symmetry.

Another comparison may be made of the proportion of observations that are typically included within the range of one Q.D., M.D. or S.D. measured both above and below the mean. For a normal distribution:

$\bar{x} \pm Q.D.$, includes 50% of the observations.

$\bar{x} \pm M.D.$, includes 57.51% of the observations.

$\bar{x} \pm S.D.$, includes 68.27% of the observations.

Example 6.8.6. Suppose $Q_1 = 15$ and $Q_3 = 40$ for a symmetrical distribution, find Q.D., M.D., and S.D. of the distribution.

$$\text{Solution. } Q.D. = \frac{Q_3 - Q_1}{2} = \frac{40 - 15}{2} = 12.5$$

We know that following relationships hold for a symmetrical distribution

$$Q.D. = \frac{2}{3}\sigma ; \quad M.D. = \frac{4}{5}\sigma ; \quad \text{and} \quad Q.D. = \frac{5}{6}M.D.$$

$$\text{Hence, } M.D. = \frac{6}{5}Q.D. = \frac{6}{5} \times 12.5 = 15$$

$$S.D. = \sigma = \frac{3}{2}Q.D. = \frac{3}{2} \times 12.5 = 18.75.$$

6.9. Coefficient of Variation

Standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the coefficient of variation. This measure developed by Karl Pearson is the most commonly used measure of relative dispersion.

Definition. If μ is the mean and σ is the standard deviation of a population data set, then coefficient of variation denoted by C.V. is defined by

$$C.V. = \frac{\sigma}{\mu} \times 100.$$

If \bar{x} is the mean and s is the standard deviation of a sample data set, then coefficient of variation is defined by

$$C.V. = \frac{s}{\bar{x}} \times 100.$$

It is a pure number and expressed as a percentage. It is useful in comparing the variability of two or more sets of data, especially if they are expressed in different units of measurement. Since it is a ratio, the units of measurement have no significance. Moreover, coefficient of variation depends on the best measure of central tendency and the best measure of dispersion. In these reasons, coefficient of variation is better than standard deviation as a measure of dispersion. Overall, coefficient of variation is the best measure of dispersion among all the absolute and relative measures.

Comment. Coefficient of variation is the best measure of dispersion.

Example 6.9.1. The grade point averages obtained by randomly selected 6 students in their H.S.C. examination are as follows: 4.9, 4.1, 4.4, 3.3, 4.6 and 4.8. Find mean, standard deviation and coefficient of variation.

Solution. We know from the problem 6.8.3 that the standard deviation of the data set is 0.59 and $\Sigma x = 26.1$. Then $\bar{x} = \frac{26.1}{6} = 4.35$.

$$\text{Hence, } C.V. = \frac{s}{\bar{x}} \times 100 = \frac{0.59}{4.35} \times 100 = 13.56.$$

Among two or more data sets, the set for which coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogenous. On the other hand, the series for which coefficient of variation is less is said to be less variability or more consistent, more uniform, more stable or more homogenous.

Example 6.9.2. Lives of two models of refrigerators in a recent survey were found as follows:

Life (no. of years)	Model A	Model B
0-2	5	2
2-4	16	7
4-6	13	12
6-8	7	19
8-10	5	9
10-12	4	1

- i) What is the average life of each model of these refrigerators?
- ii) Which of the two models shows more uniformity?
- iii) A person wants to buy a new refrigerator, which one will he prefer?

Solution. For finding the average lifetimes, we have to compute arithmetic mean and for determining the model which has greater uniformity, then compute and compare the coefficient of variations.

Class interval	Mid-points x	Model A			Model B		
		f	fx	fx ²	f	fx	fx ²
0-2	1	5	5	5	2	2	2
2-4	3	16	48	144	7	21	63
4-6	5	13	65	325	12	60	300
6-8	7	7	49	343	19	133	931
8-10	9	5	35	405	9	81	729
10-12	11	4	44	484	1	11	121
Total		50	256	1706	50	308	2146

Computations of mean, variances and co-efficient of variations of lifetimes for two models are shown below:

Model A	Model B
Arithmetic Mean, $\bar{x}_A = \frac{256}{50} = 5.12$ years	$\bar{x}_B = \frac{304}{50} = 6.16$ years
$s_A^2 = \frac{1}{49} \left[1706 - \frac{(256)^2}{50} \right]$ $= \frac{1}{49} [1706 - 1310.72]$ $= \frac{395.28}{49} = 8.07$ $s_A = \sqrt{8.07} = 2.84$ years	$s_B^2 = \frac{1}{49} \left[2146 - \frac{(308)^2}{50} \right]$ $= \frac{1}{49} [2146 - 1897.28]$ $= \frac{248.72}{49} = 5.08$ $s_B = \sqrt{5.08} = 2.25$ years

$$\text{C.V.}(A) = \frac{2.84}{5.12} \times 100 = 55.47\% \quad \text{C.V.}(B) = \frac{2.25}{6.16} \times 100 = 36.53\%.$$

- i) Average lifetimes of refrigerators of Model A is 5.12 years, while of Model B is 6.16 years.

- ii) Since coefficient of variation is less for Model B, hence refrigerators of Model B show greater uniformity as per the lifetime of the refrigerators.
- iii) Due to the greater uniformity in lifetime, the person will prefer Model B.

Example 6.9.3. The following are some of the particulars of the distribution of weights of boys and girls in a class.

	Boys	Girls
Number :	65	35
Mean weight :	60kgs	45kgs
Standard deviation :	4kgs	2kgs

The weights of which distribution is more homogenous?

Solution. We can consider the data from two populations. To compare the variability of the weights of two distributions, we will have to find the coefficient of variations of the two distributions.

Boys	Girls
$C.V. = \frac{\sigma}{\mu} \times 100 = \frac{4}{60} \times 100 = 6.67\%$	$C.V. = \frac{\sigma}{\mu} \times 100 = \frac{2}{45} \times 100 = 4.44\%$

Comment. Since the coefficient of variation is greater for boys, hence the weights of the boys is more variability and less homogenous.

Example 6.9.4. The following are some of the particulars of the distributions of weights and heights of 100 boys in a class.

	Weights	Heights
Mean weight :	60kgs	65 inches
Standard deviation :	7kgs	4 inches

Which distribution shows greater variability?

Solution. For determining variability between the two distributions, we have to compute coefficient of variations of weights and heights of the boys.

$$C.V. (\text{Weight}) = \frac{\sigma}{\mu} \times 100 = \frac{7}{60} \times 100 = 11.67\%$$

$$C.V. (\text{Heights}) = \frac{\sigma}{\mu} \times 100 = \frac{4}{65} \times 100 = 6.15\%$$

Comment. Since coefficient of variation is higher for weights, hence the distribution of weights shows greater variability than heights.

Example 6.9.5. In two factories A and B engaged in the same industry, the average weekly wages and standard deviations are as follows:

	Factory A	Factory B
Average weekly wages (in Taka) :	860	900
Standard deviation of wages (in Taka) :	50	80

- i) Which factory shows greater variability in the distribution of wages?
- ii) A person got job in both the factories, which factory he will prefer to join?

Solution. For comparing the variability of wages of the two factories, we have to compute coefficient of variations.

Factory A	Factory B
$C.V. = \frac{\sigma}{\mu} \times 100 = \frac{50}{860} \times 100 = 5.81\%$	$C.V. = \frac{\sigma}{\mu} \times 100 = \frac{80}{900} \times 100 = 8.89\%.$

Comment. The coefficient of variation is greater for factory B; hence factory B shows greater variability in the distribution of wages. The person will join factory A, since the coefficient of variation of the factory A is less than the factory B, although the average weekly wages of factory B is more than A.

Example 6.9.6. Two cricketers scored the following runs in randomly selected 10 one-day matches:

Player A	42	32	40	45	17	83	59	64	76	72
Player B	95	3	28	70	31	14	82	0	59	108

- i) Who is the better run-getter?
- ii) Who is the consistent player?
- iii) A prize is given to the best player. Who will get the prize?

Solution. In order to find out who is better run-getter, we will compare the average runs scored and to find out who is more consistent, we will compare the coefficient of variations.

Calculation of Mean and Coefficient of variation

Cricketer A : x	x^2	Cricketer B : x	x^2
42	1764	95	9025
32	1024	3	9
40	1600	28	784
45	2025	70	4900
17	289	31	961
83	6889	14	196
59	3481	82	6724
64	4096	0	0
76	5776	59	3481
72	5184	108	11664
530	32127	490	37744

Cricketer A

$$\bar{x}_A = \frac{\Sigma x}{n} = \frac{530}{10} = 53$$

Cricketer B

$$\bar{x}_B = \frac{\Sigma x}{n} = \frac{490}{10} = 49$$

(i) Since the average score for the player A is higher than player B, hence A is a better run-getter. However,

$$s_A^2 = \frac{1}{9} \left(32127 - \frac{(530)^2}{10} \right)$$

$$= \frac{1}{9} (32127 - 28090) = 448.56$$

Hence $s_A = 21.18$

$$C.V.(A) = \frac{21.18}{53} \times 100 = 39.96\%$$

$$s_B^2 = \frac{1}{9} \left(37744 - \frac{(490)^2}{10} \right)$$

$$= \frac{1}{9} (37744 - 24010) = 1526$$

Hence $s_B = 39.06$

$$C.V.(B) = \frac{39.06}{49} \times 100 = 79.71\%.$$

- (ii) The coefficient of variation for player A is less than player B; hence player A is more consistent.
 (iii) The player A will get the prize.

6.9.1. Advantage of coefficient of variation over standard deviation. Standard deviation and the original variable have the same unit of measurements. So standard deviation cannot be used to compare the variability of two or more distributions measured in different units. But coefficient of variation is a pure number. That is coefficient of variation is independent of the unit of measurement of the original variables. Therefore, coefficient of variation can be successfully used to compare the variability of two or more distributions in different units of measurement. Hence coefficient of variation is sometimes better than standard deviation as measure of dispersion.

6.9.2. Practical Significance of the standard deviation (Chebyshev's Theorem). The two summary values most often used by statisticians are the mean and standard deviation. If the observations of a data set have a small standard deviation, we would expect most of the values to be grouped around the mean. However, a large value of the standard deviation indicates a greater variability, and therefore we would expect the observations to be more spread out the mean. The Russian mathematician P.L.Chebyshev (1821-1894) discovered that the fraction of the observations falling between any two values symmetric about the mean is related to the standard deviation. Chebyshev's theorem gives a conservative estimate of the fraction of measurements falling within k standard deviation of the mean, that means within $(\mu - k\sigma)$ to $(\mu + k\sigma)$, in usual notations.

Chebyshev's Theorem.

Given a number k greater than or equal to 1, at least the fraction $(1 - 1/k^2)$ of the observations of any set of data lies within k standard deviations of the mean of the observations.

Chebyshev's theorem can be applied to any set of data and can be used to describe either a sample or a population.

Now let us choose a few numerical values for k and compute $(1 - 1/k^2)$.

k	$(1 - 1/k^2)$
1	$1 - 1 = 0$
2	$1 - \frac{1}{4} = \frac{3}{4}$
3	$1 - \frac{1}{9} = \frac{8}{9}$

Thus, the theorem states that, in case of a Population:

1. For $k = 1$, at least none of the observations lie in the interval $\mu - \sigma$ to $\mu + \sigma$.
2. For $k=2$, at least $3/4$ or 75% of the observations lie in the interval $\mu - 2\sigma$ to $\mu + 2\sigma$.
3. For $k=3$, at least $8/9$ or 89% of the observations lie in the interval $\mu - 3\sigma$ to $\mu + 3\sigma$.

And in case of a Sample

1. For $k = 1$, at least none of the observations lie in the interval $\bar{x} - s$ to $\bar{x} + s$.
2. For $k = 2$, at least $3/4$ or 75% of the observations lie in the interval $\bar{x} - 2s$ to $\bar{x} + 2s$.
3. For $k = 3$, at least $8/9$ or 89% of the observations lie in the interval $\bar{x} - 3s$ to $\bar{x} + 3s$.

The first statement does not provide any valuable information about the set of observations, but the other two values of k provide valuable information about the proportion of observations that fall in certain intervals. One can use any value of k for finding the desired proportion of observations, for example, the proportion of observations that fall within $k = 2.5$ standard deviations of the mean is at least $1 - 1/(2.5)^2 = 0.84$ or 84%.

Example 6.9.7. A population of 180 observations has a mean of 32 and a standard deviation of 4. According to Chebyshev's theorem

- (a) How many of the observations fall in the interval from 24 to 40?

- (b) Within what interval will at least 160 of the observations fall?
- (c) Outside what interval will at most 45 of the observations fall?
- (d) How many of the observations fall outside the interval from 40 to 44?

Solution. (a) Here $\mu = 32$ and $\sigma = 4$. The interval from 24 to 40 may be written as $32 \pm (2)(4)$ from which we see that $k = 2$. Hence at least $3/4$ or 75% that is 135 of the 180 observations fall in the interval from 24 to 40.

(b) Solving the equation

$$1 - 1/k^2 = 160/180 = 8/9.$$

We find that $k = 3$. Therefore, at least 160 of the observations fall in the interval $32 \pm (3)(4)$ or from 20 to 44.

(c) At most 45 observations will fall outside an interval that contains at least $180 - 45 = 135$ observations. From problem 1 we see that the desired interval goes from 24 to 40.

(d) According to problem 1, at least 160 of the observations fall in the interval from 20 to 44. Therefore, at most $180 - 160 = 20$ of the observations fall outside of this interval.

Example 6.9.8. The mean and standard deviation of sample of $n = 25$ observations are 75 and 10. Use Chebyshev's theorem to describe the distribution of observations.

Solution. We are given $\bar{x} = 75$ and $s = 10$. Chebyshev's Theorem states

- (a) At least $3/4$ or 75% of the 25 observations lie in the interval $\bar{x} - 2s$ to $\bar{x} + 2s = 75 - 2(10)$ to $75 + 2(10)$ - that is, 55 to 95.
- (b) At least $8/9$ or 89% of the 25 observations lay in the interval $\bar{x} - 3s$ to $\bar{x} + 3s = 75 - 3(10)$ to $75 + 3(10)$ - that is 45 to 105.

Matched problem

The mean and standard deviation of sample of $n = 25$ observations are 21.6 and 5.5. Describe the sample data using the Chebyshev's Theorem.

Ans. At least 75% observations will fall between 10.6 and 32.6.
At least 89% observations will fall between 5.1 and 38.1.

Chebyshev's theorem holds for any distribution of observations and, for this reason, the results are usually weak. The value given by the theorem is a lower bound only. When the data set is closer to bell-shaper or normal distribution, another rule for describing the variability of a data set works well is the Empirical Rule.

Empirical Rule

Population

For a population data set having a bell - shaped distribution:

The interval $\mu \pm \sigma$ contains approximately 68% of the observations.

The interval $\mu \pm 2\sigma$ contains approximately 95% of the observations.

The interval $\mu \pm 3\sigma$ contains all or almost all the observations.

Sample

For a sample data set having a bell - shaped distribution:

The interval $\bar{x} \pm s$ contains approximately 68% of the observations.

The interval $\bar{x} \pm 2s$ contains approximately 95% of the observations.

The interval $\bar{x} \pm 3s$ contains all or almost all the observations.

Example 6.9.9. In a time study conducted at a manufacturing plant, the length of time to complete a specific operation is measured for each of $n = 40$ workers. The mean and standard deviation are found to be 12.8 and 1.7 respectively. Describe the sample data using the Empirical Rule.

Solution. To describe the data, let us calculate the following intervals:

$$\bar{x} \pm s = 12.8 \pm 1.7, \quad \text{or } 11.1 \text{ to } 14.5$$

$$\bar{x} \pm 2s = 12.8 \pm 2(1.7) \quad \text{or } 9.4 \text{ to } 16.2$$

$$\bar{x} \pm 3s = 12.8 \pm 3(1.7) \quad \text{or } 7.7 \text{ to } 17.9$$

According to Empirical Rule, we expect approximately 68% of the observations to fall into the interval 11.12 to 14.5, approximately 95% to fall into the interval from 9.4 to 16.2 and all or almost all to fall into the interval from 7.7 to 17.9.

Comment. In case of doubt about the form of the distribution, it is wise to use Chebyshev's Theorem.

Matched problem

The distribution of a data set is relatively mound or bell-shaped with mean 50 and standard deviation 10

- What proportion of the observations will fall 40 to 60?
- What proportion of the observations will fall 30 to 70?
- What proportion of the observations will fall 30 to 60?
- If an observation is chosen at random from this distribution, what is the probability that it will be greater than 60?

Ans. a. approximately 0.68, b, approximately 0.95,
c. Approximately 0.815, d. approximately 0.16.

6.9.4 A check on the calculation of standard deviation s . Sometimes an error in the calculation of the standard deviation can go undetected.

Chebyshev's Theorem and Empirical Rule will help us to detect such an error. We know that about 75% or 95% of the data will lie in the interval $\bar{x} \pm 2s$. That is

$$\text{Range} = r = (\bar{x} + 2s) - (\bar{x} - 2s) = 4s$$

That is $s = r/4$

Similarly, about 89% or 100% of the data will lie in the interval $\bar{x} \pm 3s$.

$$\text{Range} = r = \bar{x} + 3s - (\bar{x} - 3s) = 6s$$

That is, $s = r/6$.

Thus the value of s is correct if it is close to $r/4$ or $r/6$. If it is not close to $r/4$ or $r/6$, then there is an error in the calculation of s .

Example 6.9.10. The variance of the data set given below 5, 7, 1, 2, 4 is 5.7. You might be oversight take the standard deviation to be 5.7. Check whether it is correct or not.

Solution. The range of the data set, $r = 7 - 1 = 6$

Then $s = 6/4 = 2.5$ or $s = 6/6 = 1$.

Conclusion. $s = 5.7$ is much larger than 2.5 or 1. So there is an error in the calculation of s . It is to be noted that the actual value of s is 2.29, which is close to 2.5.

6.10. Measure of Relative Standing

Consider a set of data: 1, 1, 0, 15, 2, 3, 4, 0, 1, 3

By looking at the data we observe that the value 15 is far greater than the rest of the observations. This type of observation is called outlier or extreme value or wild observation.

Definition. Values that lie far away from the rest are called outliers or extreme values or wild observations. A measure of relative standing helps in detecting the outliers of a set of observations.

6.10.1. Causes of outliers. An outlier may result from transposing digits from recording an observation, from incorrectly reading an instrument dial, from a malfunctioning piece of equipment, and from other problems. Even when there are no recording or observational errors, a data set may contain one or more valid observations that, for one reason or another, differ markedly from the others in the set. These outliers can cause a marked distortion in the values of commonly used numerical measures such as \bar{x} and s . In fact, outliers may themselves contain important information not

shared with the other observations in the set. Therefore, isolating outliers, if they are present, is an important step in any preliminary analysis of a data set.

A measure of relative standing helps in detecting the outliers of a set of observations.

The Box-and-whisker plot that will be discussed later on is also a method for isolating outliers.

Definition. The sample z - score is a measure of relative standing defined by

$$z = \frac{x - \bar{x}}{s}$$

A z-score measures the distance between the observation and the mean, measured in units of standard deviation. It is useful in determining outliers. According to Chebyshev's Theorem or the Empirical Rule about 75% observations lie in the interval $\bar{x} \pm 2s$.

That is there is a big chance that an observation x will lie in the interval

$$\bar{x} - 2s < x < \bar{x} + 2s$$

$$-2s < x - \bar{x} < 2s$$

$$-2 < \frac{x - \bar{x}}{s} < 2$$

If $z = \left| \frac{x - \bar{x}}{s} \right| \geq 2$, the observation x is suspected as an outlier. Similarly, it can be shown that the z score for all values are expected to lie within the interval - 3 to 3.

Conclusion. If z-score of an observation is greater than 2 but close 2, we say the value is suspected as an outlier. If it is above 3, we say that it is an extreme outlier. In between the value is an outlier.

Example 6.10.1. Consider this sample of $n = 10$ values

$$1, 1, 0, 15, 2, 3, 4, 0, 1, 3$$

Calculate the z score for 15 and state your conclusion.

Solution. For the sample $\Sigma x = 30$, $\Sigma x^2 = 266$, $\bar{x} = 3$ and $s^2 = 19.5556$

$$\Sigma x = 30, \Sigma x^2 = 266, \bar{x} = 3 \text{ and } s^2 = 19.5556$$

Hence, $s = 4.42$. The z - score for 15 is

$$z = \frac{15 - 3}{4.42} = 2.7$$

Conclusion. The value of z is far from 2 but less than 3. So we can say that the observation 15 is an outlier.

Matched problem

Calculate the z-score for the smallest and largest observations for these data:

8, 7, 1, 4, 6, 6, 4, 5, 7, 6, 3, 0

Identify whether they are outliers or not. Ans. -1.94, 1.33, they are not outliers.

6.11. Some Elementary Theorem and Examples

Theorem 6.11.1. Variance as well as standard deviation is independent of the shift of origin but depends on change of scale.

Proof. Suppose x_1, x_2, \dots, x_n are n values of a sample with mean \bar{x} , then variance is define by

$$s_x^2 = \frac{\sum(x - \bar{x})^2}{n-1} \quad \dots \dots \dots (6.11.1)$$

Then sample standard deviation is

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

$$\text{Let, } u = \frac{x - A}{h},$$

$$\text{Then, } x = A + hu \text{ and } \bar{x} = A + h\bar{u}$$

Now, by putting the values of x and \bar{x} in (6.11.1), we have

$$s_x^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{\sum(A + hu - A - h\bar{u})^2}{n-1} = h^2 \frac{\sum(u - \bar{u})^2}{n-1} = h^2 s_u^2.$$

That is variance of x is h^2 times variance of u. It is seen that variance of x depends on h but not on A. Hence variance is independent of the shift of origin but depends on scale. Moreover standard deviation of x is

$$s_x = hs_u$$

This shows that standard deviation also independent of the shift of origin but depends on scale.

Theorem 6.11.2. For two numbers standard deviation is the half of the range.

Proof. Let x_1 and x_2 are two quantities such that $x_1 > x_2$.

Then range is

$$R = x_1 - x_2 \text{ since } x_1 > x_2.$$

$$\text{Here, } \mu = \frac{x_1 + x_2}{2}$$

$$\begin{aligned} \text{Variance} = \sigma^2 &= \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2}{2} \\ &= \frac{\left(x_1 - \frac{x_1 + x_2}{2}\right)^2 + \left(x_2 - \frac{x_1 + x_2}{2}\right)^2}{2} = \frac{(x_1 - x_2)^2}{4} \end{aligned}$$

$$\text{Standard deviation} = \sigma = \frac{x_1 - x_2}{2} = \frac{R}{2}.$$

Hence for two quantities standard deviation is the half of the range.

From here it follows that for two positive quantities mean is always greater than standard deviation, since $\frac{x_1 + x_2}{2} \geq \frac{x_1 - x_2}{2}$.

Equality holds when any one of them is equal to zero.

Theorem 6.11.3. The variance of the first n natural numbers is $\frac{n^2 - 1}{12}$.

Proof. Let x be variable whose values are $1, 2, 3, \dots, n$.

$$\text{Then, Mean} = \mu = \frac{1+2+3+\dots+n}{n} = \frac{n(n+1)/2}{n} = \frac{n+1}{2}.$$

$$\text{Variance} = \sigma^2 = \frac{1}{n} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right]$$

$$\text{Here, } \sum x^2 = 1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\begin{aligned} \text{Variance} = \sigma^2 &= \frac{1}{n} \left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \\ &= \frac{1}{n} \left[\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \right] = \frac{n(n+1)}{n} \left[\frac{2n+1}{6} - \frac{n+1}{4} \right] \end{aligned}$$

$$= \frac{n+1}{12} \left[\frac{4n+2 - 3n-3}{12} \right] = \frac{(n+1)(n-1)}{12} = \frac{n^2 - 1}{12}$$

Hence the standard deviation of the first natural numbers is

$$\sigma = \sqrt{\frac{n^2 - 1}{12}}$$

Questions

1. What do you mean by dispersion? What are the important measures of dispersion? Define them.
2. What are the purposes of the measures of dispersion? State the desirable properties of good measures of dispersion. Which one is the important measure of absolute measure of dispersion?
3. Define standard deviation and coefficient of variation. Why sometimes coefficient of variation is better than standard deviation as a measure of dispersion?
4. What is coefficient of variation? What purpose does it serve? Distinguish between standard deviation and coefficient of variation. State some properties of standard deviation.
5. Why do we need the relative measures of dispersion? Explain the superiority of co-efficient of variation over standard deviation.
6. Define mean deviation. How can you measure mean deviation of a set of data? Which one is the best measure of mean deviation?
7. What are the relative measures of dispersion? Which one is the best and why?
8. What are the absolute measures of dispersion? Which one is the best and why?
9. State Chebyshev's theorem and Empirical rule related to it. Discuss the practical significance of standard deviation in the light of Chebyshev's theorem and Empirical rule.
10. State the five-number summary points. How can you construct Box and Whisker Plot?
11. Show that the standard deviation of a set of constants is equal to zero.

Exercise

12. Find the variance and standard deviation of the following sets of population data by using appropriate notations (i) 3, 4, 5, 6, 8, 9, 10, 12, 15; (ii) 7, 5, 9, 7, 8, 6. Ans.(i) Approx. 13.78 and 3.7; (ii) Approx. 1.67 and 1.29

13. Find the variance and standard deviation of the following sets of sample data with appropriate notations: (i) 3, 6, 7, 6, 4, 5; (ii) 20, 15, 12, 17.

Ans. (i) Approx. 2.17 (ii) Approx. 11.33

14. Find standard deviation of the sample data 1, 3, 4, 5, 3. Ans. 1.48

15. Find the standard deviation from the grouped sample data

Values of X : x	8	9	10	11	12
Frequency : f	1	2	4	2	1

Ans. Approx. 1.94

16. Find the mean deviation about mean of the sample data set 2, 3, 5, 7 and 8. Ans. 2

17. Suppose frequency distribution is almost mound-shaped with mean 50 and standard deviation 10.

- (i) What proportion of the observations will fall between 40 and 60?
(ii) What proportion of the observations will fall between 30 and 70?
(iii) What proportion of the observations will fall between 30 and 60?
(iv) If an observation is chosen at random from the distribution, what is the probability that it will be greater than 60?

Ans. (i) Approx. 68; (ii) Approx. 95 ; (iii) Approx. 815; (iv) Approx. 16.

18. The following data refer to the marks obtained by 25 students randomly selected from a class in a class test:

7, 6, 6, 11, 8, 9, 11, 9, 10, 8, 7, 7, 5,
9, 10, 7, 7, 7, 9, 12, 10, 10, 8, 6,

Compute (i) \bar{x} , s^2 and s for this sample. Ans. 8.24, 3.36, 1.83

(ii) Count the number of observations in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$. Express each as a percentage of the total number of observations.

(iii) Compare the percentages found in part (ii) to the percentages given by the data.

Application

19. The marks obtained by six students in a class are 8, 6, 7, 9, 5, 7. Find mean, variance and standard deviation of the marks obtained by the students using appropriate notations. Ans. 7, Approx. 1.67 and 1.29

20. The earnings per share (in Tk.s) for 12 companies selected at random from a list of 500 companies are:

2.35	1.42	8.05	6.71
3.11	2.56	0.72	4.17
5.33	7.74	3.88	6.21

Find mean and standard deviation. Ans. Tk.4.35, Tk.2.45

21. The reaction time (in minutes) of a drug given to a random sample of 12 patients are:

4.9	5.1	3.9	4.2
6.4	3.4	5.8	6.1
5.0	5.6	5.8	4.6

Find mean and standard deviation.

Ans. 5.1 min; 0.9 min.

22. The lives (in hours) of 100 randomly selected flashlight batteries are:

Class interval	Frequency
6.95-7.45	2
7.45-7.95	10
7.95-8.45	23
8.45-8.95	30
8.95-9.45	21
9.45-9.95	13
9.95-10.45	1

Find mean and standard.

Ans. 8.7hr.; 0.6hr

23. The following data refer to the sales in thousand takas of 25 days of a departmental store:

Sales (in thousand Tk)	Number of days
10-20	3
20-30	6
30-40	11
40-50	3
50-60	2

Find mean and mean deviation of store.

Ans. 33, 8.16

24. The following frequency distribution refers to the daily wages of 200 workers randomly selected from 2000 workers of a factory:

Daily wages (in Tk.s)	Number of workers
Less than 35	14
35-38	62
38-41	99
41-44	18
Over 44	7
44	

Compute appropriate relative measure of dispersion. Coeff. Of Q.D.= 4.4%.

25. The frequency distribution given below refers to the daily wages of 100 workers randomly selected from a factory:

Daily Wages in Tk.	Number of workers	Daily Wages in Tk.	Number of workers
210-215	8	230-235	14
215-220	13	235-240	10
220-225	16	240-245	7
225-230	29	245-250	3

- Find mean, median and standard deviation of the daily wages of the workers.
 Ans. Tk.227.55, Tk.229.74 and Tk.8.73.
26. The frequency distribution refers to the profits of 50 companies randomly selected from a country

Profits (in crores Tk.)	Number of companies
10-20	8
20-30	12
30-40	20
40-50	6
50-60	4

Find mean, standard deviation and coefficient of variation.

Ans. 32.2; 11.14; 34.6%.

27. Particulars regarding the weekly wages of workers of two factories A and B:

	Factory A	Factory B
Number of workers	600	500
Average wages (in Tk.)	1750	1850
Standard Deviation (in Tk.)	100	81

- (i) Find coefficient of variation of the weekly wages of the two factories.
 (ii) Which factory has better salary structure?

Ans. (i) $C.V(A)=5.71\%$ and $C.V.(B) = 4.355$ (ii) The factory B has less C.V., so the salary structure of B is better than A.

28. The following frequency distribution refers to the amount of annual income tax in thousand taka paid by the manager of different firms:

Annual tax paid	Number of managers
5-10	18
10-15	30
15-20	46
20-25	28
25-30	20
30-35	12
35-40	6

Compute (i) Coefficient of variation; (ii) Inter-quartile range; (iii) Modal value.
 Ans. (i) 44.35%; (ii) 10.97; (iii) 17.35

29. For two firms A and B belonging to same industry, the following details are available:

	Firm A	Firm B
Number of Employees:	100	200
Average wage per day	Tk.240	Tk.170
Standard deviation of the wage per day	Tk 6	Tk 8

Find (i) which firm pays out larger amount as daily wages? (ii) Which firm shows greater variability in the distribution of wages? (iii) Which firm has the better daily wages structure?

Ans. (i) B pays larger amount; (ii) B; (iii) A.

30. The frequency distribution given below gives the number of hours worked per month by randomly selected 360 workers of a factory:

Number of hours worked per month	Number of workers
120-130	31
131-135	44
136-140	48
141-145	51
146-150	60
151-155	55
156-160	43
161-165	28

Find (i) Arithmetic mean; (ii) Coefficient of variation; and Quartile deviation.

Ans, (i) 145.53, (ii) 7.074%; (iii) 8.355

31. Now days consumers become more careful about the foods they eat, food processors try to stay competitive by avoiding excessive amounts of fat, cholesterol. And sodium in the foods they sell. The following data are the amounts of sodium per slice (in milligrams) for each of eight brands of regular cheese available in the market. Construct a box plot for the data and look for outliers.

Ans. $Q_1 = 292.5$; $Q_2 = 325$; $Q_3 = 240$; Inner fence = 221.25, 411.25;
Outer fence = 150, 482.50; $x = 520$ is the outlier.

32. The following data give the summary measures on height and weight of 100 students of a class:

Average weight of all the students = $\mu_w = 65\text{kg}$

Standard deviation of the weight of all the student = $\sigma_w = 9\text{kg}$

Average Height of all the students = $\mu_h = 66$ inches

Standard deviation of the of all the student = $\sigma_h = 5$ inches

Average weight of boy students = $\bar{x}_w = 65\text{kg}$

Standard deviation of the weight of boy student = $s_w = 9\text{kg}$

Average weight of girl students = $\bar{x}_w = 50\text{kg}$

Standard deviation of the weight of girl student = $s_{gw} = 7\text{kg}$

- (i) Compare the variability in weight between the boy and girl.
- (ii) Compare the variability of weight and height of the students.
- (iii) Compare the variability of weight between the all students with the girl students.

33. The data listed here are the weights (in pounds) of 27 packages of ground beef in a super market display:

1.08, 0.99, 0.97, 1.18, 1.41, 1.28, 0.83, 1.06, 1.14, 1.38,
0.75, 0.96, 1.08, 0.87, 0.89, 0.89, 0.96, 1.12, 1.12, 0.93,
1.24, 0.89, 0.98, 1.14, 0.92, 1.18, 1.17

- i) Construct a stem and leaf plot. Is the distribution relatively mound-shaped?

- ii) Find mean and standard deviation of the data set.
 iii) Find the percentages of observations in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$.
 iv) How the percentages do obtained in part (iii) compare with those given by the (a) Empirical Rule and (b) Chebyshev's Rule? Explain
 34. The following table gives the length of life of electric bulbs:

Life in Hours	Number of Bulbs	Life in Hours	Number of Bulbs
0-300	3	1500-1800	52
300-600	6	1800-2100	32
600-900	19	2100-2400	23
900-1200	30	2400-2700	7
1200-1500	45	2700-3000	2

Compute men, variance, standard deviation and coefficient of variation.

$$\text{Ans. } \bar{x} = 1507.53, \sigma^2 = 285156, \sigma = 534, \text{C.V.} = 35.35\%$$

CHAPTER - 7

MOMENTS OF A DISTRIBUTION

7.1. Introduction

Moments are popularly used to describe the characteristics of a distribution. According to Karl Pearson, first four moments are sufficient to describe a distribution. There are two kinds of moments. They are known as

- i) Raw moments and
- ii) Central or corrected moments.

But moments are calculated in three different ways. They are

- i) Raw moments about origin,
- ii) Raw moments about any arbitrary value and
- iii) Moments about mean or corrected moments.

Moments about mean are also known as central moments. Moments about mean are important. They are used to describe the shape characteristics of a distribution. Actually, raw moments are used to find the corrected moments.

7.1.1. Moments from ungrouped data.

Definition. Let x_1, x_2, \dots, x_n be n values of a variable, then r th moment about origin, denoted by v_r , is defined as

$$v_r = \frac{\sum x^r}{n}$$

We get first, second, third and fourth raw moments about origin by putting $r = 1, 2, 3, 4$ in the formula of v_r . That is,

$$\text{First raw moment} = v_1 = \frac{\sum x}{n} = \bar{x} = \text{sample mean},$$

$$\text{Second raw moment} = v_2 = \frac{\sum x^2}{n},$$

$$\text{Third raw moment} = v_3 = \frac{\sum x^3}{n},$$

$$\text{Fourth raw moment} = v_4 = \frac{\sum x^4}{n}.$$

Again, the r th raw moment of a variable X about any arbitrary value A , denoted by μ'_r , is defined as

$$\mu'_r = \frac{\sum(x - A)^r}{n}$$

We get the first, second, third and fourth raw moments about arbitrary value A by putting $r = 1, 2, 3, 4$ in the formula of μ'_r . That is,

$$\mu'_1 = \frac{\sum(x - A)}{n}, \quad \mu'_2 = \frac{\sum(x - A)^2}{n}, \quad \mu'_3 = \frac{\sum(x - A)^3}{n}, \quad \mu'_4 = \frac{\sum(x - A)^4}{n}.$$

The rth corrected or central moment of the variable X, denoted by μ_r is defined as

$$\mu_r = \frac{\sum(x - \bar{x})^r}{n}$$

First, second, third and fourth central moments can be obtained by putting $r = 1, 2, 3, 4$ in the formula of μ_r , such as,

$$\mu_1 = \frac{\sum(x - \bar{x})}{n} = 0, \quad \mu_2 = \frac{\sum(x - \bar{x})^2}{n} = s^2 = \text{sample variance},$$

$$\mu_3 = \frac{\sum(x - \bar{x})^3}{n}, \quad \mu_4 = \frac{\sum(x - \bar{x})^4}{n}.$$

It is noted that first raw moment about origin is the arithmetic mean and the second corrected moment is the sample variance. Moments can also be defined for the grouped data.

7.1.2. Moments from the grouped data. Suppose x_1, x_2, \dots, x_k are k values of a variable or k mid-points of k classes with frequencies f_1, f_2, \dots, f_k respectively, then the rth raw moments about origin is defined by

$$v_r = \frac{\sum fx^r}{n}; \quad r = 1, 2, 3, 4$$

where $n = \sum f$.

Then the first, second, third and fourth raw moments about origin are

$$v_1 = \frac{\sum fx}{n}, \quad v_2 = \frac{\sum fx^2}{n}, \quad v_3 = \frac{\sum fx^3}{n}, \quad v_4 = \frac{\sum fx^4}{n}.$$

The rth raw moment about any arbitrary value A is defined

$$\mu'_r = \frac{\sum f(x - A)^r}{n}; \quad r = 1, 2, 3, 4$$

The first, second, third and fourth raw moments about A are

$$\mu'_1 = \frac{\sum f(x-A)}{n}, \quad \mu'_2 = \frac{\sum f(x-A)^2}{n}, \quad \mu'_3 = \frac{\sum f(x-A)^3}{n}, \quad \mu'_4 = \frac{\sum f(x-A)^4}{n}$$

Sometimes, raw moments about A are calculated by the following short-cut method.

Let d is a new variable defined by $d = \frac{x-A}{i}$

Then the first four raw moments are calculated by the following formulae:

$$\mu'_1 = \frac{\sum fd}{n} \times i, \quad \mu'_2 = \frac{\sum fd^2}{n} \times i^2, \quad \mu'_3 = \frac{\sum fd^3}{n} \times i^3, \quad \mu'_4 = \frac{\sum fd^4}{n} \times i^4.$$

The rth central or corrected moment is defined by

$$\mu_r = \frac{\sum f(x-\bar{x})^r}{n}; \quad r = 1, 2, 3, 4$$

The corrected first, second, third and fourth moments are

$$\mu_1 = \frac{\sum f(x-\bar{x})}{n} = 0, \quad \mu_2 = \frac{\sum f(x-\bar{x})^2}{n} = s^2 = \text{sample variance},$$

$$\mu_3 = \frac{\sum f(x-\bar{x})^3}{n}, \quad \mu_4 = \frac{\sum f(x-\bar{x})^4}{n}.$$

There is a relationship between raw moments and corrected moments and raw moments are used to compute the corrected moments using their relationship. The relationship between corrected moments and raw moments are as follows:

$$\mu_1 = 0, \quad \mu_2 = \mu'_2 - (\mu'_1)^2 = v_2 - v_1^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^2 = v_3 - 3v_2v_1 + 2v_1^3$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4 = v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4$$

7.1.3. Sheppard's correction for moments. In case of grouped frequency distribution, while calculating moments we assume that the frequencies are concentrated at the middle point of the class intervals. If the distribution is symmetrical or slightly symmetrical and the class intervals are not greater than one-twentieth of the range, this assumption is very nearly true. But since the assumption is not in general true, some error, called the "grouping error", creeps into the calculation of the moments. W.F. Sheppard proved that if

- i) The frequency distribution is continuous, and
- ii) The frequency tapers off to zero in both directions.

then, the effect due to grouping at the mid-point of the intervals can be corrected by the following formulae, known as Shepard's corrections:

$$\mu_4(\text{corrected}) = \mu_4 - \frac{i^2}{12} + \frac{7}{240} i^4.$$

Here, i is the width of the class interval.

Note that the moments μ_1 and μ_3 need no correction, because under above mentioned situations, the numerical value of these moments are always zero.

It is to be noted that the first moment about origin v_1 is the sample mean and the second central moment is the sample variance except the divisor in the denominator is n in place of $n - 1$.

Example 7.1.1. Find the first four central moments of the following frequency distribution:

Class interval	1.75-2.25	2.25-2.75	2.75-3.25	3.25-3.75	3.75-4.25	4.25-4.75
Frequency	5	38	65	92	70	40

Solution.

Class interval	Mid-point : x	Frequency : f	$d = \frac{x - 3.5}{0.5}$	fd	fd^2	fd^3	fd^4
1.75-2.25	2.00	5	-3	-15	45	-135	405
2.25-2.75	2.50	38	-2	-76	152	-304	608
2.75-3.25	3.00	65	-1	-65	65	-65	65
3.25-3.75	3.50	92	0	0	0	0	0
3.75-4.25	4.00	70	1	70	70	70	70
4.25-4.75	4.50	40	2	80	160	320	640
Total		310		-6	492	-144	1788

$$\mu'_1 = \frac{\sum fd}{n} \times i = \frac{-6}{310} \times (0.5) = -0.01$$

$$\mu'_2 = \frac{\sum fd^2}{n} \times i^2 = \frac{492}{310} \times (0.5)^2 = 1.587 \times 0.25 = 0.397$$

$$\mu'_3 = \frac{\sum fd^3}{n} \times i^3 = \frac{-144}{310} \times (0.5)^3 = -0.368 \times 0.125 = -0.046$$

$$\mu'_4 = \frac{\sum fd^4}{n} \times i^4 = \frac{1788}{310} \times (0.5)^4 = 5.768 \times 0.0625 = 0.36$$

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 0.397 - 0.0001 = 0.3969 = 0.40$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^2 = -0.046 - 3(0.397)(-0.01) + 2(-0.01)^3 = -0.3$$

$$\begin{aligned}\mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4 \\ &= 0.36 - 4(-0.046)(-0.01) + 6(0.397)(-0.01)^2 - 3(-0.01)^4 = 0.358.\end{aligned}$$

7.2. Shape Characteristics of a Distribution

The shape characteristics of a set of data describe the manner in which the data are distributed. For example, two distributions may have the same mean and standard deviation but may differ in their shape. The shape characteristics of a distribution are measured by skewness and kurtosis. The first four central moments discussed in the previous section are used to measure the skewness and kurtosis. Skewness and kurtosis are the third and fourth characteristics of a distribution.

7.2.1. Skewness. Skewness means the lack of symmetry of a distribution. That is, when a distribution is not symmetrical, it is called skewed distribution. A distribution may be symmetrical, positively skewed or negatively skewed.

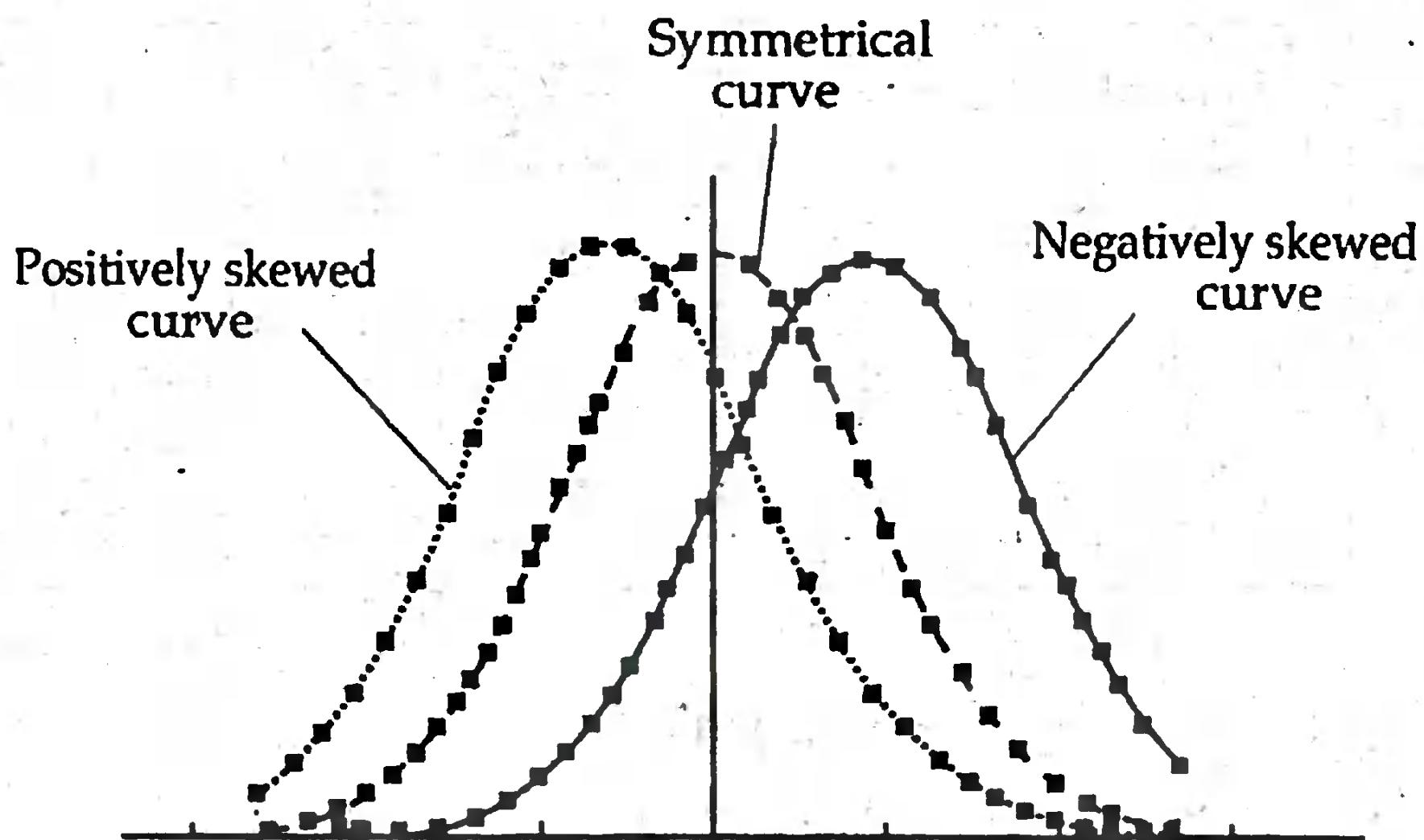


Fig. 7.1. Different types of skewed Distribution.

7.2.1.1. Symmetrical or zero-skewed distribution. A distribution is symmetric if the left and right sides of the distribution, when divided at the middle value, form mirror image. In other words, if this curve is folded at the center, two sides will coincide. In this case the values of mean, median and mode of the distribution are equal. Normal distribution is an example of a symmetrical distribution. It is also called zero-skewed distribution. The distribution of height, weight of adult people may follow this distribution. The diagram of a symmetric distribution will take the following form.

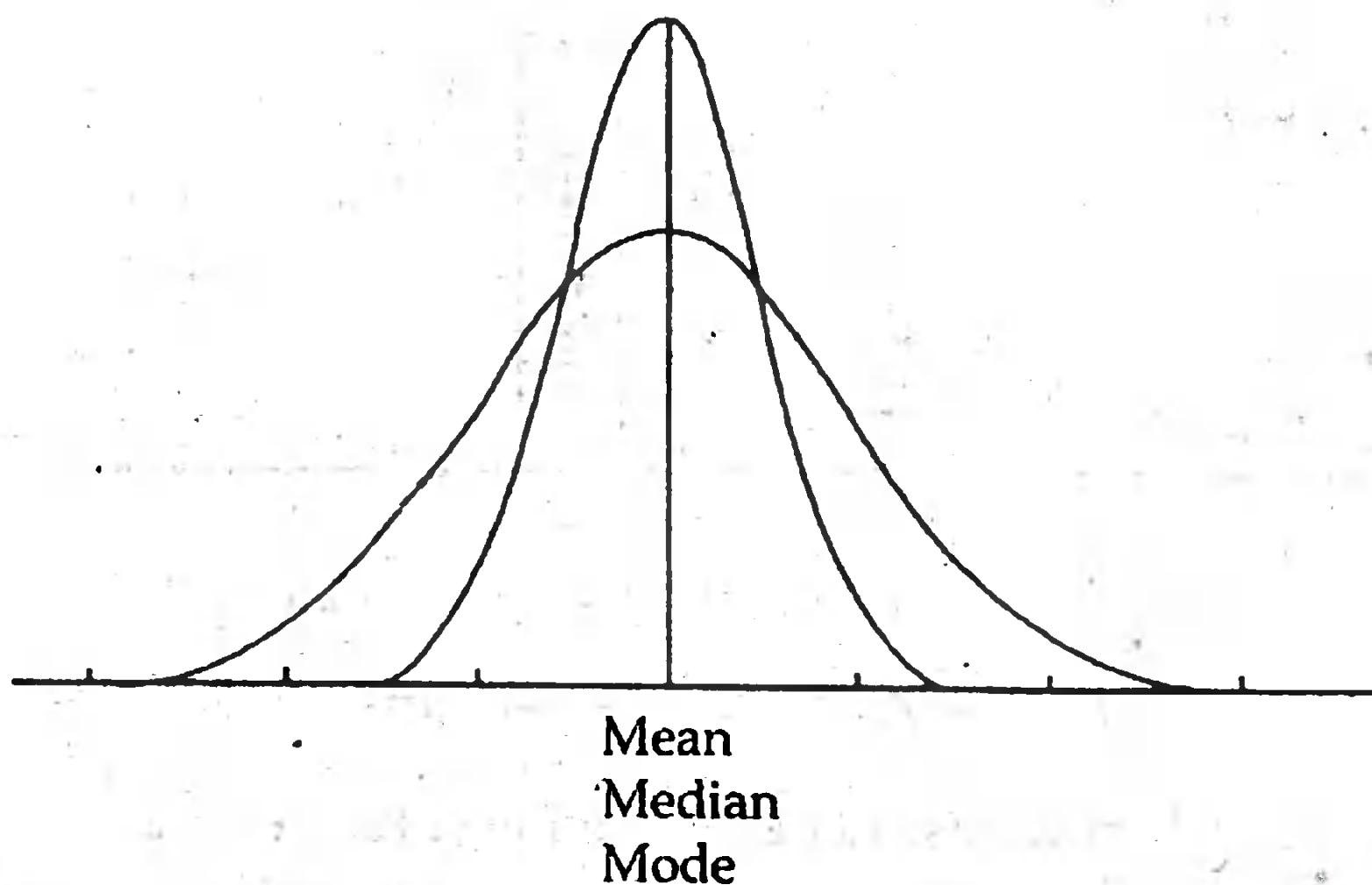


Fig. 7.2. Symmetrical curve.

7.2.1.2. Positively skewed distribution. A distribution is called positively skewed if a greater proportion of the observations lie to the right of the peak value. In this case the value of mean is greater than the value of mode and the value of median usually lies between them. Family size, female age at marriage, wage of the employees etc usually follow this type distribution. A diagram of a positively skewed is given below:

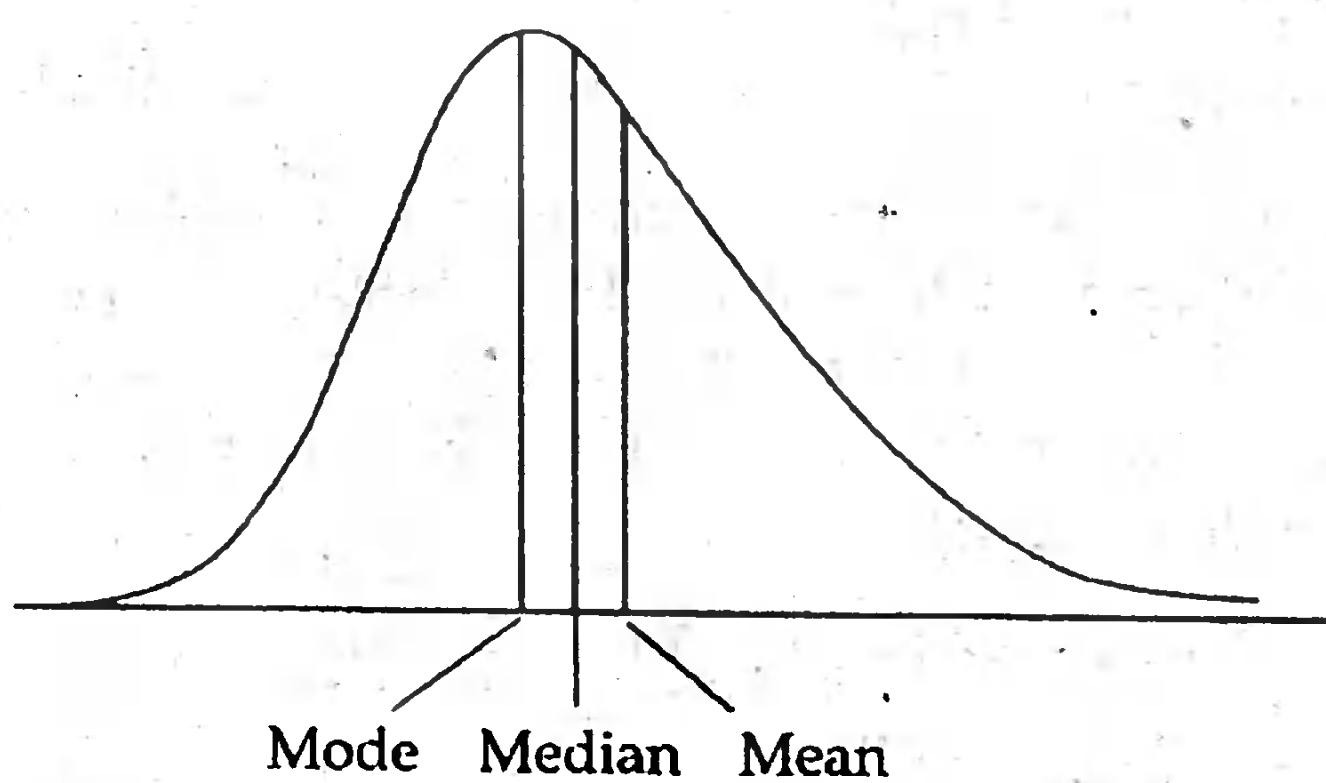


Fig. 7.3. A positively skewed curve.

7.2.1.3. Negatively skewed distribution. A distribution is called negatively skewed if a greater proportion of the observations lie to the left of the peak value. In this case the value of mode is greater than mean and the value of median lies between them. Reaction times for an experiment, daily maximum temperature for winter months etc follow this type of distribution. A diagram of a negatively skewed curve is given below:

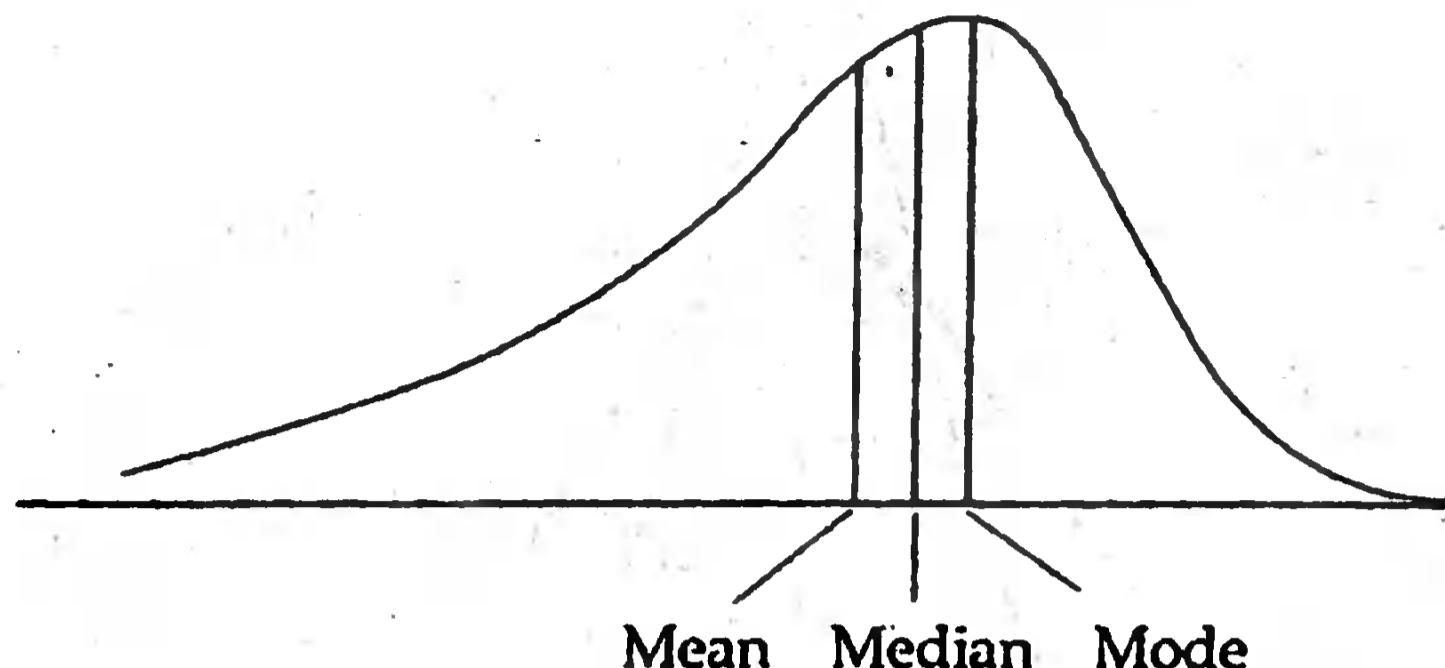


Fig. 7.4. A negatively skewed curve.

7.2.2. Some important measures of Skewness. Karl Pearson suggested a measure of skewness, based on the difference between the mean and mode. The formula is as follows:

Karl Pearson's coefficient of skewness =

$$PCS = \frac{\text{Mean} - \text{mode}}{\text{Standard deviation}} = \frac{\bar{x} - \text{mode}}{s}, \text{ for sample and}$$

$$PCS = \frac{\text{Mean} - \text{mode}}{\text{Standard deviation}} = \frac{\mu - \text{mode}}{\sigma}, \text{ for population.}$$

Here μ , \bar{x} , σ and s are the population mean, sample mean, population standard deviation and sample standard deviation respectively..

Since mode is not always obtained in routine calculations, a different formula is used for computation of the PCS. Assuming that the median is one-third of the way from the mean to the mode (for a moderately skewed distribution), the computation formula for Karl Pearson's coefficient of skewness for sample is given by

$$PCS = \frac{3(\text{Mean} - \text{median})}{\text{Standard deviation}} = \frac{3(\bar{x} - \text{median})}{s}, \text{ and}$$

for population

$$PCS = \frac{3(\text{Mean} - \text{median})}{\text{Standard deviation}} = \frac{3(\mu - \text{median})}{\sigma}.$$

Generally, its value lies between -1 to $+1$. The coefficient of skewness is zero for a symmetric curve. Its value is positive for positively skewed curve and negative for negatively skewed curve.

In case the standard deviation and mean of any distribution cannot be worked out, as is the case in open-end distribution and where extreme values are present, the coefficient of skewness, as suggested by Bowley, can be used. Bowley's coefficient of skewness (BCS) is defined as follows:

$$BCS = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

This measure is called as Quartile Measure of Skewness, and it also varies from -1 to +1.

Another measure of skewness devised by Kelly is based on deciles and percentiles.

Thus, Kelly's Coefficient of Skewness (KCS) based on deciles is computed by the following formula:

$$KCS = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1}$$

Kelly's Coefficient of Skewness (KCS) based on percentiles is computed by the following formula:

$$KCS = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

It is clear from the above two formulae that to calculate the Kelly's coefficient of skewness, we have to determine the value of 9th, 5th, 1st deciles and 90th, 50th, and 10th percentiles. However these methods are not so popular in practice.

Another important relative measure of skewness based on moments is given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

For a symmetrical distribution β_1 shall be zero. However, this coefficient β_1 as measure of skewness has a serious limitation. β_1 is always positive. So it cannot tell us the actual direction of skewness i.e., whether it is positive or negative. To remove this drawback Karl Pearson suggested γ_1 to be used as a measure of skewness as follows:

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\mu_2^{3/2}}$$

The sign of skewness would depend upon the value of μ_3 . If μ_3 is positive we will have positive skewness and if μ_3 is negative, we will have negative skewness.

Example 7.2.1. The arithmetic mean, mode and standard deviation of a distribution are 37.70, 36.67 and 8.29. Compute skewness and comment on the distribution.

Solution. Karl Pearson's coefficient of skewness is

$$PCS = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{37.25 - 36.67}{8.29} = 0.07.$$

Comment. The value of skewness is only 0.07. Hence the distribution is slightly positively skewed.

Example 7.2.2. The mean and median wages per day of a worker of an industry are Tk.157 and Tk.160. Suppose the standard deviation of wages is Tk.50. Calculate the coefficient skewness and comment.

Solution. Here we have mean = 157 and median = 160 and standard deviation = 50.

The formula for finding coefficient of skewness is

$$Sk_k = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(157 - 160)}{50} = -0.18$$

Comment. The negative value of skewness indicates the distribution is negatively skewed.

Example 7.2.3. The first and third quartiles of a distribution are 10 and 25. The median of the distribution is 20. Find its coefficient skewness.

Solution. According to definition, the Bowly's coefficient of skewness is

$$BCS = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1} = \frac{25 + 10 - 2 \times 20}{25 - 10} = \frac{35 - 40}{15} = -0.33$$

The distribution is negatively skewed.

7.2.3. Kurtosis. The concentration of observations about the central values is not same for all the symmetrical distributions. They may differ markedly in terms of peakedness what we call kurtosis.

Kurtosis is defined as the degree of flatness or peakedness of a distribution relative to a normal distribution.

Measures of kurtosis: Kurtosis is measured by the coefficient β_2 or its derivation γ_2 given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ and } \gamma_2 = \beta_2 - 3.$$

Here it is noted that β_2 is a pure number and positive.

There are three types of kurtosis. They are (i) Mesokurtic, (ii) Leptokurtic and (iii) Platykurtic.

Mesokurtic or normal curve. The curve, which is neither flat nor peaked, is called normal curve or mesokurtic curve. In this curve $\beta_2 = 3$, i.e. $\gamma_2 = 0$.

Leptokurtic curve. If a curve is more peaked than the normal curve, it is called leptokurtic. In this case $\beta_2 > 3$ i.e. $\gamma_2 > 0$.

Platykurtic curve. When a curve is less peaked than the normal curve, it is called platykurtic. For this type of curve $\beta_2 < 3$ or $\gamma_2 < 0$. The diagram below illustrates the three different types of curves mentioned above.

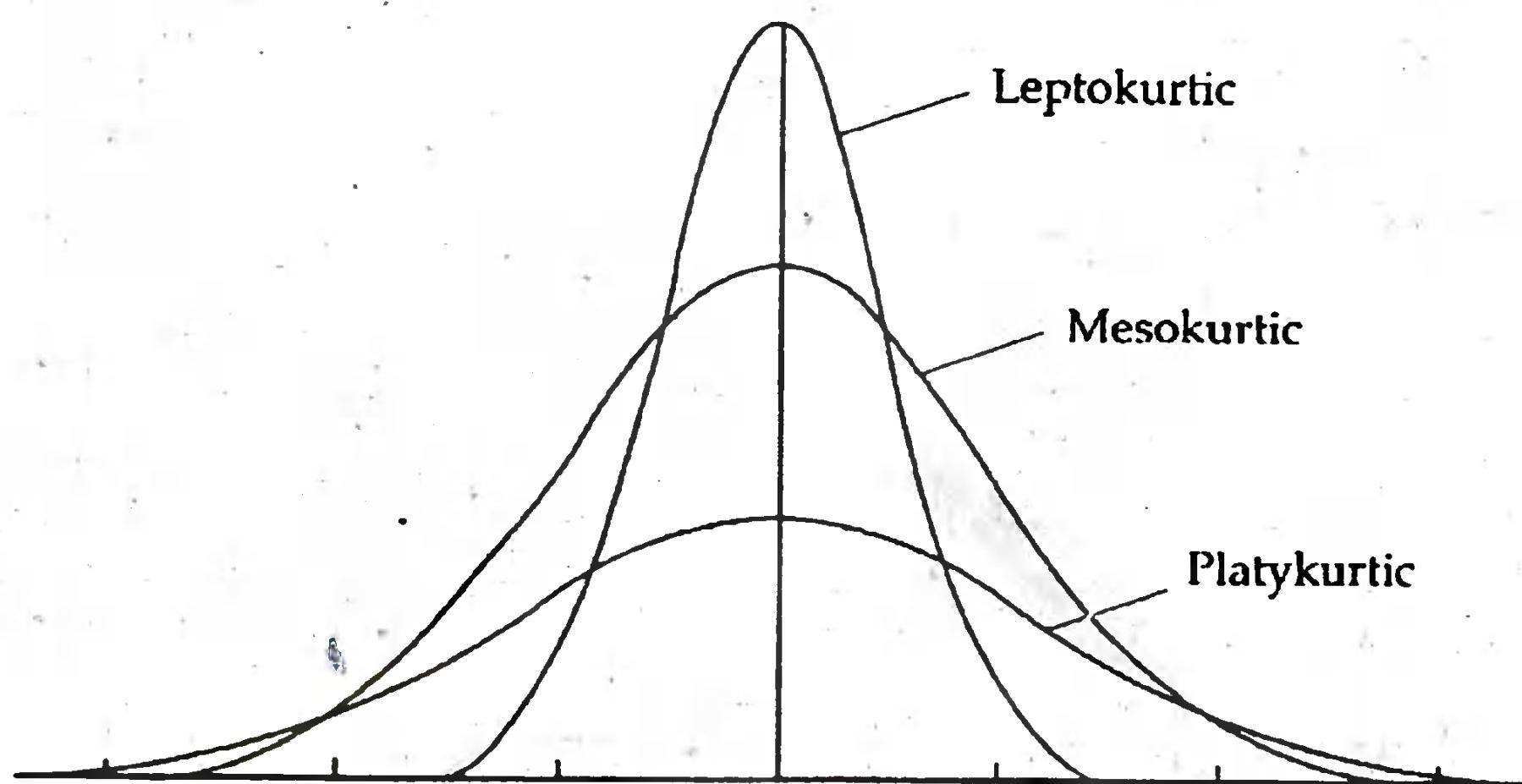


Fig. 7.5. Three types of kurtosis.

It may be noted that it is easier to interpret kurtosis by calculating β_2 instead of γ_2 . It is to be noted that the concept of kurtosis is rarely used in analyzing business data.

Example 7.2.4. The first four central moments of a distribution are 0, 16, -36 and 120. Comment on the shape characteristics of the distribution.

Solution. Skewness and Kurtosis are the shape characteristics of a distribution. They are measured by β_1 and β_2 . Since β_1 is always positive, we measure skewness by γ_1 which is computed by the formula

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Here $\mu_3 = -36$ and $\sigma = 4$. Hence $\gamma_1 = \frac{-36}{(4)^3} = \frac{-36}{64} = -0.56$

Comment. The value of γ_1 is negative. Hence the distribution is negatively skewed.

$$\text{Again, } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{120}{(16)^2} = \frac{120}{256} = 0.47.$$

Comment. Since the value of β_2 is less than 3, the distribution is platykurtic.

Matched problem. The first four central moments are 0, 2.5, 0.7 and 18.75. Find β_1 and β_2 . Comment on the shape characteristics of the distribution.

$$\text{Ans. } \beta_1 = 0.031, \beta_2 = 3$$

Example 7.2.5. The first four moments about 5 of a distribution are 2, 20, 40 and 50. Find mean and standard deviation of the distribution. Comment on the shape characteristics of the curve.

Solution. The Shape characteristics of a distribution are measured by γ_1 and β_2 . By the following formulae

$$\gamma_1 = \frac{\mu_3}{\mu_2^2} = \frac{\mu_3}{\sigma^3} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

We are given $\mu'_1 = 2, \mu'_2 = 20, \mu'_3 = 40, \mu'_4 = 50$ and $A = 5$

$$\text{Mean} = \bar{x} = A + \mu'_1 = 5 + 2 = 7$$

$$\text{Variance} = \mu'_2 - (\mu'_1)^2 = 20 - (2)^2 = 20 - 4 = 16$$

$$\text{Standard deviation} = \sigma = \sqrt{16} = 4$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3 = 40 - 3(20)(2) + 2(2)^3 = 40 - 120 + 16 = -64$$

$$\begin{aligned} \mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4 = 50 - 4(40)(2) + 6(20)(2)^2 - 3(2)^4 \\ &= 50 - 320 + 480 - 48 = 162 \end{aligned}$$

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-64}{(16)^{3/2}} = -1$$

$$\text{and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = 0.63 < 3$$

Comment. The distribution is negatively skewed and platykurtic.

Example 7.2.6. The first four moments of a distribution about the origin are 2.5, 21, 166, 1132 respectively. Find mean, variance, β_1, β_2 and comment.

'Here we are given first four moments about origin. That is

$$v_1 = 2.5, v_2 = 21, v_3 = 166 \text{ and } v_4 = 1132$$

$$\text{Mean} = v_1 = 2.5$$

$$\text{Variance} = \sigma^2 = \mu_2 = v_2' - (v_1)^2 = 21 - (2.5)^2 = 14.75$$

$$\begin{aligned}\mu_3 &= v_3 - 3v_2v_1 + 2(v_1)^2 \\ &= 166 - 3(21)(2.5) + 2(2.5)^2 = 166 - 157.5 + 31.25 = 39.75\end{aligned}$$

$$\begin{aligned}\mu_4 &= v_4 - 4v_3v_1 + 6v_2(v_1)^2 - 3(v_1)^4 \\ &= 1132 - 4(166)(2.5) + 6(21)(2.5)^2 - 3(2.5)^4 \\ &= 1132 - 1660 + 787.5 - 117.1875 = 142.312\end{aligned}$$

$$\beta_1 = \frac{\mu_2^3}{\mu_2^3} = \frac{(39.75)^2}{(14.75)^3} = \frac{1580.06}{3209.05} = 0.49$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{142.312}{(14.75)^2} = \frac{142.31}{217.56} = 0.65 < 3.$$

Since μ_3 is positive and β_1 is also positive. The curve is positively skewed. The value of β_2 is less than 3, so the curve is platykurtic.

Example 7.2.7. Suppose, the second, third and fourth central moments of a distribution are 19.28, 77.05, 1202.19 respectively. Find β_1 , β_2 and comment.

Solution. Given $\mu_2 = 19.28$, $\mu_3 = 77.05$, and $\mu_4 = 1202.19$

$$\text{So, } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(77.05)^2}{(19.28)^3} = 0.83$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{1202.19}{(19.28)^2} = 3.23 > 3$$

Comment. Since μ_3 is positive and β_1 is also positive. The curve is positively skewed. The value of β_2 is more than 3, so the curve is leptokurtic.

Matched Problem. The first four moments of a distribution about the origin are 1, 4, 10, and 46 respectively. Obtain mean, standard deviation μ_3 , μ_4 , β_1 and β_2 . Comment on the nature of the distribution.

Ans. mean = 1, $\sigma = 1.732$, $\mu_3 = 0$, $\mu_4 = 27$, $\beta_1 = 0$, $\beta_2 = 3$, normal distribution

The shape characteristics of distribution can also be shown by the five-number summary and Box and whisker plot

7.3. Exploratory Data Analysis

The techniques of exploratory data analysis (EDA) consist of simple arithmetic and easy-to-draw graphs or pictures to summarize data quickly. It was named by John Tukey. The principal graphical techniques used in EDA are

1. Five-number summary
2. Box-and-whisker plot
3. Scatter plot
4. Histogram
5. Pareto chart
6. Run chart
7. Stem-and-leaf plot
8. Multivariate chart

Some of these techniques were discussed before. In this section, here, we will first discuss the five-number summary and then Box and Whisker plot.

7.3.1. The Five-Number Summary. In a five-number summary, the following five numbers are used to summarize the data :

1. Smallest value (X_{smallest})
2. First quartile (Q_1)
3. Median (Q_2)
4. Third quartile (Q_3)
5. Largest value (X_{largest})

In ascending order, these five numbers can be written as

$X_{\text{smallest}} \quad Q_1 \quad Q_2 \quad Q_3 \quad X_{\text{largest}}$

These five numbers provide a way to determine the shape of the distribution.

If the data are perfectly symmetrical, the relationship among the various measures in the five-number summary can be exhibited as follows:

1. The distance from X_{smallest} to the median equals the distance from the median to X_{largest} .
2. The distance from X_{smallest} to Q_1 equals the distance from Q_3 to X_{largest} .

For asymmetrical distributions the relationship among the various measures of location can be exhibited as

1. In case of right-skewed distributions, the distance from the median to X_{largest} is greater than the distance from X_{smallest} to the median.
2. In case of right-skewed distributions, the distance from Q_3 to X_{largest} is greater than the distance from X_{smallest} to Q_1 .
3. In case of left-skewed distribution, the distance from X_{smallest} to the median is greater than the distance from the median to X_{largest} .

4. In case of left-skewed distribution, the distance from X_{smallest} to Q_1 is greater than the distance from Q_3 to X_{largest} .

Example 7.3.1. The monthly starting salaries in dollar for a random sample 12-business school graduates are as follows: 2890, 2930, 3425, 2860, 2960, 3060, 2880, 2765, 2720, 2900, 3260, and 2950. Find the five-number summary for the data and comment.

Solution. To find the five-number summary for the data set, we have to arrange the observations in ascending order of magnitude. The ordered array is 2720 2765 2860 2880 2890 2900 2930 2950 2960 3060 3260 3425. The smallest value, X_{smallest} is 2720 and the largest value X_{largest} 3425.

Here $n = 12$ is even. $n/2 = 12/2 = 6$. Hence the median is the mean of the 6th and 7th ordered observations, given by

$$\text{Median} = \frac{2900 + 2930}{2} = 2915.$$

Here $n/4 = 12/4 = 3$. Then first quartile is the mean of 3rd and 4th ordered observations, given by

$$Q_1 = \frac{2860 + 2880}{2} = 2870.$$

Here $3n/4 = 9$. Then third quartile Q_3 is the mean of the 9th and 10th ordered observations, given by

$$Q_3 = \frac{2960 + 3060}{2} = 3010.$$

The quartiles have divided the values into four parts, with each part consisting of 25% of the observations. The five-number summary values are shown below:

2720	2765	2860		2880	2890	2900		2930	2950	2960		3060	3260	3525
↑				Q ₁ =2870				Q ₂ =2915				Q ₃ =3010		↑
X_{smallest}														X_{largest}

Thus the five-number summary for the salary data is 2720, 2870, 2915, 3010, and 3525.

Here the distance from the median to X_{largest} is greater than the distance from X_{smallest} to the median. Hence the distribution is right-skewed.

Matched problem

Find the five-number summary for the following data set: 19, 12, 16, 0, 14, 9, 6, 1, 12, 13, 10, 19, 7, 5, 8.

$$\text{Ans. } X_{\text{smallest}} = 0, Q_1 = 6, Q_2 = 10, Q_3 = 14, X_{\text{largest}} = 19.$$

7.3.2. The box-and-whisker plot. A more sophisticated modification of the graphical five-number summary is the box-and-whisker plot. It is also sometime known as box plot. A box-and-whisker plot provides a graphical representation of the data based on the five-number summary. From box-and-whisker plot one can quickly detect any skewness in the shape of the distribution and can also observe whether there are any outliers in the data set.

To construct a box-and-whisker plot

1. Calculate median, the first quartile Q_1 , third quartile Q_3 and interquartile range IQR for the data set.
2. Draw a horizontal line representing the scale of observations. Form a box just above the horizontal line with the right and left ends at Q_3 and Q_1 . Draw a vertical line through the box at the location of the median.

A box-and-whisker plot is shown in Figure 7.3.1.

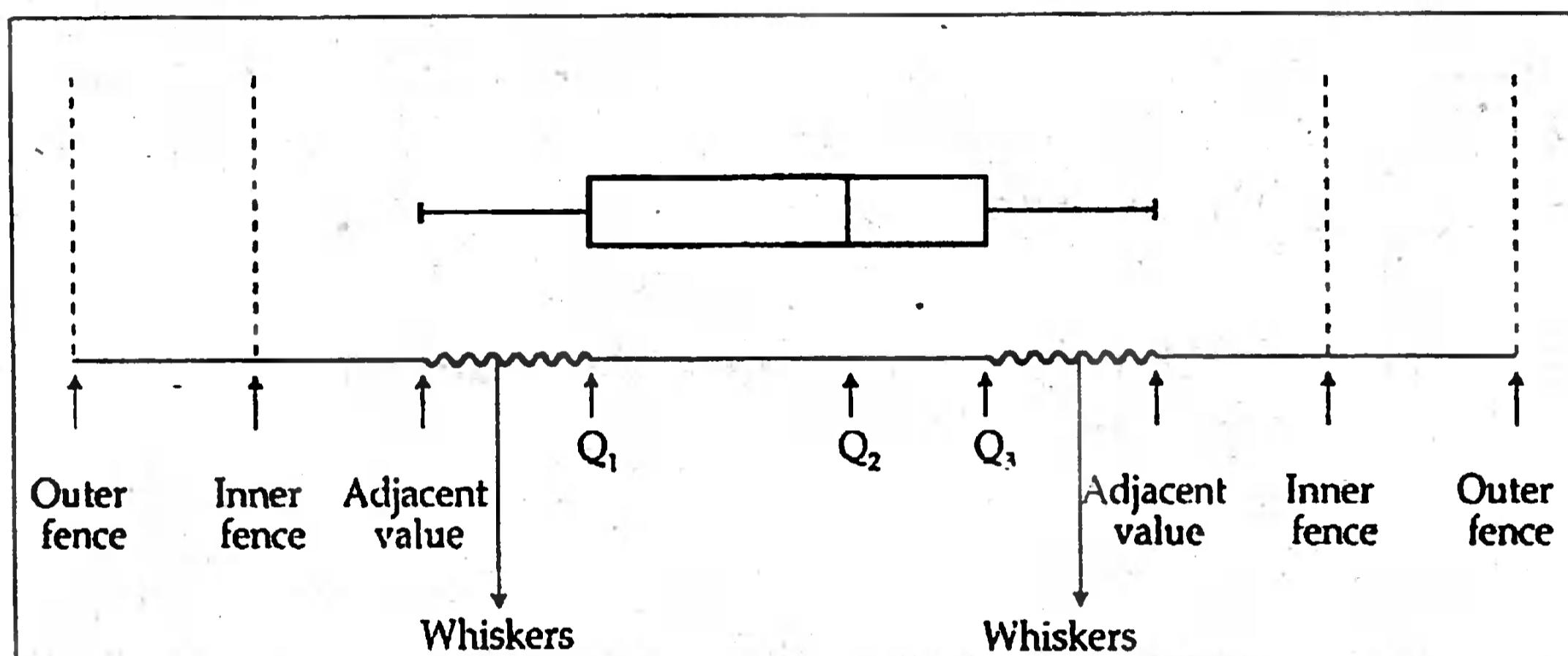


Fig. 7.3.1. A box-and-whisker plot.

3. The box plot uses the IQR to create imaginary fences to separate outliers from the rest of the data set.

$$\text{Inner fence: } Q_1 - 1.5(\text{IQR}) \quad \text{and} \quad Q_3 + 1.5(\text{IQR})$$

$$\text{Outer fence: } Q_1 - 3(\text{IQR}) \quad \text{and} \quad Q_3 + 3(\text{IQR})$$

The inner and outer fences are shown with the broken lines in Figure 6.1. Any observations that are between the inner and outer fences are called **suspect outliers**, and any observations beyond the outer fences are **extreme outliers**. The rest of the observations, inside the inner fences, are not unusual.

4. The largest and smallest values inside the inner fences are called adjacent values. Their positions are also shown in the Figure 6.1.

5. The lines that connect the ends of the box to the adjacent values are called whiskers.
6. Finally any suspect outliers are marked with an asterisk (*) on the graph, and extreme outliers are marked with a circle (o).

We can also describe the shape of the distribution of the data set by looking at the position of the median line compared to Q_1 and Q_3 , the left and right ends of the box. If the median is close to the middle of the box, the distribution is fairly symmetric. If the median line is to the left of center, the distribution is skewed to the right. If the median line is to the right of center, the distribution is skewed to the left. Also, for most skewed distributions, the whisker on the skewed side of the box tends to be longer than the whisker on the other side.

Example 7.3.2. The monthly starting salaries in dollar for a random sample 12-business school graduates are as follows: 2890, 2930, 3425, 2860, 2960, 3060, 2880, 2765, 2720, 2900, 3260, 2950. Construct Box and whisker plot for the data and identify outliers if any.

Solution. To construct box – and-whisker plot, first we have to find first quartile, median and third quartile and IQR. We arrange the data in ascending order of magnitude. The ordered observations are

2720 2765 2860 2880 2890 2900 2930 2950 2960 3060 3260 3425

Here $n = 12$ is even, and then median is the mean of the 6th and 7th ordered observations which are 2900 and 2930.

$$\text{Median} = \frac{2900 + 2930}{2} = 2915.$$

Here $n/4 = 12/4 = 3$. Then first quartile is the mean of 3rd and 4th ordered observations.

$$Q_1 = \frac{2860 + 2880}{2} = 2870$$

Here $3n/4 = 9$. Then third quartile Q_3 is the mean of the 9th and 10th ordered observations.

$$Q_3 = \frac{2960 + 3060}{2} = 3010$$

$$\text{IQR} = 3010 - 2870 = 140.$$

The quartiles have divided the values into four parts, with each part consisting of 25% of the observations.

$$2720 \ 2765 \ 2860 | 2880 \ 2890 \ 2900 | 2930 \ 2950 \ 2960 | 3060 \ 3260 \ 3525$$

$Q_1=2870 \quad Q_2=2915 \quad Q_3=3010$

Figure 7.3.2 is the box and whisker plot for the monthly starting salary data. The steps used to construct the box plot follow.

1. A box is drawn with the ends of the box located at the first quartile $Q_1 = 2870$ and third quartile $Q_3 = 3010$. The box contains the middle 50% of the data.
2. A vertical line is drawn in the box at the location of the median. Here it is 2915, which divides the data set into two equal parts.
3. Inner fences and outer fences are calculated by using the inter-quartile range, $IQR = 3010 - 2870 = 140$.

$$\text{Inner fence: } Q_1 - 1.5(IQR) = 2870 - 1.5(140) = 2870 - 210 = 2660$$

$$Q_3 + 1.5(IQR) = 3010 + 1.5(140) = 3010 + 210 = 3220$$

$$\text{Outer fence: } Q_1 - 3(IQR) = 2870 - 3(140) = 2870 - 420 = 2450$$

$$Q_3 + 3(IQR) = 3010 + 3(140) = 3010 + 420 = 3430$$

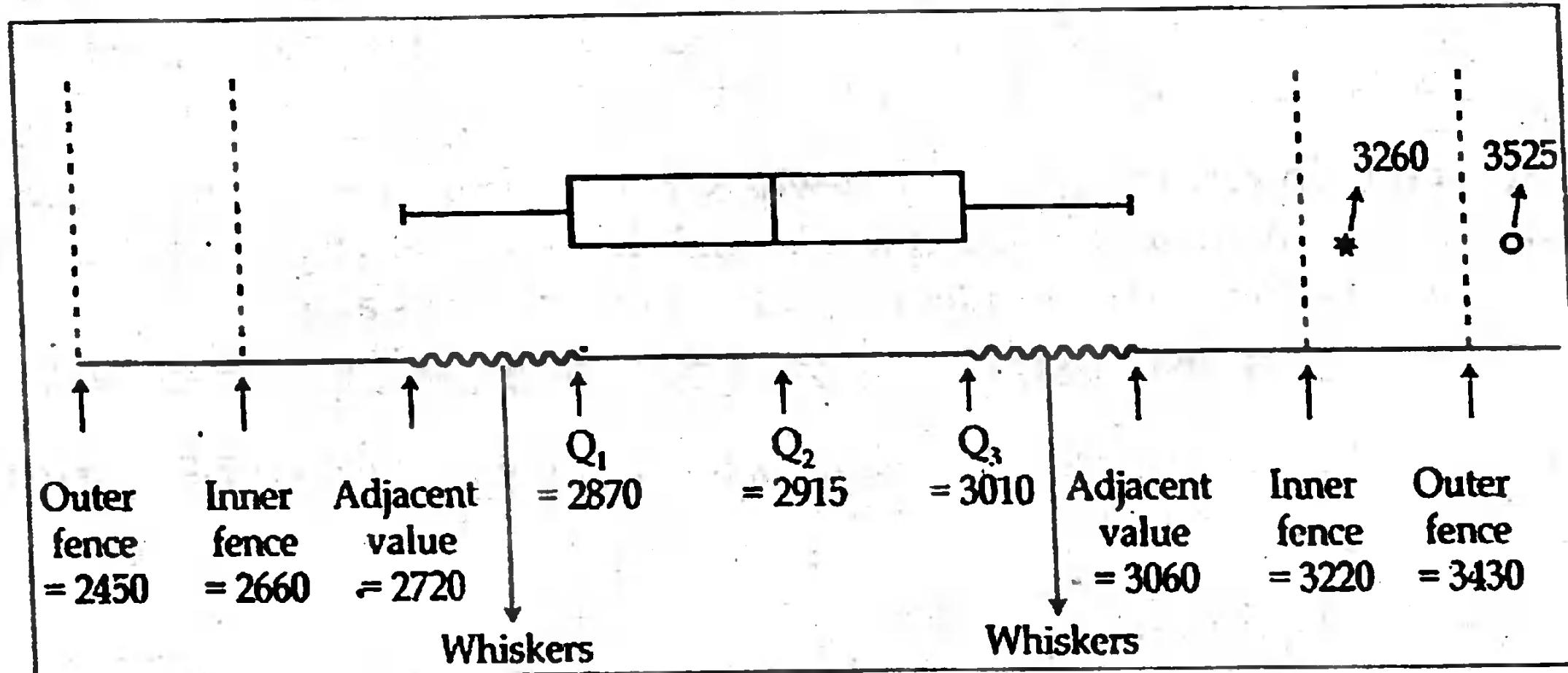


Fig. 7.3.2. A box-and-whisker plot.

The inner fences are located at 2660 and 3220 in the Figure 6.2 and are shown by two vertical lines. The outer fences are located at 2450 and 3430 and are shown by two vertical lines in the Figure 6.2. The observations lie between the inner and outer fences is considered as suspect outliers. The observation 3260 is suspected as outlier. The observations lie outside the outer fences is considered as extreme outliers. It is denoted by (*). Here the observation 3525 is considered as extreme outlier and it is denoted by (o).

4. The largest and smallest values inside the inner fences are called adjacent values. Here two adjacent values are located at 2720 and 3060.
5. The lines connect the ends of the box to the adjacent values are called whiskers. The two lines from $Q_1 = 2870$ to 2720 and $Q_3 = 3010$ to 3060 are called whiskers of the box plot.

It is seen from the box plot that the left side whisker is longer than the right side whisker. Hence the distribution is negatively skewed or left-skewed.

Comment. From the example 6.11.1 and 6.11.2, it is seen that the same data set gives different comment about the shape of the data set. This happens due to the fact that the data set contains two outliers.

Matched problem

Construct a box-and-whisker plot for these data and identify any outliers: 3, 9, 10 2, 6, 7, 5, 8, 6, 6, 4, 9, 22.

Ans. a. inner fence:-2.25 and 15.75; outer fence:-9 and 22.5; $x = 22$ is a suspect outlier

Questions

1. Define moments. Show how moments are used to describe the characteristics of a distribution.
2. Define raw and central moments of a frequency distribution. Express the second, third and fourth central moments in terms of raw moments.
3. Define skewness of a frequency distribution. Describe the different methods of measuring skewness of a distribution. Also interpret different values of skewness.
4. Discuss different types of skewness of a distribution with the help of diagrams.
5. Define kurtosis of a frequency distribution. How can you measure the kurtosis of a frequency distribution?
6. What do you mean by kurtosis? What are the different types of kurtosis? Show them with the help of a diagram.
7. What are the shape characteristics of a frequency distribution? Discuss them in brief.
8. "A frequency distribution can be described almost completely by the first central moments" Explain

Exercise

9. The first three moments of a distribution about the value 1 are 2, 25, and 80. Find its mean, three central moments, standard deviation and a measure of skewness.
Ans. Mean = 3, $\mu_1 = 0$, $\mu_2 = 21$, $\mu_3 = -54$, $\gamma_1 = -0.56$, $\gamma_2 = -0.56$.
10. The mean, median and standard deviation of a distribution are 46.4, 48.89 and 24.56 respectively. Find Karl Pearson's coefficient of skewness and comment. **Ans.** -0.101, low degree of negatively skewed distribution.

11. The mean, mode and standard deviation of a frequency distribution are 177.86, 190.56 and 25.2 Find the coefficient of skewness and comment on its value. Ans. -0.504; moderately negatively skewed distribution.
12. The first four central moments of a distribution are 0, 16, -64 and 162. Compute β_1 and β_2 , and comment on the nature of the distribution.

Ans. $\beta_1 = 0$ and $\beta_2 = -1$, negatively skewed and platykurtic.

13. Calculate the first four central moments and hence find β_1 and β_2 of the following discrete distribution:

x :	0	1	2	3	4	5	6	7	8
f :	1	8	28	56	70	56	28	8	1

Comment on your obtained values

Ans. 0, 2, 0, 11, 0, 2.75; The distribution is symmetric and platykurtic.

Applications

14. Particulars given below relating to the wage distribution of a manufacturing firm: Mean wage = Tk. 175, Modal wage = 167 and Standard Deviation = 13. Find a measure of skewness and comment.
15. The frequency distribution related to the sales of 80 randomly selected companies :

Sales in lakh Tk.	Number of companies
Below 50	8
50-60	20
60-70	40
70-80	65
80-90	80

Find a suitable measure of skewness.

Ans. Bowley's coefficient of skewness is the appropriate measure and its value is 0.111. It is case of low negatively skewed distribution.

16. The following table gives the length of life (in hours) of 400 T.V. tubers:

Length of life (in hours)	Number of Picture tubes
4000-4200	22
4200-4400	38
4400-4600	65
4600-4800	75
4800-5000	80
5000-5200	70
5200-5400	50

Compute mean, mode, standard deviation and coefficient of skewness.

Ans. 4781.5 hrs; 4866.67 hrs; 340.4 hrs; -0.25.

17. The frequency distribution refers to the profits of randomly selected 50 companies of a country:

Profits (in Lakhs Tk.)	Number of companies
70-90	8
90-110	11
110-130	18
130-150	9
150-170	4

Compute first four central moments, β_1 and β_2 .

Ans. $\mu_1 = 0$; $\mu_2 = 528$; $\mu_3 = 960$; $\mu_4 = 642816$; $\beta_1 = 0.006$ and $\beta_2 = 2.31$

Comment. Since the value of β_1 is near to zero and β_2 is less than 3, the distribution is almost symmetrical and platykurtic.

CHAPTER - 8

INTRODUCTION TO PROBABILITY

8.1. Introduction

The term probability as its origin relates with the games of chance in the seventeenth century. Girolamo Cardano (1501-1576), a gambler and physician, introduced some of the best mathematics of his time, including a systematic analysis of gambling problems. In 1654, another gambler, the Chevalier de Mere, plagued with bad luck, approached the well-known French philosopher and mathematician Blaise Pascal (1623-1662) regarding certain dice problems. Pascal became interested in these problems, studied them, and discussed them with Pierre de Fermat (1601-1665), another French mathematician. Thus, the study of probability was born out of the gaming rooms of Western Europe.

Even to-day there are many applications involving games of chance, such as various state lotteries, gambling casinos, race tracks and organized sports. Today the use of probability has gone beyond the games of chance. Nowadays government, business firms, different professional and non-profitable organizations are using probability theory into their everyday decision processes. Different insurance companies are earning a lot of money by using life table. Life table is a direct application of probability theory.

We are using the word probability in our everyday life. For example, what is the probability that a new product of a company will get good market? What is the probability that the price of Chittagong share market will rise in the next month? All these events are related with certain degree of uncertainty. In the age of free market economy, business decisions and economic analysis are mostly related with uncertainty. Probability is a numerical measure of uncertainty of such an event.

8.2. Set Theory

Before defining probability, we shall define some concepts of sets, which are useful for understanding probability theory. Set is the language of probability. The set theory was developed by the German mathematician Cantor [1845-1918].

Definition. Set. A set is well defined collection of objects which have some properties in common.

Definition. Point or element of a set. Each object of a set is called element or point or member of the set.

Example 8.2.1. Suppose A is a set with the first 8 natural numbers, and then A can be written as $A = \{ 1, 2, 3, 4, 5, 6, 7, 8 \}$

Here 1, 2, 3, 4, 5, 6, 7, 8 are the eight points or elements of the set A.

Number of elements of a set A is denoted by $N(A)$. Here $N(A) = 8$.

Now we list some more examples of set.

There are two ways to specify a particular set: (i) List method and (ii) Roster method.

Definition. List method. The method of defining a set is called list method if it is described by listing the all members separated by commas and enclosed by a bracelet { }.

Example 8.2.2. The set contains the numbers 1, 5, 9, 13 may be written as

$$A = \{ 1, 5, 9, 13 \}$$

Definition. Roster method. The method of defining a set is called roster method if it is described by a statement or rule.

Example 8.2.3. Suppose A is the set of English alphabets, we write $A = \{ x : x \text{ is an English alphabet} \}$

Whether we describe a set by list or roster method will depend on the specific problem. Suppose A is the set of all real numbers lying between 0 and 1, we write

$$A = \{ x : 0 \leq x \leq 1 \}$$

Now we list some more examples of set:

- (1) The set of odd numbers; $A = \{ 1, 3, 5, 7, \dots \}$
- (2) The set of the vowels of the English alphabet; $B = \{ a, e, i, o, u \}$
- (3) The set of the people living on the earth.
- (4) The set of the rivers of Bangladesh.
- (5) The set of the capital cities of America.

Subset. If every element of a set B is also an element of a set A, then B is known as a subset of A

Example 8.2.4. Suppose B is a set with the odd digits of A in example 8.2.1, and then B is $B = \{ 1, 3, 5, 7 \}$

Here B is a subset of A. Every set is a subset of itself.

Definition. Null Set. A set is called empty or null set if it contains no elements. It is denoted by ϕ . It is to be noted that the set { 0 } is not an empty set, since it contains 0 as its element. The empty set is a subset of every set.

Definition. Universal Set. Universal set is a set, which contains all objects of an experiment. Universal set is usually denoted by Ω or U or S . Universal set is also a subset of itself.

Example 8.2.5. If we throw a die, the universal set is the all possible outcomes which may be written as $\Omega = \{ 1, 2, 3, 4, 5, 6 \}$. Here 1, 2, 3, 4, 5 and 6 are the number of points on the faces of the die.

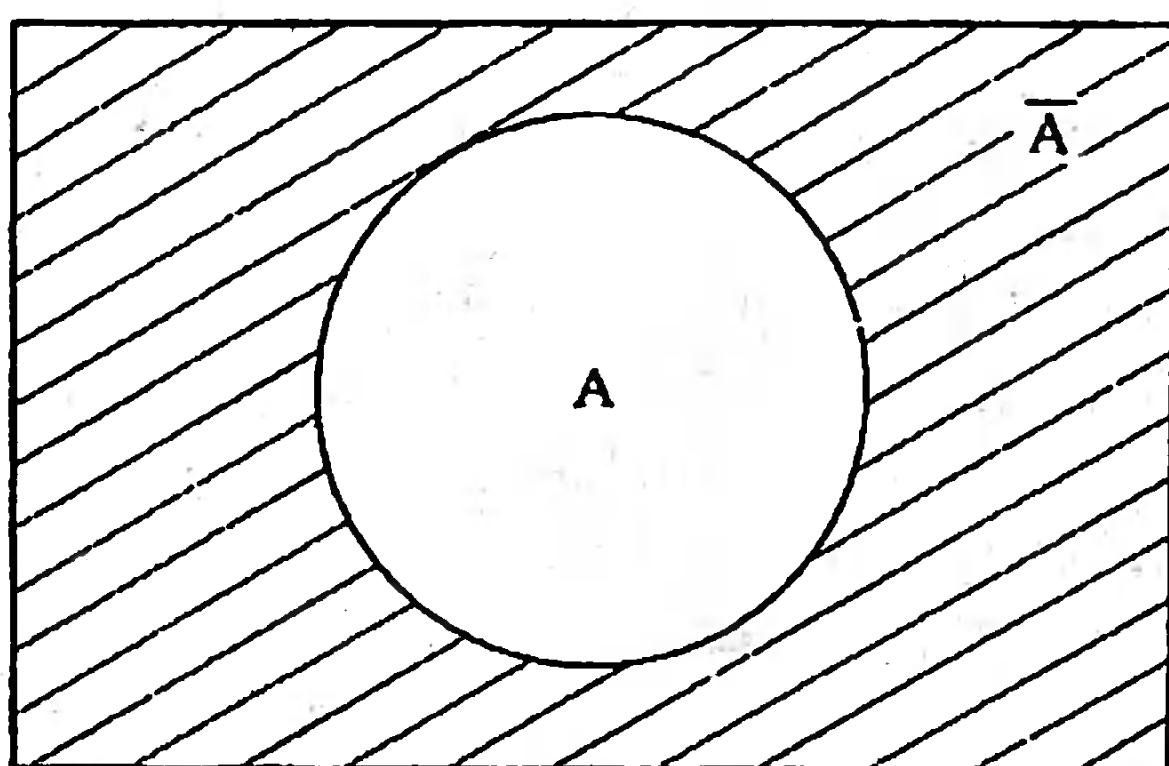
Number of subset from a set. If a set contains n elements, then the total number of subsets from A is 2^n .

Example 8.2.6. Let $A = \{ 1, 2 \}$. Find all possible subsets from A . The possible subsets from A are { 1 }, { 2 }, { 1, 2 } and ϕ .

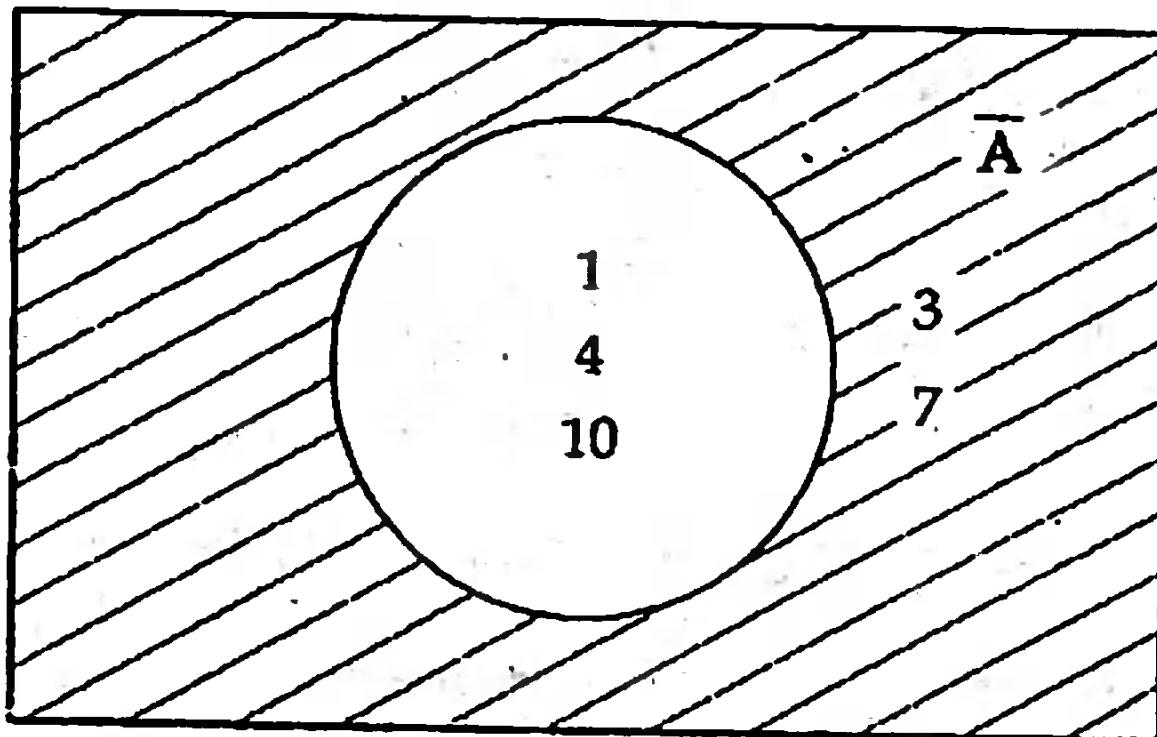
Example 8.2.7. Let $A = \{ 1, 3, 6 \}$, then the total number of subsets is $2^3 = 8$. The possible subsets are { 1 }, { 3 }, { 6 }, { 1, 3 }, { 1, 6 }, { 3, 6 }, { 1, 3, 6 } and ϕ .

Example 8.2.8. If we throw a die, the universal set is $\Omega = \{ 1, 2, 3, 4, 5, 6 \}$. The total number of subsets from Ω is $2^6 = 64$.

Complementary Set. If A is a subset of the universal set Ω , then the complement of A is the set of all elements that are in Ω but not in A . The complement of A is usually denoted by \bar{A} , A^c or A'



Example 8.2.9. Let $\Omega = \{ 1, 3, 4, 7, 10 \}$ and $A = \{ 1, 4, 10 \}$, then $\bar{A} = \{ 3, 7 \}$.



Union of two sets. If A and B are two subsets of the universal set Ω , then the union of A and B is a set which contains all the elements that are in A or B or both. Union of A and B is denoted by $A \cup B$.

Example 8.2.10. Let $A = \{1, 2, 3, 6\}$ and $B = \{2, 5, 6, 8\}$, then $A \cup B = \{1, 2, 3, 5, 6, 8\}$.

Intersection of two sets. If A and B are two subsets of the universal set Ω , then the intersection of A and B is a set that contains all the elements that are in both A and B . It is written as $A \cap B$ or AB . It is the set which contains all the common points of the sets A and B .

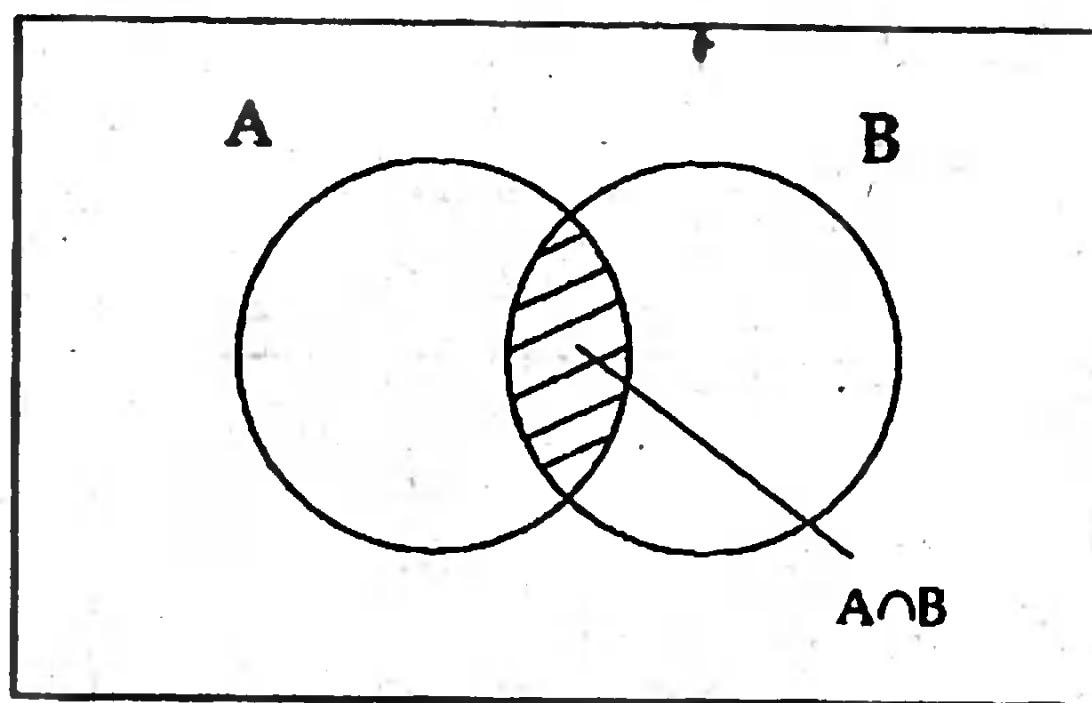


Fig. 8.2.1

Example 8.2.11. Let $A = \{1, 3, 5, 6, 8\}$ and $B = \{1, 5, 7, 8, 9\}$,
then $A \cap B = \{1, 5, 8\}$.

Disjoint or Mutually Exclusive sets. If A and B are two subsets of the universal set Ω , then A and B are said to be mutually exclusive or disjoint sets if they have no elements in common. In this case $A \cap B = \emptyset$. It is to be noted that A and \bar{A} are disjoint or mutually exclusive.

Now we state a very important law of sets which is known as addition law of set.

8.3. Addition Laws of Sets

For any two sets A and B,

$$N(A \cup B) = N(A) + N(B) - N(AB)$$

If A and B are disjoint, then $N(A \cup B) = N(A) + N(B)$

Now we shall cite two examples relating the above law.

Example 8.3.1. Employee Benefits. According to a survey of business firms in a city, 760 firms offer their employee health insurance, 650 offer dental insurance, and 285 offer health insurance and dental insurance. (i) How many firms offer their employee's health insurance or dental insurance? (ii) How many firms offer their employees only health insurance? (iii) How many firms only one insurance?

Solution. Let H be the set of firms that offer their employees health insurance and D be the set that offer dental insurance, then $H \cap D$ = Set of firms that offer health insurance and dental insurance.

Solution.

(i) $H \cup D$ = Set of firms that offer health insurance or dental insurance.

Thus, $N(H) = 760$, $N(D) = 650$, $N(H \cap D) = N(HD) = 285$ and

$$\begin{aligned} N(H \cup D) &= N(H) + N(D) - N(H \cap D) \\ &= 760 + 650 - 285 = 1,125 \end{aligned}$$

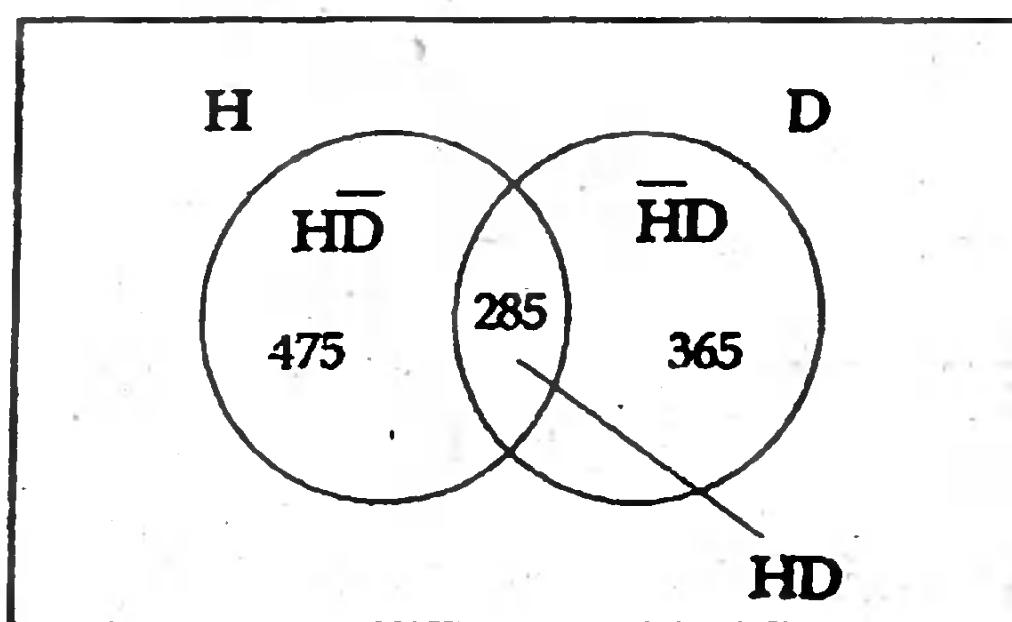


Fig. 8.3.1

Thus, 1,125 firms offer their employee's health insurance or dental insurance.

(ii) The event only health insurance = $\bar{H}\bar{D}$

$$N(\bar{H}\bar{D}) = N(H) - N(HD) = 760 - 285 = 475$$

(iii) Only one insurance = $\bar{H}\bar{D} \cup \bar{H}D$

$$\begin{aligned} N(\bar{H}\bar{D} \cup \bar{H}D) &= N(H) - N(HD) + N(D) - N(HD) \\ &= 760 - 285 + 650 - 285 = 840 \end{aligned}$$

A Matched problem to solve. A survey was conducted on some business firms. It was found that 345 firms offer their employees group life insurance, 285 offer long-term disabilities insurance, and 115 offer group life insurance and long-term disability insurance. How many firms offer their employees group life insurance or long-term disability insurance? How many firms offer their employees only group life insurance?

Ans. 515, 230

Example 8.3.2. Suppose in a class, there are 13 male students and 16 female students. Let A be the set of male students and B be the set of female students. Since these two sets have no elements in common, the intersection of sets A and B, denoted $A \cap B$, is the empty set \emptyset , then it is said that A and B are mutually exclusive sets.

Here, the number of elements in A is 13. That is $N(A) = 13$ and the number of elements in B is 16. That is $N(B) = 16$. Here $A \cap B = \emptyset$. $N(A \cap B) = N(AB) = 0$.

Union of the sets A and B, denoted $A \cup B$, is the set of all students in the class. The number of elements in $A \cup B$, denoted by $N(A \cup B)$ is given by $N(A \cup B) = N(A) + N(B) = 13 + 16 = 29$.

8.4. Venn Diagram

The diagram by which we can show the relationships among different sets and the corresponding universal set is called Venn diagram. It is named after the name of English logician John Venn (1834 - 1883). Different forms of Venn diagrams are shown below:

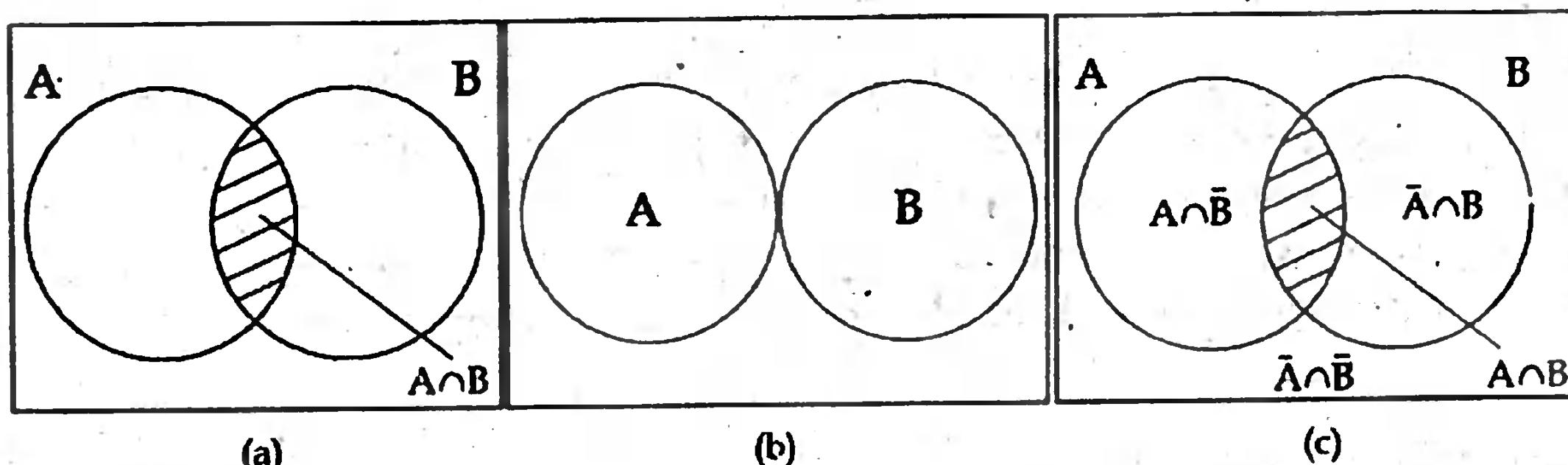


Fig. 8.5. (a). Venn diagram for two not mutually exclusive sets; (b) Venn diagram for two mutually exclusive sets and (c) Venn diagram showing different types of sub-sets of two events.

8.5. Tree Diagram

Very often the sample points of an experiment consist of more than one element. This happens when the experiment is conducted with more than

one-steps. The sample space of such experiment can be represented by a diagram which we call tree diagram. Sample points of such experiment can be displayed by a tree diagram.

Example 8.5.1. Toss two coins. Show the sample points with a tree diagram and construct the sample space of the experiment.

Solution: The experiment can be considered as two steps. There are two possible outcomes for the first coin and two possible outcomes for the second coin. The total number of outcomes will be 2×2 and the tree diagram of the sample space will be as follows:

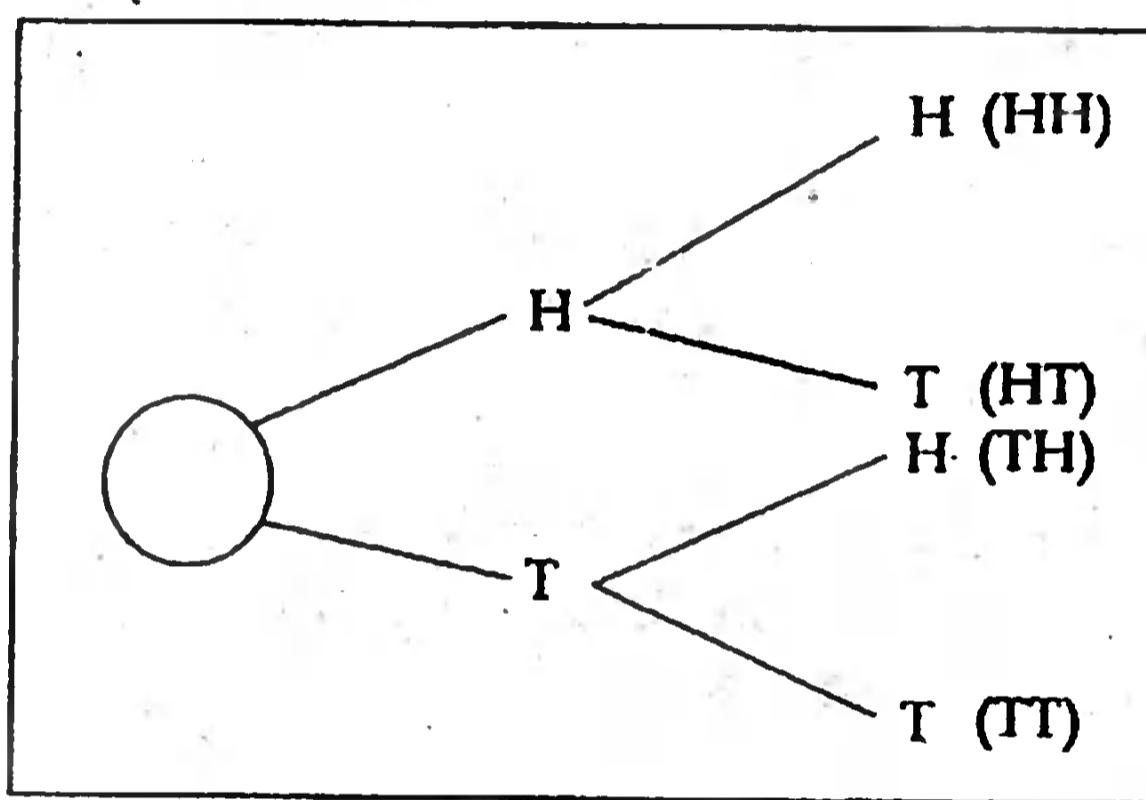


Fig. 8.5.1. Tree diagram showing outcomes of two-coin tossing experiment.

The sample space of the experiment is $S = \{ HH, HT, TH, TT \}$. There are four sample points of this random experiment and each outcome consists with two elements one from each coin. The two steps experiment can be generalized by the $m \times n$ rule.

8.5.1. The $m \times n$ rule. For a two steps experiment, sample points can be counted by the $m \times n$ rule. Suppose an experiment consists with two steps. If m is the possible outcomes of the first step and n is the possible outcomes of the second step, then the total number of outcomes of the experiment will be $m \times n$ ordered pairs. That is each outcome will contain two elements one from each step.

Example 8.5.2. Toss a coin and a die simultaneously. Show the sample points with a tree diagram and construct the sample space.

Solution. The experiment can be considered as two steps. There are two possible outcomes for the coin and six possible outcomes for the die. The total number of outcomes will be $2 \times 6 = 12$. The sample space of the experiment is

$$S = \{ 1H, 2H, 3H, 4H, 5H, 6H, 1T, 2T, 3T, 4T, 5T, 6T \}$$

The tree diagram of the experiment is

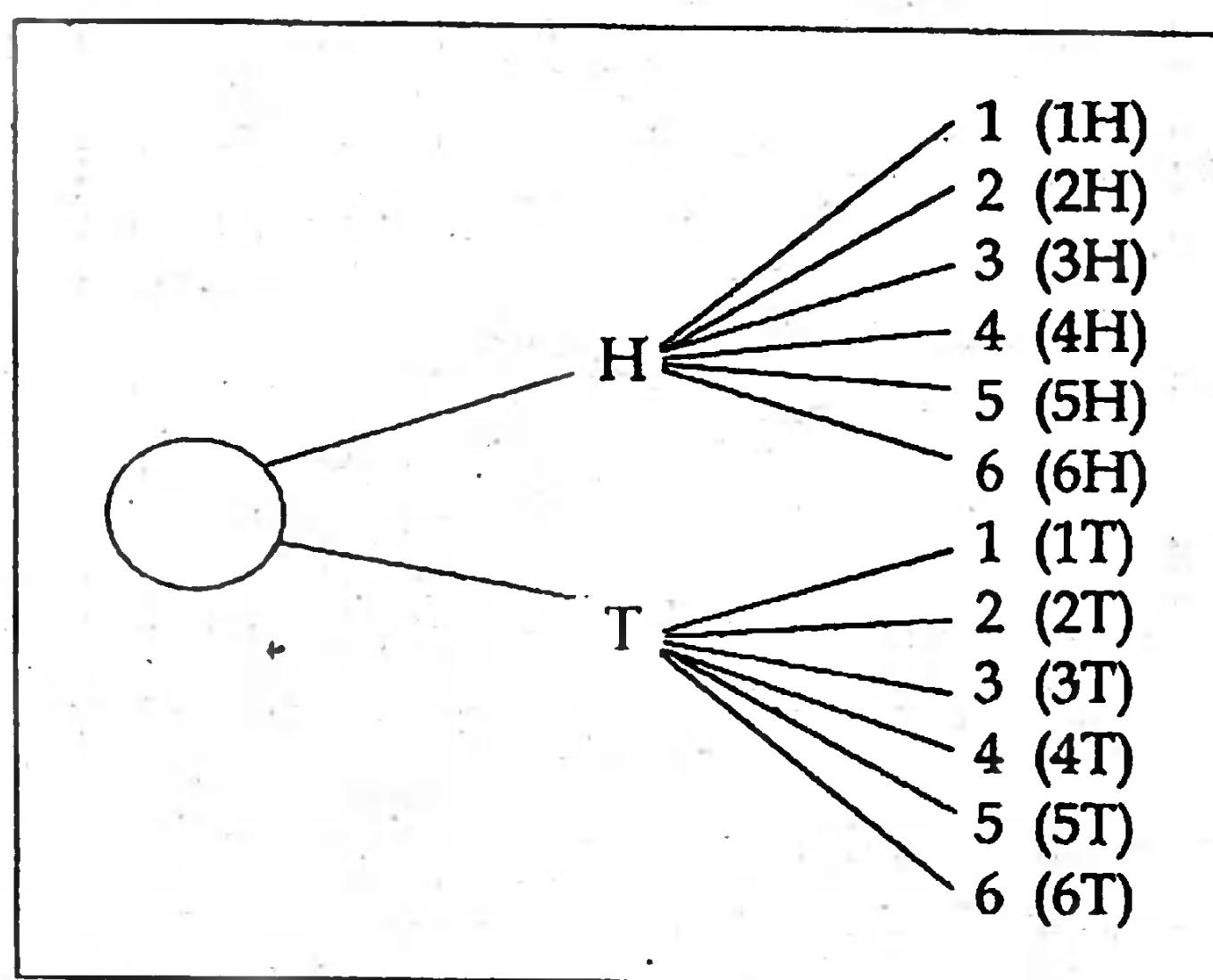


Fig. 8.5.2. Tree diagram showing the outcomes of a coin and a die

A Matched problem to solve. Four students A, B, C and D are selected at random from a statistics class and are classified as male or female. List the elements of the sample space by a tree diagram by using the letter M for male and F for female and construct the sample space. The rules can be extended for any number of steps.

If in an experiment of k steps, suppose n_1, n_2, \dots, n_k are the possible outcomes of the 1st, 2nd, ..., k th step, then the total number of outcomes of the experiment will be $n_1 \times n_2 \times \dots \times n_k$ and each outcomes of the experiment will contain k elements one from each step.

Example 8.5.3. Suppose that three items are selected at random from a manufacturing process. Each item is inspected and classified defective, D, or no defective, N. List the elements of the sample space with a tree diagram and construct the sample space of the experiment.

Solution. The experiment can be considered as three steps. There are two possible outcomes in the 1st step i.e. the selected first item may be D or N. Similarly the selected 2nd and 3rd item may be D or N. The total number of outcomes will be $2 \times 2 \times 2 = 8$. The sample space of the experiment is

$$S = \{ \text{DDD}, \text{DDN}, \text{DND}, \text{DNN}, \text{NDD}, \text{NDN}, \text{NND}, \text{NNN} \}$$

The tree diagram of the outcomes will be as follows:

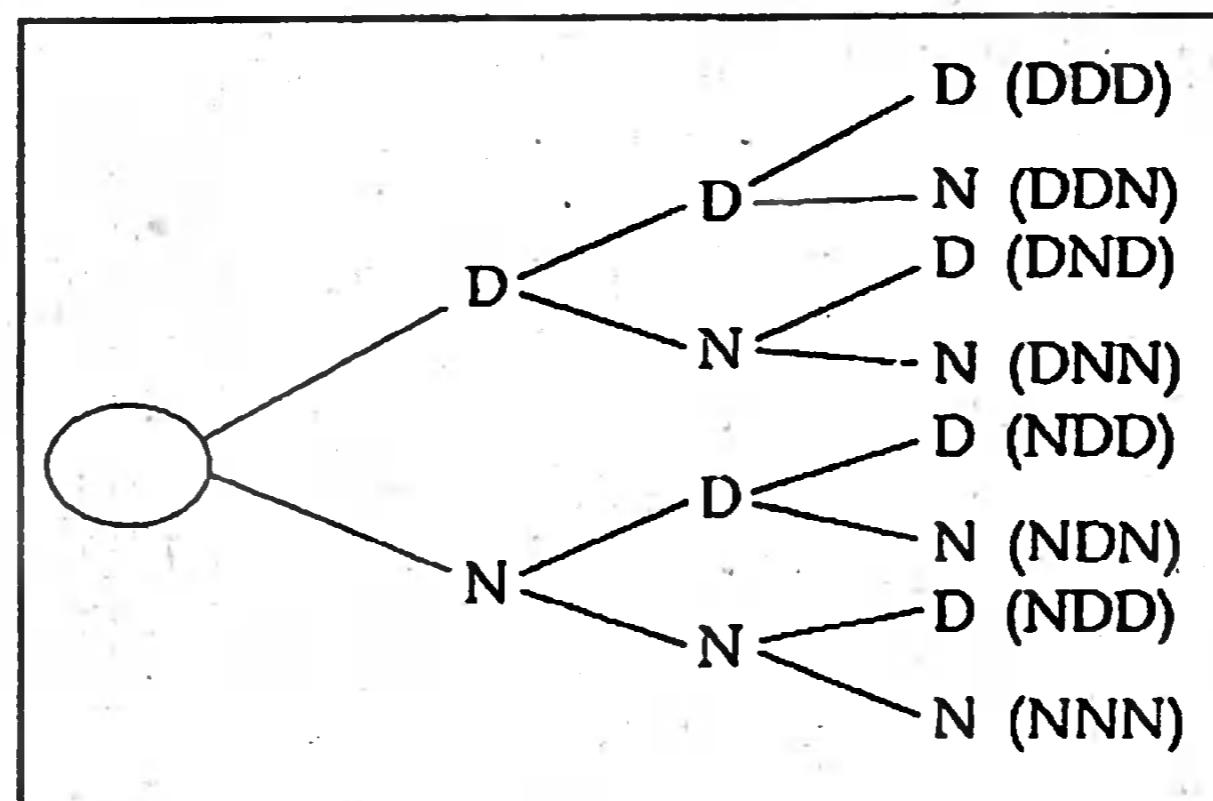


Fig. 8.5.3. Tree Diagram showing possible outcomes of three items.

A Matched Problem to solve. An experiment consists of asking three shoppers A, B and C at random if they wash their dishes with brand X detergent. List the elements of the sample space with a tree diagram using the letter Y for yes and N for no and write the sample space.

8.6. Concepts Related to Probability

Now we shall define some concepts, which are necessary to define probability. Every event is related with some experiment.

Definition. Experiment. Experiment is an act that can be repeated under certain conditions.

Definition. Outcomes. The results of an experiment are called outcomes. Experiment may be (i) Non-random or deterministic experiment and (ii) Random experiment.

Definition. Deterministic Experiment. The experiment is called deterministic when the outcome or result is unique or certain.

By a deterministic experiment we mean one in which the result can be predicted with certainty. For example

- (a) For a perfect gas, if P is the pressure and V is the volume, then

$$PV = \text{constant}$$

Provided the temperature remains constant.

- (b) If H is the hydrogen and O is the oxygen, then



A remarks. The possible outcome of a deterministic experiment is unique.

Sometimes the result of an experiment cannot be predicted with certainty. These types of experiments are called random experiment. Probability is related with random experiment.

Definition. Random Experiment. An experiment is called random experiment whose outcomes cannot be predicted with certainty

The possible results of a random experiment are not unique but may be one of several possible outcomes. For example

- (i) In tossing a coin one is not sure if a head or a tail will be obtained.
- (ii) If a light-tube has lasted for t hours, nothing can be said about its further life. It may fail to function at any moment.
- (iii) Number of defective items produced by machine per hour.
- (iv) Drawing a card from a pack etc.

Remarks. The possible outcomes of a random experiment is more than one.

Definition. Sample Space. The collection or totality of all possible outcomes of a random experiment is called sample space. S or Ω usually denotes sample space.

Example 8.6.1. If we toss a coin, the sample space is $S = \{ H, T \}$.

Where H and T denote the head and tail of the coin respectively.

Definition. Sample point. Each outcome in a sample space is called a sample point. It is also called an element or a member of the sample space.

Event : One or more outcomes of a random experiment constitute an event. A, B, C , etc usually denote events.

In the language of set, an event is a subset of a sample space.

Definition. Event. An event is a subset of the sample space.

There are two kinds of events: (i) simple event and (ii) compound event.

Definition. Simple event. An event is called simple event if it contains only one sample point.

This means each sample point of a sample space is a simple event. Simple events are also called elementary events. E usually denotes elementary event.

Definition. Compound Event. An event is called compound event if it contains more than one sample points.

That is, a compound event is the union of more than one simple event. For example in a die throwing experiment, the event of odd numbers is a compound event $A = \{ 1, 3, 5 \}$.

There are two extreme events. They are called impossible event and sure event.

Example 8.6.2. Toss a coin. The sample space of the experiment is, $S = \{ H, T \}$

The possible subsets of the sample space are { H }, {T}, {H,T}, ϕ . Here there are four events. Each of the first two events contains single sample point. They are known as simple event. The third event {H, T} contains two sample points. It is called compound event. It is also called sure event, since it must occur. The last event contains no sample point. It is called impossible event.

Definition. Impossible Event. The event that contains no sample points is called impossible event. That means impossible event will never happen.

Example 8.6.3. Toss a coin. The event H and T will happen together is an impossible event.

Definition. Sure Event. The event that contains all the sample points of an experiment is called sure event. It must happen. The sample space is a sure event.

Example 8.6.4. Toss a coin. The outcome H or T is a sure event.

Remark. A set may be considered as event and the universal set may be considered as a sample space. The main differences among sets and universal set with the events and sample space is that the sample space and events are related with the random experiments. That is, all events are subset but all subsets are not events. Similarly, all sample spaces are universal sets but all universal sets are not sample spaces.

Venn diagram is also used to show the relationship between sample space and events.

Usually, sample space is shown by a rectangle, and circles within the rectangle show events.

Union of two events : Union of two events A and B is also an event, which contains all the elements of A or B or both. It is denoted by $A \cup B$. Union of events means at least one event will happen.

Intersection of Events : Intersection of two events is also an event, which contains all elements of both A and B. It is denoted by $A \cap B$. Intersection of events means events they happen simultaneously.

Mutually Exclusive events. Two events are said to be mutually exclusive if they have no sample points in common. If A and B are two mutually exclusive events, then $AB = \phi$.

It is to be noted that all simple events of a sample space are mutually exclusive.

Example 8.6.5. A business firm wants to appoint two salesmen. Four persons applied for the post. Each person has an equal chance of being selected. However, after a test the applicants were ranked as 1, 2, 3, and 4. Construct a sample space of the experiment. Let A be the event that best two candidates will be selected. Let B be the event that at least one best candidate will be selected. Find $A \cup B$ and $A \cap B$.

Solution. The sample space of the experiment is

$$S = \{ (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4) \}; N(S) = 6$$

$$\text{Here } A = \{ (1, 2) \}; N(A) = 1$$

$$B = \{ (1, 2), (1, 3), (1, 4), (2, 3), (2, 4) \}; N(B) = 5$$

$$A \cup B = \{ (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4) \}; N(A \cup B) = 6$$

$$A \cap B = \{ \} ; N(A \cap B) = 0$$

Definition. Complementary event. Let A be any event defined on a sample space S, then the complementary of A denoted by \bar{A} , or A^c or A' is the event consisting of all the sample points in S but not in A.

That is, $A \cup \bar{A} = S$. This means A and \bar{A} are mutually exclusive and $A \cap \bar{A} = \emptyset$.

Example 8.6.6. Roll a die and observe the number that appears on the upper face of the die. Construct the sample space and define some simple and compound events from it.

Solution. The sample space of the experiment is $S = \{ 1, 2, 3, 4, 5, 6 \}$

Let us define some events.

A : observe a number 5

B : observe an odd number

C : observe a number less than 3

D : observe an even number

Here $A = \{ 5 \}$, $B = \{ 1, 3, 5 \}$, $C = \{ 1, 2 \}$, $D = \{ 2, 4, 6 \}$

Here the event A is a simple event since it contains only one element of the sample space. In this experiment, there are six simple events, such as $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$. All these simple events are mutually exclusive. The rest of the events B, C and D are compound events since each of them contains more than one sample points. Here, the events B and D are mutually exclusive events since they contain no elements in common! That is $BD = \emptyset$. The event BD is an impossible event. If you roll a die, the faces with odd and even numbers cannot occur simultaneously. It is to be mentioned that B is the complementary of D and vice-versa since $B \cup D = S$. Here S is a sure event.

Definition. Event Space. The class of all events associated with a given experiment is called event space.

Number of events from a sample space. If a sample space contains n sample points, then the number of all possible events from the sample space is 2^n . It is just like the number of subsets from a set.

Example 8.6.7. Toss a coin. The sample space of the experiment is $S = \{H, T\}$. The number of events from this sample space is $2^2 = 4$. The possible events are $\{H\}, \{T\}, \{H, T\}, \emptyset$. That is, event space contains four events.

Example 8.6.8. Toss two coins. Write down the names of all the members of the event space. The sample space of the random experiment is

$$S = \{ HH, HT, TH, TT \}$$

The total number of events in event space is $2^4 = 16$. The possible events in event space are $\{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, T\}, \{HH, HT, TH, TT\}, \emptyset$.

In this experiment, there are four simple events, eleven compound events and one impossible event.

Definition. Mutually Exclusive Outcomes. Outcomes of an experiment are called mutually exclusive when no two outcomes can happen simultaneously.

Example 8.6.9. (i) In tossing a coin, the outcomes head and tail are mutually exclusive, (ii) A finished product may be defective and non-defective are mutually exclusive.

Definition. Equally Likely outcomes. Outcomes are called equally likely if one does not occur more often than the other.

In this case the sample points of a sample space are all equal probable. This type of sample space is called simple sample space.

Simple Sample Space. A sample space is called simple sample space if all the sample points are equal probable.

Example 8.6.10. (i) In tossing a fair coin, the outcomes head and tail are equally likely, (ii) In case of a finished product, the outcomes defective and non-defective are not equally likely.

Definition. Favourable Outcomes. Outcomes of an experiment are called favourable to an event, which entail the happening of the event.

Example 8.6.11. In throwing a die, the favourable outcomes corresponding even number of points on the faces of the die is 3. They are 2, 4 and 6.

Definition. Exhaustive Outcomes. Outcomes are called collectively exhaustive if no other outcomes are possible for a given experiment.

There are two exhaustive outcomes if we toss a coin. There are six exhaustive outcomes if we throw a die.

8.6.1. Meaning and statement of some events in probability

$A \cup B$: A or B or both occurs,

: At least one occurs

$A \cap B = AB$: A and B both occurs

$\bar{A}\bar{B}$: Neither A nor B occurs

$A\bar{B}$: only A occurs

$A\bar{B} \cup \bar{A}B$: only one occurs

$A\bar{B} \cup \bar{A}B \cup \bar{A}\bar{B}$: At most one occurs

: Not more than A or B occurs

\bar{A} : Complementary of A

$AB = \phi$: A and B are disjoint or mutually exclusive

$AB \neq \phi$: A and B may be dependent or independent and not disjoint.

Example 8.6.12. Suppose two bands of refrigerator, say A and B are available in the market. A survey was conducted on 1000 people, 500 liked A, 400 liked B, 200 liked both A and B. A person is selected at random from this 1000 people. How many persons liked (i) A or B, (ii) only A, (iii) only one, (iv) neither A nor B. Draw a Venn diagram and count the sample points.

Solution. Let A be the event that a person liked band A and B be the event that a person liked band B.

$$N(A) = 500, N(B) = 400, N(AB) = 200$$

The Venn-diagram of the problem is

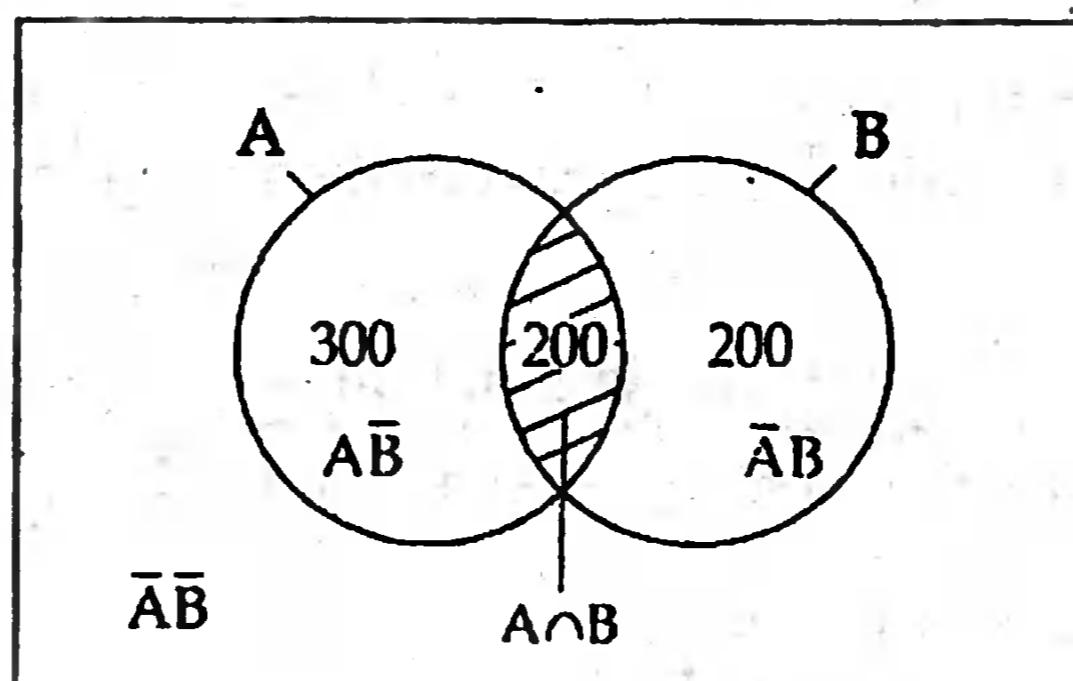


Fig. 8.6.1. Venn diagram.

- (i) $N(A \cup B) = N(A) + N(B) - N(AB) = 500 + 400 - 200 = 700$
- (ii) $N(A \bar{B}) = N(A) - N(AB) = 500 - 200 = 300$
- (iii) $N(A \bar{B} \cup \bar{A}B) = N(A \bar{B}) + N(\bar{A}B) = N(A) - N(AB) + N(B) - N(AB)$
 $= 500 - 200 + 400 - 200 = 500$
- (iv) $N(\bar{A}\bar{B}) = N(S) - N(A \cup B) = 1000 - 700 = 300.$

8.7. Definition of Probability

There are four approaches of defining probability. They are

- i) Classical or mathematical probability,
- ii) Empirical or statistical or frequency probability,
- iii) Subjective probability
- iv) Axiomatic probability

8.7.1. Classical or mathematical probability. If there are n mutually exclusive, equally likely and exhaustive outcomes of a random experiment and if m of these outcomes are favourable to an event A , then the probability of the event A , denoted by $P[A]$ is defined as

$$P[A] = \frac{m}{n}; \quad 0 \leq P[A] \leq 1$$

Laplace gave this definition of probability in his classical work on the subject.

The classical probability does not require the 'actual experimentation or previous experience. It enables us to obtain probability by logical reasoning even without conducting actual trial and hence it is known as a prior or mathematical probability.

Classical probability can also be defined with the help of sample space as in following way.

If a sample space S contains n equally likely sample points and m of this sample points are favourable to an event A , then the probability of the event A is defined by

$$P[A] = \frac{\text{Number of sample points in } A}{\text{Number of sample points in } S} = \frac{N(A)}{N(S)} = \frac{m}{n}$$

Here the sample space is simple.

The mutually exclusive and exhaustive conditions of events do not require here, since the sample points of a sample space are all mutually exclusive and exhaustive.

Let \bar{A} be the complementary of the event A, then the probability of the event not A, denoted by $P[\bar{A}]$ is defined by

$$P[\bar{A}] = \frac{\text{Number of sample points in } \bar{A}}{\text{Number of sample points in } S} = \frac{N(\bar{A})}{N(S)} = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P[A]$$

Therefore, $P[A] + P[\bar{A}] = 1$.

This gives the numerical measure of probability. Clearly $P[A]$ is non-negative and not greater than unity. That is

$$0 \leq P[A] \leq 1$$

That is, probability of an event may be zero and one in some extremes cases.

Probability of a sure event

Sample space is called sure event and it contains all the sample points. The probability of a sure event is

$$P[S] = \frac{N[S]}{N[S]} = \frac{n}{n} = 1.$$

Here n is the number of sample points in S.

Probability of an impossible event

Impossible event contains no sample points. The number of sample points in favour of an impossible event is zero. Hence the probability of an impossible event is

$$P[\phi] = \frac{N[\phi]}{N[S]} = \frac{0}{n} = 0.$$

Here n is the number of sample points in S.

Actually there are three category of events; i) Sure event; ii) impossible event and iii) uncertain events.

Sure Event

When the outcome of an event is certain then the event is called sure event. The probability of a sure event is one. Sample space is a sure event.

Impossible Event

When the outcome of an event will never happen, then this type of event is called impossible event. The probability of an impossible event is zero.

Uncertain event

When the outcomes of an event may or may not happen, then this type of event is called uncertain event. This kind of event is important. In most of the cases, we need to find the probabilities of these kinds of events. The probabilities of these kinds of events lie between 0 and 1.

Example 8.7.1. A box contains 12 items of which two are defectives. An item is selected at random from this box. Find the probability that the selected item is (i) non-defective, (ii) defective, (iii) defective or non-defective (iv) defective and non-defective. Comment on the nature of the events.

Solution. Here there are 12 equally likely, mutually exclusive and exhaustive outcomes. That is $N(S) = 12$.

(i) Let A be the event that the selected item is non-defective. Then $N(A) = 12 - 2 = 10$.

$$P[A] = \frac{\text{# of non-defective item}}{\text{Total # of items}} = \frac{N(A)}{N(S)} = \frac{10}{12} = \frac{5}{6}.$$

Here the event A is an uncertain event.

(ii) Let B be the event that the selected item is defective. Then $N(B) = 2$.

$$P[B] = \frac{\text{# of defective item}}{\text{Total # of items}} = \frac{N(B)}{N(S)} = \frac{2}{12} = \frac{1}{6}.$$

Here B is also an uncertain event.

(iii) Let C be the event that the item is defective or non-defective. Then $N(C) = 12$

$$P[C] = \frac{N(C)}{N(S)} = \frac{12}{12} = 1.$$

Here C is a sure event.

(iv) Let D be the event that the selected item is defective and non-defective. Then $N(D) = 0$.

$$P[D] = \frac{N(D)}{N(S)} = \frac{0}{12} = 0.$$

Here D is an impossible event.

A Matched problem to solve. A bag contains 4 red and 5 black marbles. One marble is selected at random from this bag, what is the probability that the selected marble is i) red, ii) black, iii) red or black and vi) red and black. Comment on the nature of the events.

Ans. i) $4/9$, uncertain event, ii) $5/9$, uncertain event,
iii) 1, sure event and iv) 0, impossible event.

8.7.1.1. Drawbacks or limitations of classical probability. Mainly, there are three drawbacks of classical probability.

- i) The classical probability fails to define probability, when the total number of possible outcomes is infinite.
- ii) The classical definition leaves us completely helpless when the possible outcomes are not equally likely.
- iii) It is not always possible to enumerate all the equally likely cases.

Empirical or statistical probability has been developed to overcome the limitations of classical probability.

8.7.2. Empirical or Statistical probability. The outcomes of an experiment are not always equally likely. For example, defective and non-defective items produced by a machine are not equally likely, the occurrence of rain on 1st July and on 1st January are not equally likely, the outcomes of a cricket or football (win, loss or draw) are not equally likely. Hence, we cannot find the probability of the event like obtaining a defective item or occurrence of rain on 1st July, win of a team with the help of classical probability. We have to find the probability of such events using empirical data. That is by using the concept of empirical or statistical probability.

Empirical or Statistical or frequency probability. If a trial is repeated n number of times under the same conditions and if an event A occurs m times, then the probability of the event A is then unique limiting value p of the ratio m/n as n tends to infinity. Symbolically,

$$P[A] = p = \lim_{n \rightarrow \infty} \frac{m}{n}$$

Von Mises gave this definition of probability.

The empirical probability is the most widely accepted definition of probability. It is the definition that most frequently comes to mind when we are confronted with a probability statement. For example, when the weather bureau predicts a 0.9 or 90 % probability of rain, we assume that 90% of the day with identical weather conditions will be rained.

Example 8.7.2. A component goes through five operations before it is completed. Basic units of raw material introduced are 3200 units. Information relative to production is given below:

Basic units of raw material introduced 3200

Operation	Number of rejection
1st operation	136
2nd operation	126
3rd operation	56
4th operation	48
5th operation	12

- i) What is the probability that a basic unit of raw material will become a finished component?
- ii) What is the probability that a basic unit of raw material getting beyond the second operation?
- iii) If a further 600 finished components are required how many basic units of raw material should be introduced?

Solution. Total number of rejected units = $136 + 126 + 56 + 48 + 12 = 378$

Total number of finished components = $3200 - 378 = 2822$

- i) Let A be the event that a basic unit of raw material will become a finished component, then,

$$P[A] = \frac{2822}{3200} = 0.882.$$

- ii) Let B be the event that a basic unit of raw material getting beyond the second operation

Number of rejections at first two operations = $136 + 126 = 262$

Number of basic units beyond the second operation = $3200 - 262 =$

$$2938. \text{ So, } P[B] = \frac{2938}{3200} = 0.918$$

- iii) Total number of more basic unit required to get 600 finished components is

$$\frac{600}{0.882} = 680.27 \approx 681 \text{ Units}$$

Matched problem. The following table gives a distribution of monthly wages of 4000 employees of a firm:

Wages in taka	Below 500	500-750	750-1000	1000-1250	1250-1500	1500-1750	1750-and above
Number of workers	36	472	1912	800	568	140	72

An employee is selected at random from this firm. What is the probability that his wages are (i) under Tk.750, (ii) above Tk.1250 and (iii) between 750 and 1250?

Ans.(i) 0.127, (ii) 0.195, and (iii) 0.678.

8.7.2.1. Drawbacks of empirical probability. Empirical probability is not also free from drawbacks. The following are the drawbacks of this definition.

- i) In practice, it is not possible to repeat an experiment an infinite number of times under the same conditions to get the probability.
- ii) Even if it were possible to repeat an experiment an infinite number of times, it is conceivable that a different infinite sequence of performance of the same experiment could produce a different value of the probability.
- iii) It is not clear how large n should be before we are certain that the probability is close to the limiting of m/n as n tends to infinity.

8.7.3. Axiomatic probability. The Russian mathematician A.N. Kolmogorov in 1933 first introduced axiomatic probability and now universally accepted by all probabilists and mathematical statisticians. Perhaps, it is the simplest of all the definitions and least controversial. This definition is based on a number of axioms, which allow rigorous development of the mathematics of probability.

Axioms: An axiom is a statement that is assumed to be true.

Theorem: A theorem is a statement that can be deduced either from axioms or from previously proved theorem.

In axiomatic probability, several simple statements concerning probability are assumed to be true.

Axiomatic definition of probability

Definition. Suppose S is a sample space and A is an event of this sample space. Then the probability of the event A , denoted by $P[A]$ must satisfy the following four axioms:

- i) $P[A] \geq 0$;
- ii) $P[S] = 1$;
- iii) If A and B are two mutually exclusive events, then

$$P[A \cup B] = P[A] + P[B]$$
- iv) Let $A_1, A_2, \dots, A_k, \dots$ be a sequence of mutually exclusive events, then

$$P[A_1 \cup A_2 \cup \dots \cup A_k \cup \dots] = P[A_1] + P[A_2] + \dots + P[A_k] + \dots$$

The axiom (i) is referred as the axiom of positiveness, which states that probability of an event can never be negative.

The axiom (ii) is known as the axiom of certainty and states that the probability of sure event is always one.

The axiom (iii) and (iv) are known as the axioms of additivity.

8.7.4. Subjective probability. Subjective probability is a personal evaluation of the likelihood of chance of phenomena. Keynes and Jeffreys developed subjective probability. It is an important element in many decision-making processes and is a basic ingredient for Bayesian decision theory.

Definition. The probability of an event based on the degree of a person's belief or judgment is called subjective probability.

Like other definitions, subjective probability of an event is a number ranging from 0 to 1. As it depends on individual's judgment and belief, it may vary from individual to individual even when they are confronted with the same set of evidence. For example, one fine morning Mr. Ali will be prepared for rain, but his friend Mr. Ahmed may not.

8.8. Joint Probability and Marginal Probability

Definition. Joint probability. The probability of the simultaneous occurrence of two or more events is called joint probability.

Actually intersection of two or more events is called joint event .The event AB is called the joint event. If A and B are two events, then the joint probability of the events A and B is denoted by P [AB].

Definition. Marginal Probability. The probability of an event for all possible values of other events is called marginal probability. Probability P[A] is called the marginal probability of A.

Example 8.8.1. A survey was conducted on 100 people of which 60 are male and 40 are female. 40 male and 30 female like Sony Television. A person is selected at random from this 100 people. What is the probability that the selected person (i) likes Sony Television, (ii) is male and likes Sony Television? Also mention the kind of probability.

Solution. Let us define the following events:

M : person is a male, F : person is a female,

L : likes Sony T.V.; \bar{L} : does not like Sony T.V.

Events and their corresponding sample points in them are depicted in the table given below:

	M	F	Total
L	LM(40)	LF(30)	L(70)
\bar{L}	$\bar{L} M(20)$	$\bar{L} F(10)$	$\bar{L} (30)$
Total	M(60)	F(40)	100

$$(i) P[L] = \frac{\text{# of persons like Sony T.V.}}{\text{Total # of persons}} = \frac{N(L)}{N(S)} = \frac{70}{100} = 0.70.$$

$$(ii) P[ML] = \frac{N(ML)}{N(S)} = \frac{40}{100} = 0.40.$$

Here the probability $P[L]$ is called marginal probability of the event L and $P[ML]$ is called the joint probability of events M and L.

8.9. Conditional Probability

Conditional probability arises from the dependent events. Two events A' and B are said to be dependent with each other if the happening of one event depends on the happening of the other event. Conditional probability is obtained when the probability of one event is calculated for a given value of the other event, which has already occurred.

Definition. Conditional Probability. Let A and B be two events, then the conditional probability of A for given value of B, denoted by $P[A | B]$ is defined by

$$P[A | B] = \frac{P[AB]}{P[B]} ; \text{ Provided } P[B] > 0$$

$$P[B | A] = \frac{P[AB]}{P[A]} ; \text{ Provided } P[A] > 0$$

To read 'probability of A given B', which means, conditional probability of the event A given that another event B has already occurred.

The concept of the conditional probability is clarified with the following illustration. Suppose the sell of a particular item is dependent upon the weather condition. On a rainy day, the probability that the item will be sold is 0.4, but if it is not rained, the probability increases to 0.9. The weather forecast gives a probability 0.75 that it will be rained on a particular day. Let R be the event that it rains and S be the event that the item is sold, then the information can be represented by the following tree diagram.

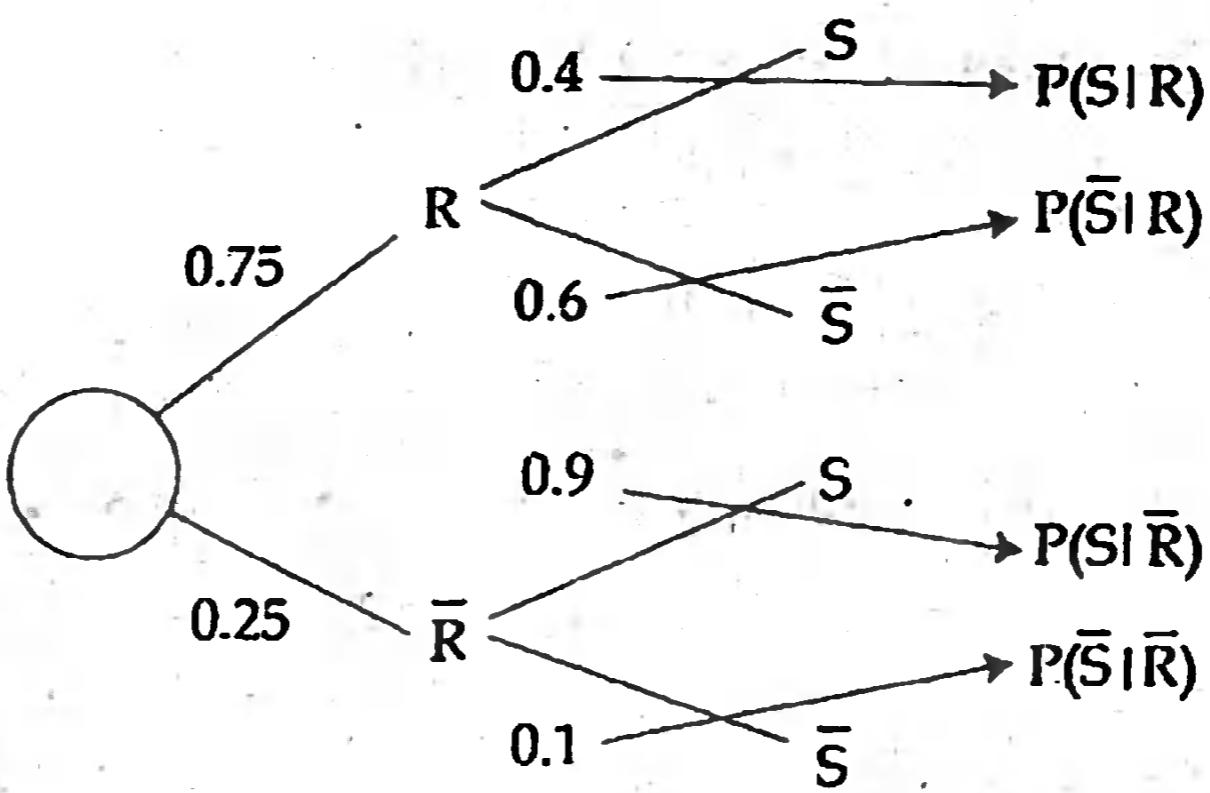


Fig. 8.9.1

Here the events in second branches are all conditional events, because occurrence of second event S in second branch depends on the occurrence of first event R. The corresponding probabilities of occurrences are the conditional probabilities.

Remark. Multiply the probabilities while moving along the branches of a tree, and add the probabilities while moving between the branches.

Example 8.9.1. The probability that a student, selected at random from the first year class at Chittagong University, will pass a statistics course is $\frac{4}{5}$, and the probability that he passes both statistics and mathematics is $\frac{1}{2}$. What is the probability that he will pass mathematics if it is known that he has passed statistics?

Solution. Let S be the event that a student passes statistics and M be the event that a student passes mathematics. Here we have to find $P[M|S]$.

$$P[M|S] = \frac{P[M \cap S]}{P[S]} = \frac{\frac{1}{2}}{\frac{4}{5}} = \frac{5}{8}.$$

8.10. Independent Events

Events are said to be independent if the happening of one event does not affect the happening of the others. If A and B are two independent events, then $P[AB] = P[A]P[B]$.

In this case $P[A] = P[A|B]$ and $P[B] = P[B|A]$.

That is, in the case of independent events marginal probability of an event is equal to its conditional probability.

Definition. Independent Events. Two events A and B are said to be independent if and only if any one of the following conditions holds

- i) $P[AB] = P[A]P[B]$, ii) $P[A] = P[A|B]$, iii) $P[B] = P[B|A]$.

It can be easily showed that if A and B are independent, then

- (i) A and \bar{B} , (ii) \bar{A} and B, (iii) \bar{A} and \bar{B} are also independent.

Interested students are advised to concern the book [2].

Example 8.10.1. Suppose we have a box containing 24 bulbs of which 3 are defective. Two bulbs are drawn one by one at random from the box with replacement. What is the probability that both bulbs are defective?

Solution. Let A be the event that the first bulb is defective and B be the event that the second bulb is defective. We have to find AB. The probability of first drawn bulb defective is $3/24$. That is $P[A] = 3/24$. The first drawn bulb is replaced in the box and then a second bulb is drawn from the box. The probability of second bulb defective is also $3/24$. That is $P[B] = 3/24$. Since A and B are independent and hence

$$P[AB] = P[A] P[B] = \frac{3}{24} \times \frac{3}{24} = \frac{1}{64}.$$

8.10.1. Formula for finding conditional probability for equally likely outcomes or simple sample space

The conditional probability of A given B is defined by $P[A|B] = \frac{P[AB]}{P[B]}$.

Let $N(S)$ be the number of sample points in sample space S. Let $N(A)$ and $N(AB)$ be the sample points in A and AB. Now,

$$P[A|B] = \frac{P[AB]}{P[B]} = \frac{\frac{N(AB)}{N(S)}}{\frac{N(B)}{N(S)}} = \frac{N(AB)}{N(B)}.$$

This formula is applicable when the sample space is simple or the outcomes are equally likely.

Example 8.10.2. Twenty students went on a picnic, 15 played football, 10 cooked lunch and 7 students took part both in cooking and playing football. It is given that a student played football what is the probability that he cooked lunch.

Solution. Let F be the event that a student played football and L be the event that a student cooked lunched. Let \bar{F} and \bar{L} be the complementary of F and L.

The given information can be displayed in the following table:

Football/Lunch	F	\bar{F}	Total
L	7	3	10
\bar{L}	8	2	10
Total	15	5	20

The conditional probability of L given F is the required probability. The sample space is simple. Hence $P[L|F] = \frac{P[LF]}{P[F]} = \frac{N(LF)}{N(F)} = \frac{7}{15}$.

8.10.2. Conditional probability when the sample space is not simple. In this case, conditional probability will be found with the help of the original formula. Now we shall cite examples when the sample points of a sample space are not equally likely or the sample space is not simple.

Example 8.10.3. A coin is made in such a way that $P[H] = 2P[T]$. The coin is tossed twice. It is given that head did not appear in the first toss, what is the probability of getting two heads.

Solution. Let p be probability of tail. Then

$$P[H] + P[T] = 1 \quad \text{or,} \quad 2p + p = 1 \quad \text{or,} \quad 3p = 1$$

Then $p = 1/3$. That is $P[H] = 2/3$ and $P[T] = 1/3$.

The sample space of the experiment is

$$S = \{HH, HT, TH, TT\}$$

$$P[HH] = P[H]P[H] = (2/3)(2/3) = 4/9$$

$$P[HT] = P[TH] = (2/3)(1/3) = 2/9$$

$$P[TT] = P[T]P[T] = (1/3)(1/3) = 1/9.$$

Here the sample points are not equally likely. That is sample space is not simple. Let A be the event of getting heads and B be the event that head did not appear in the first toss. Then

$$A = \{HH, HT, TH\}; B = \{TH, TT\} \text{ and } AB = \{TH\}$$

$$P[B] = P[TH] + P[TT] = 2/9 + 1/9 = 3/9; P[AB] = P[TH] = 2/9$$

$$P[A|B] = \frac{P[AB]}{P[B]} = \frac{2/9}{3/9} = \frac{2}{3}.$$

Example 8.10.4. A fair coin is tossed until a head appears or it has been tossed three times. Given that a head did not appear on the first toss, find the probability that the coin was tossed 3 times.

Solution. Let H and T denote the head and tail of the coin respectively. The sample space of the experiment is $S = \{H, TH, TTH, TTT\}$

Since the coin is fair, $P[H] = P[T] = 1/2$. The probabilities corresponding to the different sample points are

$$P[H] = 1/2, P[TH] = 1/4, P[TTH] = 1/8 \text{ and } P[TTT] = 1/8.$$

Here the sample points are not equally likely.

Let A be the event that the coin was tossed three times and B be the event that the head did not appear on the first toss.

The number of sample points in A and B are

$$A = \{ TTH, TTT \} \text{ and } B = \{ TH, TTH, TTT \}$$

$$\text{Then } AB = \{ TTH, TTT \}$$

$$\text{Therefore, } P[A|B] = \frac{P[AB]}{P[B]}$$

$$P[AB] = P[TTH \cup TTT] = P[TTH] + P[TTT] = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

$$P[B] = P[TH \cup TTH \cup TTT] = P[TH] + P[TTH] + P[TTT] = \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

$$P[A|B] = \frac{P[AB]}{P[B]} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

Example 8.10.5. The probability that Mr. X will die in the next 20 years is $1/5$ and the probability that Mr. Y will die in the next 20 years is $1/7$. What is the probability that (i) both will die in the next 20 years, (ii) neither will die in the next 20 years, (iii) at least one will die in the next 20 years, (iv) and only Mr. X will die the next 20 years?

Solution. Let A be the event that Mr. X will die in the next 20 years and B is the event that Mr. Y will die in the next 20 years. Let \bar{A} and \bar{B} be the complementary of A and B respectively. Here A and B are independent. Hence \bar{A} and \bar{B} are also independent.

$$\text{We have, } P[A] = \frac{1}{5}, P[B] = \frac{1}{7}. \text{ Then } P[\bar{A}] = 1 - P[A] = 1 - \frac{1}{5} = \frac{4}{5}.$$

$$\text{Similarly, } P[\bar{B}] = \frac{6}{7}$$

$$(i) P[\text{both will die}] = P[AB] = \frac{1}{5} \times \frac{1}{7} = \frac{1}{35}.$$

$$(ii) P[\bar{A}\bar{B}] = P[\bar{A}]P[\bar{B}] = \frac{4}{5} \times \frac{6}{7} = \frac{24}{35}$$

$$(iii) P[A \cup B] = P[A] + P[B] - P[AB] = \frac{1}{5} + \frac{1}{7} - \frac{1}{5} \times \frac{1}{7} = \frac{11}{35}$$

$$(iv) P[A\bar{B}] = P[A]P[\bar{B}] = \frac{1}{5} \times \frac{6}{7} = \frac{6}{35}$$

8.10.3. Mutually exclusive events can never be independent or independent events can never be mutually exclusive. If A and B are two independent events, then

$$P[AB] = P[A]P[B] > 0$$

For $P[A] > 0$ and $P[B] > 0$.

This means the events A and B have some common sample points and the Venn diagram looks like as in Fig 8.10.1.

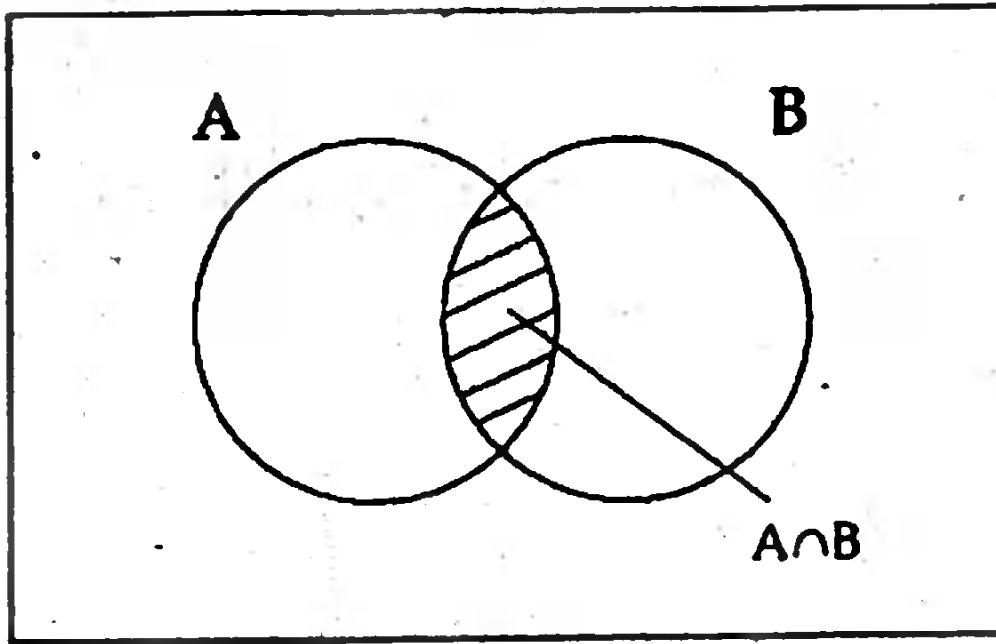


Fig. 8.10.1. Two not mutually exclusive events.

But if A and B are mutually exclusive, then $AB = \emptyset$. That is there is no common sample points between A and B as in Fig. 8.10.2.

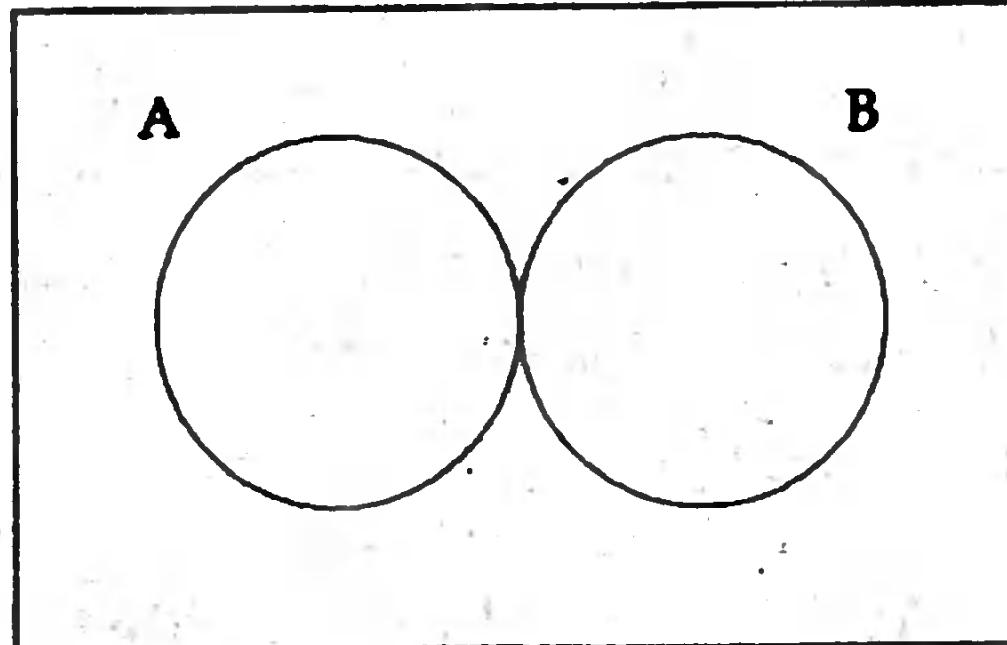


Fig. 8.10.2. Two mutually exclusive events.

In this case, $P[AB] = 0$.

Therefore the conditions shown in figures (810.1) and (8.10.2.) contradict each other. This means two independent events can never be mutually exclusive or two mutually exclusive events can never be independent.

Important Remarks. Two independent events can happen simultaneously, so $P[AB] \neq 0$. But two mutually exclusive events can never happen simultaneously. In this case $P[AB] = 0$. Therefore two mutually exclusive events can never be independent or two independent events can never be mutually exclusive.

8.11. Laws of Probability

There are two important rules or laws of probability;

- i) Addition laws of probability and
- ii) Multiplication laws of probability.

8.11.1. Addition laws of probability. If A and B are two events, then

$$P[A \cup B] = P[A] + P[B] - P[AB]$$

Proof. Let A and B be two events, \bar{A} and \bar{B} be their complementary events respectively. From the Venn-diagram, it is seen that

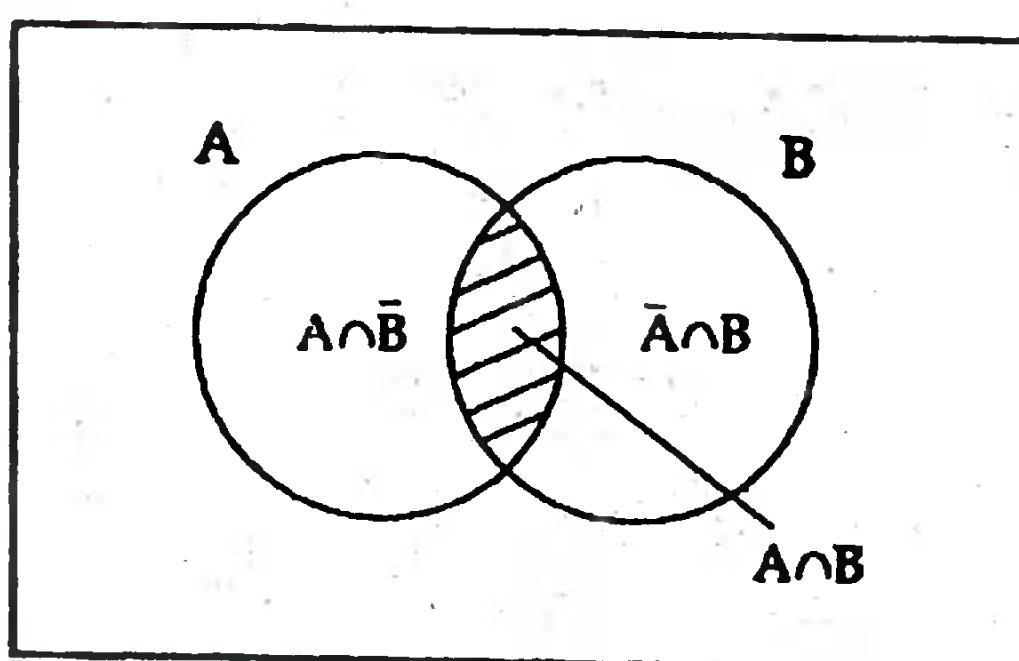


Fig. 8.11.1 Venn diagram showing two events A and B.

$$A = AB \cup A\bar{B}$$

$$\text{Then } P[A] = P[AB \cup A\bar{B}] = P[AB] + P[A\bar{B}] \quad (8.1)$$

Since AB and $A\bar{B}$ are mutually exclusive.

$$\text{Similarly, } B = AB \cup \bar{A}B$$

$$P[B] = P[AB] + P[\bar{A}B] \quad (8.2)$$

It is also seen that

$$A \cup B = \bar{A}\bar{B} \cup AB \cup \bar{A}B$$

$$\text{Then, } P[A \cup B] = P[\bar{A}\bar{B} \cup AB \cup \bar{A}B] = P[\bar{A}\bar{B}] + P[AB] + P[\bar{A}B] \quad (8.3)$$

Since $\bar{A}\bar{B}$, AB and $\bar{A}B$ are all mutually exclusive.

By adding (8.1) and (8.2) we have

$$\begin{aligned} P[A] + P[B] &= P[AB] + P[\bar{A}\bar{B}] + P[\bar{A}B] + P[A\bar{B}] \\ &= P[AB] + P[A \cup B] \end{aligned} \quad (8.4)$$

From (8.4) we have, $P[A \cup B] = P[A] + P[B] - P[AB]$. This proves the theorem.

But if A and B are mutually exclusive, then $P[AB] = 0$.

In that case, $P[A \cup B] = P[A] + P[B]$.

Example 8.11.1. The probability that a contractor will get a plumbing contract is $2/3$ and the probability that he will get an electric contract is $4/9$. If the probability of getting both the contract is $14/45$, what is the probability that he will get at least one contract.

Solution. Let A and B denote the event that a contractor will get a plumbing and electric contract respectively. Then $P[A] = 2/3$; $P[B] = 4/9$; $P[A \cap B] = 14/45$.

$$\text{Hence } P[A \cup B] = P[A] + P[B] - P[AB]$$

$$\begin{aligned} &= \frac{2}{3} + \frac{4}{9} - \frac{14}{45} \\ &= \frac{30 + 20 - 14}{45} \\ &= \frac{36}{45} = \frac{4}{5} = 0.75. \end{aligned}$$

Example 8.11.2. The probability that Parul passes economics is $7/8$ and the probability that she passes statistics is $3/4$. If the probability of passing both the course is $3/5$, what is the probability that Parul will pass at least one of the courses?

Solution. Let E be the event that Parul passes economics and S be the event that she passes statistics, then by the addition rule we have

$$\begin{aligned} P[E \cup S] &= P[E] + P[S] - P[ES] \\ &= \frac{7}{8} + \frac{3}{4} - \frac{3}{5} = \frac{31}{40}. \end{aligned}$$

Example 8.11.3. The probability that a garment industry will locate in Chittagong is 0.7, the probability that it will locate in Dhaka is 0.8 and the probability that it will locate in either Chittagong or Dhaka or both is 0.95. What is the probability that the industry will locate

- (i) in both cities?
- (ii) in neither city?

Solution. Let A be the event that the industry will locate in Chittagong and B be the event that it will locate in Dhaka. We have

$$P[A] = 0.7, P[B] = 0.8 \text{ and } P[A \cup B] = 0.95.$$

(i) According to addition law of probability, we have

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[AB] \\ \Rightarrow 0.95 &= 0.70 + 0.80 - P[AB] \\ \Rightarrow P[AB] &= 1.5 - 0.95 = 0.55 \end{aligned}$$

$$(ii) P[\overline{AB}] = 1 - P[A \cup B] = 1 - 0.95 = 0.05.$$

Example 8.11.4. Mr. Rahman wants to attend a seminar at Dhaka. The probability that he will fly to Dhaka is $3/7$ and the probability that he will go to Dhaka by train is $1/2$. What is the probability that Mr. Rahman will attend the seminar by plain or by train?

Solution. Let A be that Mr. Rahman will fly and B be the event that he will go by train. The two events A and B are mutually exclusive since he cannot go by train and plane simultaneously. That is $P[AB] = 0$. Hence,

$$P[A \cup B] = P[A] + P[B] = \frac{3}{7} + \frac{1}{2} = \frac{13}{14}.$$

Example 8.11.5. Suppose the probabilities are 0.16, 0.22 and 0.24 that a person purchasing a new automobile will choose the colour green, red and blue respectively. What is the probability that a given buyer will purchase a new automobile that comes in one of those colours?

Solution. Let G, R and B be the events that a buyer selects a green, red or blue automobile respectively. Since these three events are mutually exclusive, the probability is

$$P[G \cup R \cup B] = P[G] + P[R] + P[B] = 0.16 + 0.22 + 0.24 = 0.62.$$

8.11.2. Multiplication law of probability. If A and B are two events, then

$$P[AB] = P[A] P[B|A] = P[B] P[A|B]$$

Proof. From the definition of conditional probability, we have

$$P[A|B] = \frac{P[AB]}{P[B]}, \text{ Provided } P[B] > 0$$

It follows that $P[AB] = P[B] P[A|B]$.

(8.5)

Similarly, $P[B|A] = \frac{P[AB]}{P[A]}$; provided $P[A] > 0$.

It follows that $P[AB] = P[A] P[B|A]$ (8.6)

From (8.5) and (8.6) we have

$$P[AB] = P[A] P[B|A] = P[B] P[A|B]$$

This proves the theorem.

But if A and B are independent events, then

$$P[A] = P[A|B] \text{ and } P[B] = P[B|A]$$

Hence $P[AB] = P[A] P[B]$.

Example 8.11.6. The distribution of number of stores according to size in 3 areas is given in the following table:

Area	Store size		
	Large (L)	Medium (M)	Small (S)
A	30	45	75
B	150	125	275
C	20	130	150

Find the probabilities: (i) $P(M)$; (ii) $P(B \cap M)$; (iii) $P(A|M)$; (iv) $P(B \cup L)$; (v) Are A and L independent?

Solution.

Area	Store size			Total
	Large (L)	Medium (M)	Small (S)	
A	30	45	75	150
B	150	125	275	550
C	20	130	150	300
Total	200	300	500	1000

Here the total number of stores is 1000. That is $N(S) = 1000$

(i) Here M is the event of medium size store. Then $N(M) = 300$

$$\text{Then, } P[M] = \frac{N(M)}{N(S)} = \frac{300}{1000} = 0.300.$$

(ii) $B \cap M = BM$ is the joint event of the medium stores in the area B. Here $N(BM) = 125$.

$$P[BM] = \frac{\text{# of stores in } BM}{\text{Total # of stores}} = \frac{N(BM)}{N(S)} = \frac{125}{1000} = 0.125.$$

(iii) Here A is the event of area A. We have to find the conditional probability of A for a given value of M. According to definition,

$$P[A|M] = \frac{P[AM]}{P[M]}$$

$$\text{We have } P[M] = \frac{300}{1000}, P[AM] = \frac{N(AM)}{N(S)} = \frac{45}{1000}$$

$$P[A|M] = \frac{P[AM]}{P[M]} = \frac{N(AM)}{N(M)} = \frac{45}{300} = 0.15$$

(iv) $P[B \cup L] = P[B] + P[L] - P[BL]$

$$\text{Here } N(B) = 550, N(L) = 200, N(BL) = 150$$

$$P[B \cup L] = P[B] + P[L] - P[BL]$$

$$\begin{aligned} &= \frac{N(B)}{N(S)} + \frac{N(L)}{N(S)} - \frac{N(BL)}{N(S)} \\ &= \frac{550}{1000} + \frac{200}{1000} - \frac{150}{1000} = \frac{600}{1000} = 0.60. \end{aligned}$$

(v) The events A and L will be independent if $P[AL] = P[A]P[L]$

$$\text{Here } P[A] = \frac{N(A)}{N(S)} = \frac{150}{1000}; P[L] = \frac{200}{1000}; P[AL] = \frac{N(AL)}{N(S)} = \frac{30}{1000}$$

$$P[A] \times P[L] = \frac{150}{1000} \times \frac{200}{1000} = \frac{30}{1000} = P[AL].$$

Hence the events A and L are independent.

Example 8.11.7. A lot of 10,000 parts produced on four machines were inspected and classified into three grades. The results were given in the following table:

Grade	Machine				Total
	I	II	III	IV	
Satisfactory	2400	1600	2400	1600	8000
Rework	450	300	450	300	1500
Scrap	150	100	150	100	500
Total	3000	2000	3000	2000	10000

Requirement: If a part is selected at random from this lot, then find the following probabilities that:

- it is produced by Machine III;
- it is produced on Machine I given that it is scrapped
- it is scrapped given that it is produced on Machine IV;
- a satisfactory part is produced on Machine II.

Solution. Total number of parts of the lot is 10,000. That is sample space contains 10,000 sample points. Here $N(S) = 10,000$.

- (i) Let A be the event that the part is produced by machine III.

Here $N(A) = 3000$.

$$P[A] = \frac{N(A)}{N(S)} = \frac{3000}{10000} = 0.30.$$

- (ii) Let B be the event that the part is produced by machine I and C be the event that part is scrapped.

We have to find $P[B | C]$. By definition

$$P[B | C] = \frac{P[BC]}{P[C]}$$

Here $N(BC) = 150$, $N(C) = 500$

$$P[B | C] = \frac{P[BC]}{P[C]} = \frac{N(BC)}{N(C)} = \frac{150}{500} = 0.30.$$

- (iii) Let D be the event that the part is produced by machine IV. We have

to find $P[C | D] = \frac{P[CD]}{P[D]}$.

Here $N(CD) = 100$, $N(D) = 2000$

$$\text{Hence, } P[C | D] = \frac{P[CD]}{P[D]} = \frac{N(CD)}{N(D)} = \frac{100}{2000} = 0.100.$$

- (iv) Let E be an event that a part is produced by machine II. Here $N(E) = 2000$ and $N(S) = 10000$. Let F be the event that a part is satisfactory. Then $N(EF) = 1600$.

$$\text{Hence, } P[EF] = \frac{N(EF)}{N(S)} = \frac{1600}{10000} = 0.16.$$

Example 8.11.8. A coin is biased so that a head is twice as likely to occur as a tail. If the coin is tossed twice, what is the probability of getting (i) exactly 2 heads (ii) a head and a tail?

Solution. Let H and T denote occurring the head and tail of the coin respectively. The sample space of the experiment is $S = \{HH, HT, TH, TT\}$.

Since the coin is not unbiased, it is not possible to assign equal weights to each sample point. By assigning weights of w and $2w$ for getting a tail and a head, respectively, we have $3w = 1$ or $w = 1/3$. Hence $P[H] = 2/3$ and $P[T] = 1/3$.

- (i) Let A be the event of getting exactly two heads. Then $A = \{ HH \}$
 And since the outcomes for the two tosses are independent,

$$P[A] = P[HH] = P[H]P[H] = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}.$$

- (ii) Let B be the event of getting a head and a tail. Then $B = \{ HT, TH \}$
 $P[B] = P[HT \cup TH] = P[HT] + P[TH] = P[H]P[T] + P[T]P[H]$

$$= \frac{2}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{3} = \frac{4}{9}.$$

Now shall state two important theorem on probability

8.11.3. Theorem of Total probabilities. Let B_1, B_2, \dots, B_n be n mutually exclusive and exhaustive events in a random experiment and A be any event in the same sample space, then

$$\begin{aligned} P[A] &= P[B_1]P[A|B_1] + P[B_2]P[A|B_2] + \dots + P[B_n]P[A|B_n] \\ &= \sum_{i=1}^n P[B_i]P[A|B_i] \end{aligned}$$

This theorem is called theorem of total probabilities.

Proof. The Venn-diagram of the theorem is

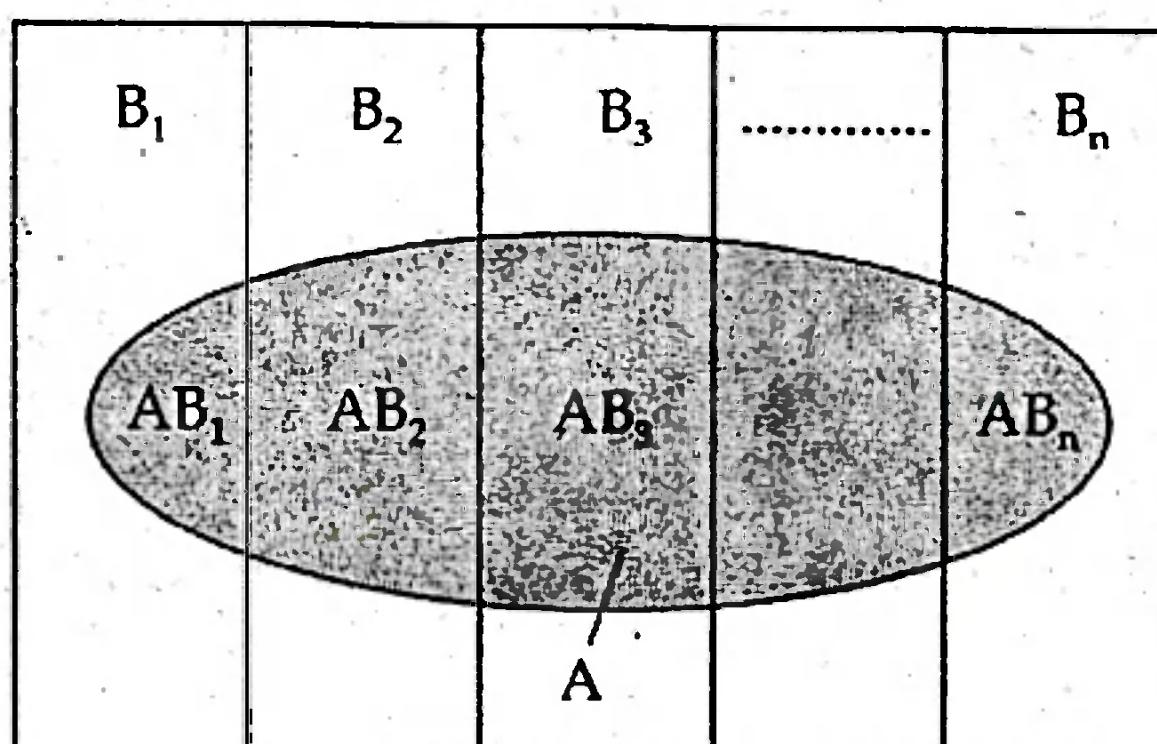


Fig. 8.11.2. Venn-diagram for total probability.

It is obvious from the Venn diagram,

$$A = AB_1 \cup AB_2 \cup \dots \cup AB_n$$

$$\text{Then, } P[A] = P[AB_1 \cup AB_2 \cup \dots \cup AB_n]$$

$$= P[AB_1] + P[AB_2] + \dots + P[AB_n] \quad (8.7)$$

Since the events AB_1, AB_2, \dots, AB_n are all mutually exclusive as B_1, B_2, \dots, B_n are mutually exclusive. From the multiplication law of probability, we know $P[AB] = P[B]P[A|B] = P[A]P[B|A]$ (8.8)

Substituting (8.8) in (8.7), we have

$$\begin{aligned} P[A] &= P[B_1] P[A|B_1] + P[B_2] P[A|B_2] + \dots + P[B_n] P[A|B_n] \\ &= \sum_{i=1}^n P[B_i] P[A|B_i]. \quad \text{This proves the theorem.} \end{aligned}$$

Remarks. B_1, B_2, \dots, B_n are mutually exclusive and exhaustive means $B_1 \cup B_2 \cup \dots \cup B_n = S$ and $B_i \cap B_j = \emptyset$ for $i, j = 1, 2, \dots, n$.

Example 8.11.9. Mr. Ali wants to buy a car this year. He applied for a bank loan. The probability that he will get the bank loan is $2/3$. If he will get the bank loan, the probability that he will buy a car is $3/4$. If he will not get the bank loan, the probability that he will buy a car is $1/4$. What is the probability that Mr. Ali will buy a car this year?

Solution. Let B_1 be the event that Mr. Ali will get the bank loan, B_2 be the event that he will not get the bank loan and A be the event that he will buy a car this year.

We are given $P[B_1] = 2/3$, $P[B_2] = 1/3$, $P[A|B_1] = 3/4$ and $P[A|B_2] = 1/4$

We have to find the probability of A . According to the theorem of total probabilities

$$\begin{aligned} P[A] &= P[AB_1 \cup AB_2] = P[B_1] P[A|B_1] + P[B_2] P[A|B_2] \\ &= (2/3)(3/4) + (1/3)(1/4) = 7/12. \end{aligned}$$

Matched Problem. Mr. Karim wants to build a house this year. He applied for a bank loan. The probability that he will get the bank loan is $4/5$. If he will get the bank loan the probability that he will build a house is $8/9$. However, if he will not get the bank loan, the probability that he will build a house is $2/9$. What is the probability that Mr. Karim will build a house this year?

Ans. 34/45.

8.12. Bayes Theorem

Now we shall discuss a very important theorem in probability, which is known as Bayes theorem after the name of Reverend Thomas Bayes. Actually he published a paper in 1763 where he outlined this theorem as inverse probability. The theorem has a wide range of application in medicine, industry, decision-making etc. Bayes theorem is the basic ingredient in Bayesian inference.

Statement of Bayes Theorem. Let B_1, B_2, \dots, B_n be n mutually exclusive and exhaustive events in a random experiment and A be any event in the same experiment such that $P[A] > 0$, then Bayes theorem states that

$$P[B_i | A] = \frac{P[B_i]P[A | B_i]}{\sum_{j=1}^n P[B_j]P[A | B_j]}, \quad i = 1, 2, \dots, n$$

Proof. The Venn-diagram of the theorem is shown in Figure 8.4.

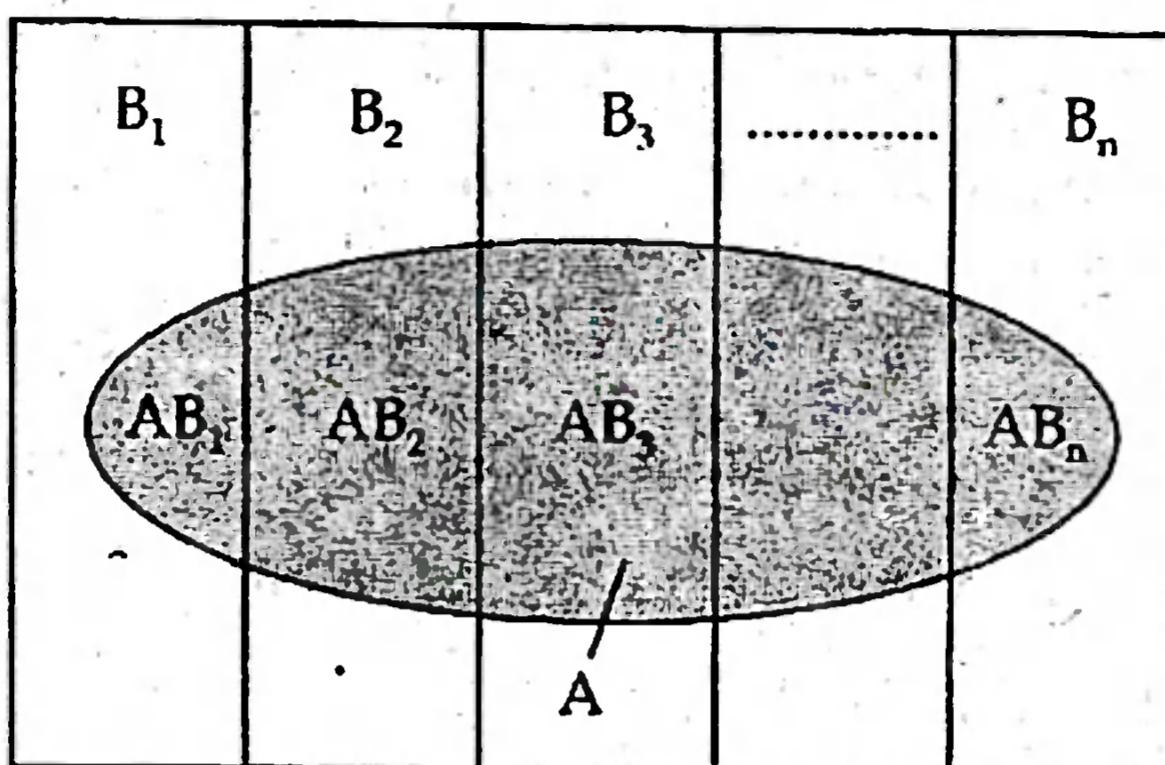


Fig. 8.12.1. Venn-diagram for Bayes theorem.

From the definition of conditional probability, we have

$$P[B_i | A] = \frac{P[AB_i]}{P[A]} \quad (8.9)$$

We know from the theorem of total probabilities,

$$P[A] = \sum_{j=1}^n P[B_j]P[A | B_j] \quad (8.10)$$

From the multiplication law of events, we have

$$P[AB_i] = P[B_i]P[A | B_i] \quad (8.11)$$

Now by putting (8.10) and (8.11) in (8.9), we have

$$P[B_i | A] = \frac{P[B_i]P[A | B_i]}{\sum_{j=1}^n P[B_j]P[A | B_j]}, \quad i = 1, 2, \dots, n$$

This completes the proof.

The probabilities $P[B_1], P[B_2], \dots, P[B_n]$ in Bayes theorem are known as prior probabilities, since they are known to the statistician before the experiment on the basis of some prior knowledge. The probabilities

$P[B_1 | A], P[B_2 | A] \dots P[B_n | A]$ in Bayes theorem are known as posterior probabilities, since they are known after the experiment on the basis of the prior probabilities.

Example 8.12.1. In a factory machine A produces 60% of the output and machine B produces the rest. 1% of the output of machine A is defective and 2% of the output of machine B is defective. An item is selected at random from a day's output and is found to be defective, what is the probability that the defective item was produced by machine B.

Solution. Let us define the following events :

B_1 : An item produced by machine A

B_2 : An item produced by machine B

A : A defective item produced by the machines.

We have $P[B_1] = 60\% = 0.60$, $P[B_2] = 0.40$, $P[A|B_1] = 0.01$, $P[A|B_2] = 0.02$

We have to find $P[B_2|A]$. According to Bayes theorem

$$P[B_2|A] = \frac{P[B_2]P[A|B_2]}{P[A]}$$

According to theorem of total probabilities,

$$\begin{aligned} P[A] &= P[AB_1 \cup AB_2] = P[B_1]P[A|B_1] + P[B_2]P[A|B_2] \\ &= (0.60)(0.01) + (0.40)(0.02) = 0.014. \end{aligned}$$

$$P[B_2|A] = \frac{P[B_2]P[A|B_2]}{P[A]} = \frac{(0.40)(0.02)}{0.014} = \frac{0.008}{0.014} = \frac{8}{14} = \frac{4}{7}.$$

Example 8.12.2. Mr. Rahman wants to buy a car this year. He applied for a bank loan. The probability that he will get the bank loan is $2/3$. If he will get the bank loan, the probability that he will buy a car is $3/4$. If he will not get the bank loan, the probability that he will buy a car is $1/4$. After one year, one fine morning it is seen that Mr. Rahman is driving a new car. What is the probability that Mr. Rahman got the bank loan?

Solution. Let B_1 be the event that Mr. Rahman will get the bank loan, B_2 be the event that he will not get the bank loan and A be the event that he will buy a car this year. We have to find the conditional probability of B_1 for given A. According to Bayes theorem

$$P[B_1|A] = \frac{P[B_1]P[A|B_1]}{P[A]}$$

We are given $P[B_1] = 2/3$, $P[B_2] = 1/3$, $P[A|B_1] = 3/4$ and $P[A|B_2] = 1/4$. According to the theorem of total probabilities

$$\begin{aligned} P[A] &= P[AB_1 \cup AB_2] = P[B_1]P[A|B_1] + P[B_2]P[A|B_2] \\ &= (2/3)(3/4) + (1/3)(1/4) = 7/12. \end{aligned}$$

$$\text{Hence } P[B_1 | A] = \frac{P[B_1]P[A | B_1]}{P[A]} = \frac{(2/3)(3/4)}{7/12} = \frac{6/12}{7/12} = \frac{6}{7}.$$

Matched Problem. Mr. Karim wants to build a house this year. He applied for a bank loan. The probability that he will get the bank loan is $4/5$. If he will get the bank loan the probability that he will build a house is $8/9$. However, if he will not get the bank loan, the probability that he will build a house is $2/9$. After one year Mr. Rahman, a friend of Mr. Karim residing abroad heard that Mr. Karim is living a beautiful new house, what is the probability that Mr. Karim got the bank loan?

Ans. 16/17.

8.13. Some Solved Problems on Probability

Example 8.13.1. Suppose $P[A] = 0.3$ and $P[B] = 0.4$ and $P[AB] = 0.2$. (i) Are A and B independent? (ii) Are A and B mutually Exclusive? (iii) Find $P[A | B]$ (iv) State the nature of the events A and B.

Solution. (i) A and B are independent if

$$P[AB] = P[A]P[B]$$

$$\text{Here } P[A] \times P[B] = 0.3 \times 0.4 = 0.12 \neq 0.2 = P[AB]$$

Hence A and B are not independent.

(ii) A and B are mutually exclusive if $P[AB] = 0$.

$$\text{Here } P[AB] = 0.2 \neq 0$$

Hence A and B are not mutually exclusive.

$$(iii) P[A | B] = \frac{P[AB]}{P[B]} = \frac{0.2}{0.4} = 0.5.$$

Hence A and B are dependent events.

Example 8.13.2. Suppose $P[A] = 0.5$ and $P[B] = 0.6$ and $P[AB] = 0.3$. (i) Are A and B independent? (ii) Find $P[A | B]$, $P[B | A]$ and comment on A and B.

Solution. (i) The events A and B are independent if $P[AB] = P[A]P[B]$

$$\text{Here } P[A]P[B] = (0.5)(0.6) = 0.30 = P[AB]$$

Hence A and B are independent.

$$(ii) P[A | B] = \frac{P[AB]}{P[B]} = \frac{0.3}{0.6} = 0.5.$$

$$P[A | B] = P[A].$$

Here conditional probability is equal to its unconditional. This is true since A and B are independent

$$P[B|A] = \frac{P[AB]}{P[A]} = \frac{0.3}{0.5} = 0.6 = P[B].$$

This happens since A and B are independent.

Example 8.13.3. Let $P[A] = 0.4$, $P[B] = 0.5$. Find (a) $P[\bar{B}]$, (b) Find $P[A \cup B]$ when (i) A and B are mutually exclusive and (ii) A and B are independent.

Solution. (a) We know $P[B] + P[\bar{B}] = 1$

$$\text{Hence } P[\bar{B}] = 1 - P[B] = 1 - 0.5 = 0.5.$$

(i) When A and B are mutually exclusive,

$$P[A \cup B] = P[A] + P[B] = 0.4 + 0.5 = 0.9.$$

(ii) Here the events A and B are independent, $P[AB] = P[A]P[B]$

According to addition law of probability,

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[AB] = P[A] + P[B] - P[A]P[B] \\ &= 0.4 + 0.5 - (0.4)(0.5) = 0.7. \end{aligned}$$

Example 8.13.4. Let $P[A] = 0.6$, $P[B] = 0.8$ and $P[AB] = 0.50$. Find (i) $P[\bar{A}]$, (ii) $P[A \cup B]$, (iii) $P[A \bar{B}]$, (iv) $P[A|B]$ and (v) Are the events A and B independent?

Solution. (i) We know $P[A] + P[\bar{A}] = 1$

$$\text{Here } P[A] = 0.6. \text{ Hence, } P[\bar{A}] = 1 - 0.60 = 0.40.$$

(ii) By additive law of probability, we know

$$P[A \cup B] = P[A] + P[B] - P[AB] = 0.60 + 0.80 - 0.50 = 0.90.$$

$$(iii) P[A \bar{B}] = P[A] - P[AB] = 0.60 - 0.50 = 0.10.$$

(iv) The conditional probability of A given B is defined by

$$P[A|B] = \frac{P[AB]}{P[B]} = \frac{0.50}{0.80} = \frac{5}{8}.$$

(v) The events A and B are independent if

$$P[AB] = P[A]P[B]$$

$$\text{Here } P[A]P[B] = (0.60)(0.80) = 0.48 \neq 0.50 = P[AB]$$

Hence the events A and B are not independent.

Example 8.13.5. Let $P[A] = 0.4$ and $P[A \cup B] = 0.7$. Find $P[B]$ for which (i) A and B are mutually exclusive, (ii) A and B are independent and (iii) $P[A|B] = 0.6$

Solution. (i) If A and B are mutually exclusive,

$$\text{then } P[A \cup B] = P[A] + P[B] = 0.7$$

$$\text{Here } P[A] = 0.4, \text{ then } P[B] = 0.7 - 0.4 = 0.3.$$

(ii) If A and B are independent, then $P[AB] = P[A] P[B]$

$$P[A \cup B] = P[A] + P[B] - P[AB] = P[A] + P[B] - P[A] P[B] = 0.7$$

$$\text{That is, } P[A] + P[B] - P[A] P[B] = 0.7$$

$$0.4 + P[B] - 0.4P[B] = 0.7$$

$$0.6P[B] = 0.3.$$

$$\text{Hence } P[B] = \frac{0.3}{0.6} = \frac{3}{6} = \frac{1}{2}.$$

(iii) Here we have $P[A|B] = 0.6$

$$P[A|B] = \frac{P[AB]}{P[B]} = 0.6 \Rightarrow P[AB] = 0.6 P[B]$$

$$\text{Now, } P[A \cup B] = P[A] + P[B] - P[AB]$$

$$0.7 = 0.4 + P[B] - 0.6 P[B] \Rightarrow 0.4 P[B] = 0.3$$

$$\text{Hence } P[B] = \frac{3}{4}.$$

Example 8.13.6. Suppose $P[A \cup B] = 5/8$, $P[A] = 1/2$ and $P[B] = 1/5$. Find (i) $P[AB]$ and (ii) $P[A|B]$.

Solution. (i) We know, $P[A \cup B] = P[A] + P[B] - P[AB]$

$$\Rightarrow \frac{5}{8} = \frac{1}{2} + \frac{1}{5} - P[AB] \Rightarrow P[AB] = \frac{1}{2} + \frac{1}{5} - \frac{5}{8} = \frac{3}{40}.$$

$$\text{Hence } P[AB] = \frac{3}{40}$$

$$(ii) P[A|B] = \frac{P[AB]}{P[B]} = \frac{\frac{3}{40}}{\frac{1}{5}} = \frac{3}{8}.$$

Example 8.13.7. Three events A, B and C of a random experiment are mutually exclusive and exhaustive. If $P[A] = 2P[B]$ and $P[B] = 2P[C]$. Find the probabilities of A, B and C.

Solution. We have, $P[A \cup B \cup C] = 1$

Since the events A, B and C are exhaustive. The events are also mutually exclusive. Hence $P[A \cup B \cup C] = P[A] + P[B] + P[C] = 1$

$$\Rightarrow 4P[C] + 2P[C] + P[C] = 1;$$

This follows from the condition of the problem.

$$\Rightarrow 7P[C] = 1$$

Hence $P[C] = \frac{1}{7}$, then $P[B] = \frac{2}{7}$ and $P[A] = \frac{4}{7}$.

Example 8.13.18. A study showed that 75% of managers had some business education and 55% had engineering education. Further, 35% of the managers had some business education but no engineering education.

- (i) What is the probability that a manager has some business and engineering education?
- (ii) What is the probability that a manager has some engineering education but no business education?
- (iii) It is known that a manager has some engineering education, what is the probability that he has some business education?

Solution. Let A be the event that the manager has some business education and B be the event that he has some engineering education. Let \bar{A} and \bar{B} be the complementary of the events A and B. Then $P[A] = 0.75$, $P[B] = 0.55$, $P[A \bar{B}] = 0.35$,

$$(i) P[AB] = P[A] - P[A \bar{B}] = 0.75 - 0.35 = 0.40$$

$$(ii) P[\bar{A}B] = P[B] - P[AB] = 0.55 - 0.40 = 0.15$$

$$(iii) P[A|B] = \frac{P[AB]}{P[B]} = \frac{0.40}{0.55} = \frac{8}{11}$$

Example 8.13.9. The personal department of a company has records, which show the following analysis of its 200 engineers.

Age	Bachelor's degree only	Master's degree	Total
Under 30	90	10	100
30 - 40	20	30	50
over 40	40	10	50
Total	150	50	200

An engineer is selected at random from the company. Find the probability that the engineer is a

- (i) Master's degree,
- (ii) Master's degree and age between 30 - 40,
- (iii) Master's degree or age over 40
- (iv) Master's degree given that he is over 40

Solution. Let us define the following events:

- A : an engineer has a master's degree
- B : an engineer has a bachelor's degree
- C : an engineer is between 30-40 years of age
- D : an engineer is over 40 years of age
- E : an engineer is under 30 years of age

We have, $N(S) = 200$, $N(A) = 50$, $N(B) = 150$, $N(C) = 50$, $N(D) = 50$, $N(E) = 100$, $N(AC) = 30$, $N(AD) = 10$ and $N(BE) = 90$.

$$(i) P[A] = \frac{N(A)}{N(S)} = \frac{50}{200} = 0.25.$$

$$(ii) P[AC] = \frac{N(AC)}{N(S)} = \frac{30}{200} = \frac{3}{20}.$$

$$(iii) P[A \cup D] = P[A] + P[D] - P[AD] = \frac{N(A)}{N(S)} + \frac{N(D)}{N(S)} - \frac{N(AD)}{N(S)}$$

$$= \frac{50}{200} + \frac{50}{200} - \frac{10}{200} = \frac{90}{200} = \frac{9}{20}.$$

$$(iv) P[A | D] = \frac{P[AD]}{P[D]} = \frac{N(AD)}{N(D)} = \frac{10}{50} = 0.20.$$

$$(v) P[B | E] = \frac{P[BE]}{P[E]} = \frac{N(BE)}{N(E)} = \frac{90}{100} = 0.90.$$

Example 8.13.10. Suppose two popular models of cars, say A and B are available in the market. A survey was conducted on 1000 people. 500 liked A, 400 liked B, 200 liked both A and B. A person is selected at random from these 1000 people. What is the probability that he liked (i) A or B, (ii) only A, (iii) only one, (iv) neither A or B.

Solution. Let A be the event that a person liked model A and B be the event that a person liked model B. Let \bar{A} and \bar{B} be the complementary of the events A and B respectively.

Here $N(S) = 1000$, $N(A) = 500$, $N(B) = 400$, $N(AB) = 200$

	B	\bar{B}	Total
A	200	300	500
\bar{A}	200	300	500
Total	400	600	1000

$$(i) P[A \cup B] = P[A] + P[B] - P[AB]$$

$$= \frac{N(A)}{N(S)} + \frac{N(B)}{N(S)} - \frac{N(AB)}{N(S)} = \frac{500}{1000} + \frac{400}{1000} - \frac{200}{1000} = \frac{700}{1000} = 0.70.$$

$$(ii) P[A\bar{B}] = \frac{N(A\bar{B})}{N(S)}$$

$$\text{Here } N(A\bar{B}) = N(A) - N(AB) = 500 - 200 = 300.$$

$$P[A\bar{B}] = \frac{N(A\bar{B})}{N(S)} = \frac{300}{1000} = 0.30.$$

(iii) Only one means only A or only B. That is $A\bar{B} \cup \bar{A}B$.

$$N(A\bar{B} \cup \bar{A}B) = N(A\bar{B}) + N(\bar{A}B) = N(A) - N(AB) + N(B) - N(AB)$$

$$= 500 - 200 + 400 - 200 = 500$$

$$\text{Hence } P[A\bar{B} \cup \bar{A}B] = \frac{N(A\bar{B} \cup \bar{A}B)}{N(S)} = \frac{500}{1000} = 0.50.$$

$$(iv) P[\bar{A}\bar{B}] = \frac{N(\bar{A}\bar{B})}{N(S)} = \frac{300}{1000} = 0.30.$$

Example 8.13.11. In a class of 100 students 90 took statistics and 80 took mathematics and 75 took both the courses. A student is selected at random from this class. (a) What is the probability that the selected student took

- (i) at least one course,
- (ii) both the courses
- (iii) only statistics
- (iv) Only one course.

(b) It is given that the student took mathematics, what is the probability that he (i) took statistics, (ii) did not take statistics.

(c) It is given that the selected student did not take statistics, what is the probability that he (i) took mathematics and (ii) did not take mathematics.

Solution. Let A be the event that the selected student took statistics and B be the event that he took mathematics. We have $N(S) = 100$, $N(A) = 90$, $N(B) = 80$ and $N(AB) = 75$.

The possible events and their sample point are shown in the tables given below

	B	\bar{B}
A	AB	$A\bar{B}$
\bar{A}	$\bar{A}B$	$\bar{A}\bar{B}$

a. (i) $A \cup B$ = Selected student took at least one course

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[AB] = \frac{N(A)}{N(S)} + \frac{N(B)}{N(S)} - \frac{N(AB)}{N(S)} \\ &= \frac{90}{100} + \frac{80}{100} - \frac{75}{100} = \frac{95}{100} = 0.95 = 95\%. \end{aligned}$$

That is 95% student took at least one course.

(ii) AB = Selected student took both the courses

$$P[AB] = \frac{N(AB)}{N(S)} = \frac{75}{100} = 0.75 = 75\%.$$

That is 75% student took both the courses.

$$(iii) P[\text{only statistics}] = P[A\bar{B}] = \frac{N(\bar{A}\bar{B})}{N(S)} = \frac{15}{100} = 0.15 = 15\%.$$

15% student took only statistics.

$$(iv) P[\text{only one course}] = P[\bar{A}\bar{B} \cup \bar{A}B] = P[\bar{A}\bar{B}] + P[\bar{A}B]$$

$$= \frac{N(\bar{A}\bar{B})}{N(S)} + \frac{N(\bar{A}B)}{N(S)} = \frac{15}{100} + \frac{5}{100} = \frac{20}{100} = 0.20 = 20\%.$$

$$b. (i) P[A|B] = \frac{P[AB]}{P[B]} = \frac{N(AB)}{N(B)} = \frac{75}{80} = \frac{15}{16}.$$

$$(ii) P[\bar{A}|B] = \frac{P[\bar{A}B]}{P[B]} = \frac{N(\bar{A}B)}{N(B)} = \frac{5}{80} = \frac{1}{16}.$$

$$c. (i) P[B|\bar{A}] = \frac{P[\bar{A}B]}{P[\bar{A}]} = \frac{N(\bar{A}B)}{N(\bar{A})} = \frac{5}{10} = 0.5.$$

$$(ii) P[\bar{B}|\bar{A}] = \frac{P[\bar{A}\bar{B}]}{P[\bar{A}]} = \frac{N(\bar{A}\bar{B})}{N(\bar{A})} = \frac{5}{10} = 0.5.$$

Example 8.13.12. Employee Benefits. In a city, there are 1200 business firms. According to a survey it is seen that 760 firms offer their employee health insurance, 650 offer dental insurance, and 285 offer health insurance and dental insurance. A firm is selected at random from the city. (a) What is the probability that the selected firm offers

- (i) at least one insurance,
- (ii) only one insurance,
- (iii) only health insurance.

(b) It is known (given) that the firm offers health insurance, what is the probability that

- (i) It offers dental insurance.
- (ii) It does not offer dental insurance.

Solution. Let H be the event that a firm offers health insurance and D be the event that a firm offers dental insurance. Then $N(S) = 1200$, $N(H) = 760$, $N(D) = 650$ and $N(HD) = 285$. $H \cup D$ = at least one insurance, $H\bar{D} \cup \bar{H}D$ = only one insurance, $\bar{H}\bar{D}$ = only health insurance. The Venn-diagram of the problem is

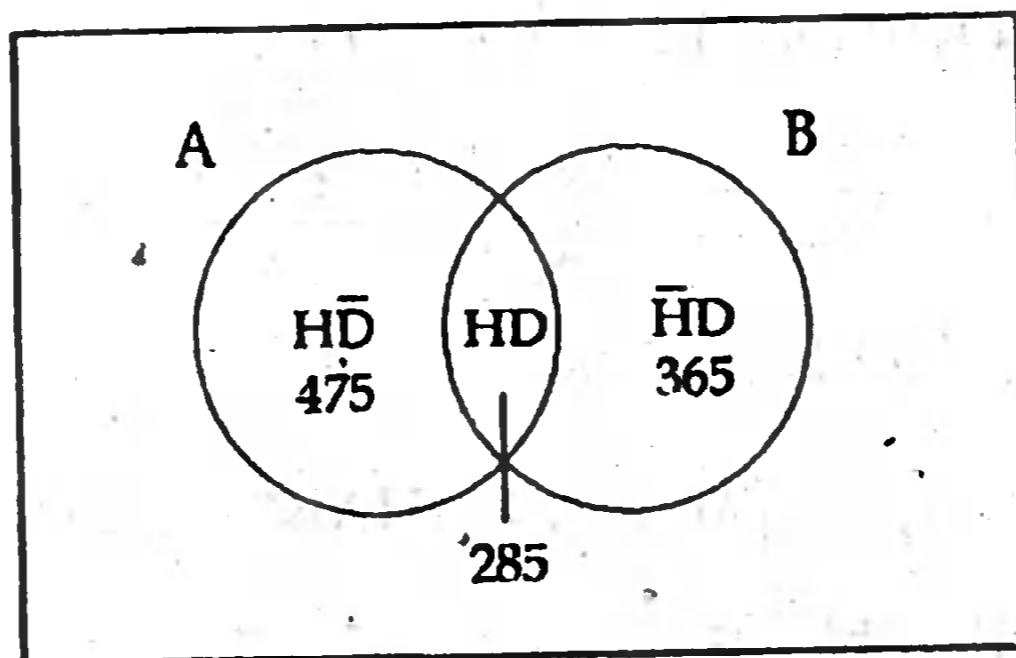


Fig. 8.13.1. Venn-diagram.

a. (i) According to additive law of events, we have

$$\begin{aligned} P[H \cup D] &= P[H] + P[D] - P[HD] \\ &= \frac{N(H)}{N(S)} + \frac{N(D)}{N(S)} - \frac{N(HD)}{N(S)} = \frac{760}{1200} + \frac{650}{1200} - \frac{285}{1200} = \frac{1125}{1200} = \frac{15}{16}. \end{aligned}$$

$$(ii) P[H\bar{D} \cup \bar{H}D] = P[H\bar{D}] + P[\bar{H}D]$$

Since $H\bar{D}$ and $\bar{H}D$ are mutually exclusive.

$$N(H\bar{D}) = N(H) - N(HD) = 760 - 285 = 475$$

$$N(\bar{H}D) = N(D) - N(HD) = 650 - 285 = 365$$

$$\text{Hence, } P[H\bar{D} \cup \bar{H}D] =$$

$$P[\bar{H}\bar{D}] + P[\bar{H}\bar{D}] = \frac{N(\bar{H}\bar{D})}{N(S)} + \frac{N(\bar{H}\bar{D})}{N(S)} = \frac{475}{1200} + \frac{365}{1200} = \frac{840}{1200} = \frac{7}{10}$$

$$(iii) P[\text{Only health}] = P[\bar{H}\bar{D}] = \frac{N(\bar{H}\bar{D})}{N(S)} = \frac{475}{1200} = \frac{19}{48}$$

$$b. (i) P[D|H] = \frac{P[DH]}{P[H]} = \frac{N(DH)}{N(H)} = \frac{285}{760} = \frac{57}{152}$$

$$(ii) P[\bar{D}|H] = \frac{P[\bar{D}H]}{P[H]} = \frac{N(\bar{D}H)}{N(H)} = \frac{475}{760} = \frac{95}{152}$$

$$\text{As a check, } P[D|H] + P[\bar{D}|H] = 1.$$

Example 8.13.13. In a city, 60% of the people moves by bus, 25% by rickshaw, 15% by car, 3% of the accidents committed by bus, 5% by rickshaw and 1% by car. A person of the city falls in an accident. What is the probability that the accident has committed by rickshaw?

Solution. Let us define the following events:

- B₁: a person moves by bus
- B₂: a person moves by rickshaw
- B₃: a person moves by car
- A : a person falls in an accident

We have $P[B_1] = 0.60$, $P[B_2] = 0.25$, $P[B_3] = 0.15$,

$P[A|B_1] = 0.03$, $P[A|B_2] = 0.05$, $P[A|B_3] = 0.01$

We have to find $P[B_2|A]$. According to Bayes theorem

$$P[B_2|A] = \frac{P[B_2]P[A|B_2]}{P[A]}$$

$$\begin{aligned} P[A] &= P[AB_1 \cup AB_2 \cup AB_3] \\ &= P[B_1]P[A|B_1] + P[B_2]P[A|B_2] + P[B_3]P[A|B_3] \\ &= (0.60)(0.03) + (0.25)(0.05) + (0.15)(0.01) \\ &= 0.0180 + 0.0125 + 0.0015 = 0.0320 \end{aligned}$$

$$P[B_2|A] = \frac{P[B_2]P[A|B_2]}{P[A]} = \frac{(0.25)(0.05)}{0.032} = \frac{0.0125}{0.0320} = \frac{125}{320}$$

Example 8.13.14. A fan assembler uses motors from two suppliers. Company A supplies 60% and company B supplies the rest. It is known from previous experiences that 2% of the motors supplied by company A are defective and 3% of the motors supplied by company B are defective.

An assembled fan is found to have a defective motor, what is the probability that company A supplied this motor.

Solution. Let us define the following events:

- B_1 : company A supplies motor
- B_2 : company B supplies motor
- A : the company supplies a defective motor

We have $P[B_1] = 0.60$, $P[B_2] = 0.40$,

$$P[A | B_1] = 0.02, P[A | B_2] = 0.03$$

$$\text{According to Bayes theorem } P[B_1 | A] = \frac{P[B_1]P[A | B_1]}{P[A]}$$

$$\begin{aligned} P[A] &= P[AB_1 \cup AB_2] = P[B_1]P[A | B_1] + P[B_2]P[A | B_2] \\ &= (0.60)(0.02) + (0.40)(0.03) = 0.012 + 0.012 = 0.024. \end{aligned}$$

$$P[B_1 | A] = \frac{P[B_1]P[A | B_1]}{P[A]} = \frac{(0.60)(0.02)}{0.024} = \frac{0.012}{0.024} = \frac{12}{24} = 0.5.$$

Example 8.13.15. (Man is not free from mistakes) In a bank 45% and 55% of the monthly statements are prepared by Mrs Ali and Miss Karim respectively. These employees are very reliable. However, they are in error sometimes. The probabilities of committing their errors are 0.05% and 0.01% respectively. A monthly statement was found to be erroneous, what is the probability that it was done by Miss. Karim?

Solution. Let us define the following events:

- B_1 : monthly statement prepared by Mrs. Ali
- B_2 : monthly statement prepared by Miss. Karim
- A : a statement prepared by them

We have $P[B_1] = 0.45$, $P[B_2] = 0.55$,

$$P[A | B_1] = 0.0005, P[A | B_2] = 0.0001$$

We have to find $P[B_2 | A]$.

$$\text{According to Bayes theorem } P[B_2 | A] = \frac{P[B_2]P[A | B_2]}{P[A]}$$

$$\begin{aligned} P[A] &= P[B_1]P[A | B_1] + P[B_2]P[A | B_2] \\ &= (0.45)(0.0005) + (0.55)(0.0001) = 0.000225 + 0.000055 = 0.00028. \end{aligned}$$

$$P[B_2 | A] = \frac{P[B_2]P[A | B_2]}{P[A]} = \frac{(0.45)(0.0005)}{0.00028} = \frac{0.000055}{0.00028} = \frac{55}{280} = \frac{11}{56}.$$

Example 8.13.16. (Machines do mistakes) A factory produces certain types of output by three machines. The respective daily production figures are machine A = 3000 units, Machine B = 2500 units, Machine C = 4500 units. Past experience shows that 2 percent of the output produced by Machine A, 1.2 percent by Machine B and 1 percent by Machine C are defective. An item is drawn at random from a day's production run and is found to be defective. What is the probability that it comes from the output of (i) Machine A; (ii) Machine B and (iii) Machine C?

Solution. Let us define the following events:

- B₁ : an unit is produced by machine A
- B₂ : an unit is produced by machine B
- B₃ : an unit is produced by machine C
- A : a defective unit is produced by the machines

Here, we have N(S) = 10,000, N(B₁) = 3000, N(B₂) = 2500, N(B₃) = 4500

$$\text{We have } P[B_1] = \frac{N(B_1)}{N(S)} = \frac{3000}{10000} = 0.30, P[B_2] = \frac{N(B_2)}{N(S)} = \frac{2500}{10000} = 0.25,$$

$$P[B_3] = \frac{N(B_3)}{N(S)} = \frac{4500}{10000} = 0.45,$$

$$P[A | B_1] = 0.02, P[A | B_2] = 0.012, P[A | B_3] = 0.01$$

We have to find P[B₁ | A], P[B₂ | A] and P[B₃ | A].

$$(i) \text{ According to Bayes theorem } P[B_1 | A] = \frac{P[B_1]P[A | B_1]}{P[A]}$$

$$\begin{aligned} P[A] &= P[B_1]P[A | B_1] + P[B_2]P[A | B_2] + P[B_3]P[A | B_3] \\ &= (0.30)(0.02) + (0.25)(0.012) + (0.45)(0.01) \\ &= 0.006 + 0.003 + 0.0045 = 0.0135. \end{aligned}$$

$$P[B_1 | A] = \frac{P[B_1]P[A | B_1]}{P[A]} = \frac{(0.30)(0.02)}{0.0135} = \frac{0.0060}{0.0135} = \frac{60}{135}.$$

$$(ii) P[B_2 | A] = \frac{P[B_2]P[A | B_2]}{P[A]} = \frac{(0.25)(0.012)}{0.0135} = \frac{0.0030}{0.0135} = \frac{30}{135}.$$

$$(iii) P[B_3 | A] = \frac{P[B_3]P[A | B_3]}{P[A]} = \frac{(0.45)(0.01)}{0.0135} = \frac{0.0045}{0.0135} = \frac{45}{135}.$$

Example 8.13.17. A piece of equipment will function only when all the three components A, B, and C are working. The probability of A failing during one year is 0.15, that of B failing is 0.05 and that of C failing is 0.10. What is the probability that the equipment will fail before the end of the year?

Solution. Let A be the event that the component A does not fail during one year.

Let B be the event that the component B does not fail during one year.

Let C be the event that the component C does not fail during one year.

\bar{A} , \bar{B} and \bar{C} are the complementary events of A, B and C respectively.

Here $P[\bar{A}] = 0.15$ then $P[A] = 1 - 0.15 = 0.85$.

$P[\bar{B}] = 0.05$ then $P[B] = 1 - 0.05 = 0.95$

$P[\bar{C}] = 0.10$ then $P[C] = 1 - 0.10 = 0.90$

Hence $P[ABC] = P[A] P[B] P[C] = (0.85)(0.95)(0.90) = 0.727$.

Hence the probability that the equipment will fail before the end of the year is

$$1 - 0.727 = 0.272.$$

Example 8.13.18. Consumer Survey. If 60% customers of a departmental store are female and 75% of the female customers have charge accounts at the store, while 80% of the male customers have charge accounts at the store. Draw a tree diagram and find the probability that a customer selected at random is

- i) a charge account holder given that she is a female
- ii) a female and has a charge account ?
- iii) a female given that she has no charge account.

Solution. Let S = All store customers

F = Female customers

M = Male customer

C = Customers with charge account

We have $P[F] = 0.60$. $P[M] = 1 - 0.60 = 0.40$.

The probabilities of different types of events are shown in following tree diagram

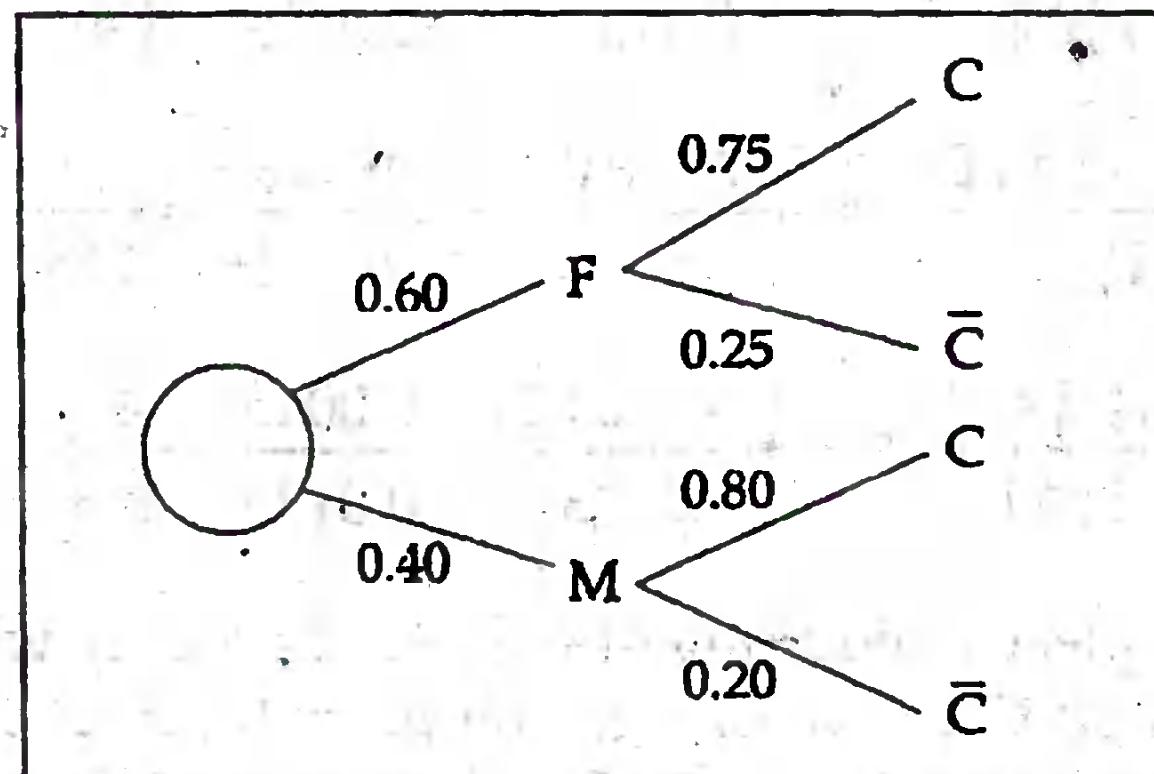


Fig. 8.13.2. Tree diagram of different types of customers.

- (i) Since 75% of the female customers have charge accounts, the probability that a customer has a charge account, given that the customer is a female, is

$$P[C|F] = 0.75.$$

- (ii) The probability that a customer is a female and has a charge account is given by

$$P[FC] = P[F] P[C|F] = (0.60)(0.75) = 0.45.$$

- (iii) The probability that the selected customer is a female given that she has no charge account is a case of reverse conditional probability and to compute the required probability we have to use Bayes theorem, which is given by

$$\begin{aligned} P[F|\bar{C}] &= \frac{P(F \cap \bar{C})}{P(\bar{C})} = \frac{P(F)P(\bar{C}|F)}{P(F)P(\bar{C}|F) + P(M)P(\bar{C}|M)} \\ &= \frac{0.60 \times 0.25}{0.60 \times 0.25 + 0.40 \times 0.20} = 0.65. \end{aligned}$$

Example 8.13.19. Market Research From a survey involving 1,000 people in a certain city, it was found that 500 people had tried a certain brand of diet cola, 600 had tried a certain brand of regular cola, and 200 had tried both types of cola. If a resident of the city is selected at random, (a) what is the probability that the resident has tried

- (i) the diet or regular cola?
- (ii) only diet cola but not regular cola?
- (iii) only one cola ?

- (b) It is known that the resident has tried diet cola, what is the probability that the resident (i) has tried regular cola and (ii) has not tried regular cola.

Solution. Let D is the event that a person has tried the diet cola and R is the event that a person has tried the regular cola. Let \bar{D} and \bar{R} be the complementary of the events D and R. The information can be presented by the following table :

	R	\bar{R}	Total
D	$N(DR) = 200$	$N(D\bar{R}) = 300$	500
\bar{D}	$N(\bar{D}R) = 400$	$N(\bar{D}\bar{R}) = 100$	500
Total	600	400	1000

Here $N(S) = 1,000$; $N(D) = 500$; $N(R) = 600$; $N(DR) = 200$; $N(D\bar{R}) = 300$; $N(\bar{D}R) = 400$; $N(\bar{D}\bar{R}) = 100$.

a. (i) The probability that the selected person has tried diet or regular cola is

$$\begin{aligned} P[D \cup R] &= P[D] + P[R] - P[DR] \\ &= \frac{N(D)}{N(S)} + \frac{N(R)}{N(S)} - \frac{N(DR)}{N(S)} = \frac{500}{1000} + \frac{600}{1000} - \frac{200}{1000} = \frac{900}{1000} = 0.90. \end{aligned}$$

$$(ii) P[D \bar{R}] = \frac{N(D\bar{R})}{N(S)} = \frac{300}{1000} = 0.30.$$

$$(iii) P[\bar{D} \cup \bar{R}] = P[\bar{D}] + P[\bar{R}]$$

$$= \frac{N(\bar{D})}{N(S)} + \frac{N(\bar{R})}{N(S)} = \frac{300}{1000} + \frac{400}{1000} = \frac{700}{1000} = 0.70.$$

$$b. (i) P[D|R] = \frac{P[DR]}{P[R]} = \frac{N(DR)}{N(R)} = \frac{200}{600} = \frac{1}{3}.$$

$$(ii) P[D | \bar{R}] = \frac{P[\bar{D}\bar{R}]}{P[\bar{R}]} = \frac{N(\bar{D}\bar{R})}{N(\bar{R})} = \frac{300}{400} = 0.75.$$

Questions

- Define the following with examples: Random experiment, sample space, sample point, event, sure event, impossible event, uncertain event, simple event and compound event.
- What are different approaches of defining probability? Define classical probability. State some of its drawbacks.
- Define mutually exclusive and independent event. Can two events be mutually exclusive and independent simultaneously? Justify your answer.
- State and prove additive law of probability for three events A, B and C.
- Define conditional probability. State and prove multiplicative law of probability.
- What do you mean by complementary event? Distinguish between a simple event and a compound event.
- State and prove Bayes theorem.
- What is a tree diagram? Three fair coins are tossed. Construct the sample space with the help of a tree diagram.

Exercise

- Toss a coin and a die. Construct the sample space of the experiment.
- If $P[A] = 0.5$, $P[B] = 0.4$ and $P[A \cup B] = 0.7$, then find (i) $P[AB]$, and P $[A \bar{B}]$

Ans. (i) 0.2 and (ii) 0.3

11. If $P[A] = 0.5$, $P[B] = 0.7$ and $P[AB] = 0.4$. Find (i) $P[\bar{A}]$, (ii) $P[A\bar{B}]$, (iii) $P[A \cup B]$, (iv) $P[A|B]$ and (v) $P[B|A]$, (vi) Are A and B independent?
12. Three horses A, B and C are in a race. A is twice as likely to win as B, and B is twice as likely to win as C. What are their respective chances of winning? Ans. $4/7, 2/7$ and $1/7$.
13. Toss a fair coin three times and write down the sample space. Suppose, A be the event that first coin shows head and B be the event that second coin shows tail. Check whether the events A and B are independent or not. Ans. Independent
14. The events A and B are such that $P[A] = 0.3$, $P[B] = 0.25$ and $P[A \cup B] = 0.5$
- Show that A and B are independent
 - Represent these probabilities in a Venn diagram
 - Find $P[A|\bar{B}]$ Ans. c. 0.67
15. Events A, B and C are defined in the sample space S such that $P[A] = 0.4$, $P[B] = 0.2$, $P[A \cap C] = 0.04$, and $P[B \cup C] = 0.44$. The events A and B are mutually exclusive, and B and C are independent.
- Draw a Venn diagram to illustrate the relationship between the three events and the sample space.
 - Find $P[C]$, $P(B \cap C)$, $P(B|C)$ Ans. 0.3, 0.06, 0.2

Applications

16. In a high school graduating class of 100 students 54 studied mathematics, 69 studied history, and 35 studied both mathematics and history. If one of these students is selected at random, find the probability that
- the student takes mathematics or history;
 - the student does not take either of these subjects;
 - the student takes history but not mathematics.
- Ans. 0.88, 0.12, 0.34
17. Employee Benefits. A survey was conducted on 550 business firms of a city. It was found that 345 firms offer their employees group life insurance, 285 offer long-term disabilities insurance, and 115 offer group life insurance and long-term disability insurance. A firm is selected at random from this city. What is the probability that the selected offers
- (i) at least one insurance, (ii) only life insurance, (iii) only one insurance.
 - It is given that the firm offers life insurance what is the probability (i) it offers disability insurance, (ii) it does not offer disability insurance?
- Ans. a(i) $103/110$, (ii) $23/55$, (iii) $8/11$, b(i) $23/69$, (ii) $46/69$.

18. The probability that a contractor will get a building contract is $2/3$ and the probability that he will not get an electric contract is $5/9$. If the probability of getting at least one contract is $4/5$. What is the probability that he will get both the contracts? Ans. 14/45
19. Two hundred students were classified according to their faculty (science and commerce) and sex. Among them 120 are male and 140 students belong to science faculty. Among the male students 84 belong to science. A student is selected at random from the class.
- What is the probability that the selected student is (i) male, (ii) female and belongs to science faculty.
 - It is known that the selected student belongs to commerce, find the probability that the student is (i) male and (ii) female.
- Ans. a (i) $3/5$, (ii) $7/25$, b(i) $3/5$, (ii) $2/5$.
20. In a city 55% of the population is male. It is known that 20% of the male and 30% of the female is unemployed. A research student studying the employment situation selects a person at random; find the probability that the student is a male given that he is unemployed.
21. In a class of 100 students 75 play football, 50 play cricket and 40 play both the games. A student is selected at random from this class. What is the probability that the selected student (i) plays only cricket but not football? (ii) Plays at least one of the games? (iii) Does not play any of the games?
- Ans. (i) 0.10, (ii) 0.85, (ii) 0.1
22. **Market Research.** From a survey involving 1,000 students at a large university, a market research company found that 750 students owned houses, 350 student owned cars and 250 owned houses and cars. If a student at the university is selected at random, what is the probability that:
- The selected student owns a car or a house? Ans. 0.85
 - The selected student owns neither a house nor a car. Ans. 0.1
 - It is known that the selected student owns a house, what is the probability that he owns a car. Ans. 1/3.
23. **Machine is not free from mistakes.** A manufacturing firm produces steel pipes in three plants with daily production of 500, 1,000 and 2,000 units respectively. According to past experience it is known that the defective outputs produced by the three plants are respectively 0.005, 0.004 and 0.010. If a pipe is selected at random from a day's total production and found to be defective, what is the probability that the first plant produced the selected pipe? Ans. 0.0897.
24. The personal manager of a large manufacturing firm finds that 15% of the firm's employees are junior executive and 25% of the firm's employees are MBAs. He also discovers that 5% of the firm's employees are both junior executives and MBAs. An executive is selected from the firm and is known that he is an MBA, what is the probability that he is a junior executive. Ans. 0.20.

25. **Product Defects.** In a bolt factory, machines A, B and C manufacture respectively 25%, 30% and 40% of the product. Of the total of their output 5%, 4%, and 2% are defective bolts respectively. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that machine i) A, ii) B and iii) C manufactured it.

Ans. 0.37, 0.40 and 0.23.

26. Asma, Rokeya and Salma are three sisters. They do the family job 40%, 30% and 30% respectively. The probability that one dish will be broken by Asma when she is washing them is 0.02, for Rokeya and Salma the probabilities are 0.03 and 0.02 respectively. One night the parent hears one break but they don't know who was washing them. What is the probability that Asma was washing them?

Ans. 8/23.

27. **Product defects.** A manufacturer obtains clock-radios from three different subcontractors: 20% from A, 40% from B and 40% from C. The defective rates for these subcontractors are 1%, 3%, and 2% respectively. Draw a tree diagram to represent this information. If a defective clock-radio is returned by a customer, find the probability that it came from subcontractor A? From B? From C.

Ans. .091, .545, .364

28. There are 24 ice creams in a shop of which 10 are lemon and rest are orange. Monmoy chooses an ice cream at random and eats it, followed by another one. Find the probability that Monmoy eats
 a. Two lemon ice creams
 b. One lemon and one orange ice cream

Ans. a. 0.163, b. 0.507

29. Two brands of ice cream, say A and B are available in the market. A survey was conducted on 500 customers. 200 liked A, 150 liked B, 40 liked both A and B. A customer is selected at random from these 500 customers. What is the probability that s/he liked (i) A or B, (ii) only A, (iii) only one, (iv) neither A or B.

30. Suppose the following table is constructed from a survey on employees attitude towards the management of a company:

Sex of employee	Attitude toward management		
	Good	Fair	Bad
Male	20	35	45
Female	36	44	20

If an employee is selected at random from the company, find the probability that

- i) The employee is a female.
- ii) Opinion the employee toward management is good.
- iii) The employee is a male with bad attitude towards management.
- iv) The employee's attitude is fair given that the person is a female.

CHAPTER - 9

RANDOM VARIABLE

9.1. Introduction

It is not rare that the results of random experiments are expressed in terms of numerical values. For example, number of heads in case of tossing two coins, number of defective items among three items produced by a machine, number of boys in case of three children family, etc. The number of heads, number of defective items, and number of children are expressed in numerical quantities. The specific values of these quantities depend on the outcome of the random experiments and cannot be predicted before the experiment. These variable quantities, which vary from outcomes to outcomes, are known as random variables. Random variables and their distributions called, probability distributions, play an important role in business decision-making. The expected values of these random variables, also play important role in business decision making. In this chapter, we shall discuss random variable and its expectation.

9.2. Random Variable

Definition. A variable whose values are determined by the outcomes of a random experiment is called random variable.

Definition . A random variable is a function that assigns a real number to each sample point in a sample space.

We shall use capital letters, such as X, Y, Z etc. to denote random variables and the corresponding small letters x, y, z to denote any specific value of a random variable. Like variable, there are two types of random variable. They are

- i) Discrete random variable and
- ii) Continuous random variable.

9.3. Discrete Random Variable

Definition. A discrete random variable is a function that assigns a numerical value to each sample point in a discrete sample space.

That is, a variable, which takes different discrete values with specified probabilities, is called discrete random variable. Now we shall cite one example.

Example 9.3.1. Toss a fair coin twice. Construct the sample space and define a random variable.

Solution. Let H and T denote the head and tail of the coin respectively. The sample space of the experiment is $S = \{ TT, TH, HT, HH \}$

Let X denotes the number of heads. Then X can take values 0, 1 and 2, which are the quantities determined by the outcomes of the random experiment. Here $P[H] = P[T] = 1/2$, since the coins are fair. Let $p(x)$ be the probability that the random variable X takes value x . Here X , number of heads is a random variable. The sample points of the experiment, number of heads and the probabilities of the number of heads are presented in a tabular form given below:

Sample point	Number of heads	Probability of X
TT	0	1/4
TH	1	1/4
HT	1	1/4
HH	2	1/4
Total		1

It is seen that the number of heads equals to 1 has probability $1/2$. The different values of X with the probabilities will be as follows:

Values of $X : x$	0	1	2	Total
$p(x)$	1/4	1/2	1/4	1

Here it is seen that the value of $p(x)$ is greater than 0 and the sum of $p(x)$ is 1. Here $p(x)$ is the probability that X takes value x . That is $P[X = x] = p(x)$.

It is seen that (i) $p(x) > 0$ and (ii) $\sum p(x) = 1$.

The set of ordered pairs $(x, p(x))$ is called probability function or probability mass function. It is also called probability distribution.

Usually discrete random variable takes integer value. But sometimes it can take fractional isolated values too.

9.3.1. Probability function, probability mass function or probability distribution. The probability function of a discrete random variable can be shown in a table or by a formula.

A table or a formula by which the different values of a random variable with their associated probabilities are shown is called discrete probability distribution. Table 9.1 is an example of a probability distribution.

Let X be a random variable which can take values x_1, x_2, \dots, x_n with associate probabilities $p(x_1), p(x_2), \dots, p(x_n)$, then the probability function of X can be defined by the following table:

Table 9.1. Discrete probability distribution

Variable X : x	x_1	x_2	x_n
$p(x)$	$p(x_1)$	$p(x_2)$	$p(x_n)$

Probability function of a random variable X can also be expressed by a formula. Suppose X is a discrete random variable, then the probability function of X may be defined by the formula

$$p(x) = \binom{4}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{4-x}; x = 1, 2, 3, 4.$$

Definition. The set of ordered pairs $(x, p(x))$ is called probability function, probability mass function or probability distribution of the discrete random variable X, if for each value of x

- i) $p(x) \geq 0,$
- ii) $\sum p(x) = 1$ and
- iii) $P[X=x] = p(x).$

Example 9.3.2. Suppose that three items are selected at random from a manufacturing process. Each item is inspected and classified as defective D, or non-defective N. Let X denotes the number of defective items. Suppose the probability of a defective item is 0.02. (i) Find the probability function of X. Also compute, (ii) $P[X > 2]$, (iii) $P[X \geq 2]$ and (iv) $P[X = 2].$

Solution. (i) The sample space of the experiment is

$$S = \{ NNN, NND, NDN, DNN, NDD, DND, DDN, DDD \}$$

Here $P[D] = 0.02$, then $P[N] = 1 - P[D] = 1 - 0.02 = 0.98.$

It is easily seen that the possible values of X are 0, 1, 2, 3

$$\begin{aligned} P[NNN] &= P[X=0] = P(0) = P[N] P[N] P[N] \\ &= (0.98)(0.98)(0.98) = 0.941192. \end{aligned}$$

$$\begin{aligned} P[NND] &= P[NDN] = P[DNN] = P[X=1] = p(1) \\ &= P[N] P[N] P[D] = (0.98)(0.98)(0.02) = 0.19208 \end{aligned}$$

$$\begin{aligned} P[NDD] &= P[DND] = P[DDN] = P[X=2] = p(2) \\ &= P[N] P[D] P[D] = (0.98)(0.02)(0.02) = 0.000392 \end{aligned}$$

$$\begin{aligned} P[DDD] &= P[X=3] = p(3) = P[D] P[D] P[D] \\ &= (0.02)(0.02)(0.02) = 0.000008. \end{aligned}$$

The probability function of X is

Values of X : x	0	1	2	3
$p(x)$	0.941192	0.19208	0.000392	0.000008

The probability function of the random variable X can also be defined as

$$P[X=x] = p(x) = \binom{4}{x} (0.02)^x (0.98)^{3-x}; x = 0, 1, 2, 3.$$

$$(ii) P[X > 2] = P[X = 3] = P(3) = 0.000008.$$

$$\begin{aligned} (iii) P[X \geq 2] &= P[X = 2] + P[X = 3] = p(2) + P(3) \\ &= 0.001176 + 0.000008 = 0.001184. \end{aligned}$$

$$(iv) P[X = 2] = P(2) = 0.001176.$$

Example 9.3.3. Suppose the different sizes of shoes sold by a departmental store in the last month are as follows:

Sizes of Shoes	5	5.5	6	6.5	7	7.5	8	8.5	9
Number of pairs sold	25	30	95	105	155	210	170	125	85

(i) Find the probability function of sizes of the shoes sold by the departmental store. Compute the probability of the sizes of the shoes sold by the shop (ii) more than 7, (iii) less than 6, (iv) between 6 and 8.

Solution. (i) Here the sizes of the shoes may be considered as a random variable X . The possible values of X are 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9 and their relative frequencies are 0.025, 0.030, 0.095, 0.105, 0.155, 0.210, 0.170, 0.125, and 0.085. The relative frequencies for the different sizes of the shoes are considered as their respective probabilities. Then the possible values of X and their probabilities constitute the probability distribution of X and is given by the following table:

Table 9.3. Probability distribution of the sizes of the shoes

Sizes of Shoes $X : x$	5	5.5	6	6.5	7	7.5	8	8.5	9
$p(x)$	0.025	0.030	0.095	0.105	0.155	0.210	0.170	0.125	0.085

$$\begin{aligned} (ii) P[X > 7] &= P[X = 7.5] + P[X = 8] + P[X = 8.5] + P[x = 9] \\ &= p(7.5) + p(8) + p(8.5) + p(9) \\ &= 0.210 + 0.170 + 0.125 + 0.085 = 0.615. \end{aligned}$$

$$\begin{aligned} (iii) P[X < 7] &= P[X = 6.5] + P[X = 6] + P[X = 5.5] + P[X = 5] \\ &= p(6.5) + p(6) + p(5.5) + p(5) \\ &= 0.105 + 0.095 + 0.030 + 0.025 = 0.255. \end{aligned}$$

$$\begin{aligned} (iv) P[6 < X < 8] &= P[X = 6.5] + P[X = 7] + P[7.5] = p(6.5) + p(7) + p(7.5) \\ &= 0.105 + 0.155 + 0.210 = 0.47. \end{aligned}$$

Example 9.3.4. Suppose a lot contains 4 items. Let the random variable X denote the number of defective items in the lot. Suppose the probability function of the random variable X is

Values of $X : x$	0	1	2	3
$p(x)$	0.50	0.3	0.15	0.05

Find (i) $P[X < 1]$, (ii) $P[X \leq 1]$, (iii) $P[1 \leq X < 3]$

Solution. (i) $P[X < 1] = P[X = 0] = p(0) = 0.50$.

(ii) $P[X \leq 1] = P[X = 0] + P[X = 1] = p(0) + p(1) = 0.50 + 0.3 = 0.8$.

(iii) $P[1 \leq X < 3] = P[X = 1] + P[X = 2] = p(1) + p(2) = 0.30 + 0.15 = 0.45$.

Example 9.3.5. A random variable has the following probability function:

Values of $X : x$	-2	-1	0	1	2	3
$p(x)$	0.1	K	0.2	$2k$	0.3	0.1

(i) Find the value of k.

(ii) $P[X > 1]$, (iii) $P[-1 < X < 2]$, (iv) $P[X < 1]$.

Solution. (i) Since $\sum p(x) = 1$, we have

$$0.1 + k + 0.2 + 2k + 0.3 + 0.1 = 1$$

$$\Rightarrow 0.7 + 3k = 1$$

$$\Rightarrow 3k = 0.3$$

$$\Rightarrow k = 0.1$$

(ii) $P[X > 1] = P[X = 2] + P[X = 3] = p(2) + p(3) = 0.3 + 0.1 = 0.4$

(iii) $P[-1 < X < 2] = P[X = 0] + P[X = 1] = p(0) + p(1) = 0.2 + 0.2 = 0.4$

(iv) $P[X < 1] = P[X = 0] + P[X = -1] + P[X = -2]$

$$= p(0) + p(-1) + p(-2) = 0.2 + 0.1 + 0.1 = 0.4$$

Example 9.3.6. Consider the sample space with three children family. Suppose the probability of a boy is $4/7$. Let X be the number of boys. Compute the probability function of the number of boys. A family with three children is selected at random, what is the probability that the family contains (i) exactly 2 boys, (ii) at most two boys, (iii) at least two boys and (iv) no boys.

Solution. B denotes a boy and G denotes a girl. The sample space of the experiment is

$$S = \{GGG, GGB, GBG, BGG, GBB, BGB, BBG, BBB\}$$

Let X denotes the number of boys.

Here, $P[B] = 4/7$, then $P[G] = 1 - 4/7 = 3/7$.

$$P[GGG] = P[X = 0] = p(0) = P[G] P[G] P[G] = (3/7)(3/7)(3/7) = 27/343$$

Since the events are independent.

$$P[GGB] = P[GBG] = P[BGG] = P[B] P[G] P[G]$$

Here $P[GGB] = (3/7)(3/7)(4/7) = 36/343$

So, $P[X = 1] = p(1) = 36/343 + 36/343 + 36/343 = 108/343$.

$$P[GBB] = P[BGB] = P[BBG] = P[B]P[B]P[G]$$

Here, $P[GBB] = (3/7)(4/7)(4/7) = 48/343$

$$P(X=2) = p(2) = 48/343 + 48/343 + 48/343 = 144/343.$$

$$p(3) = P[BBB] = P[B]P[B]P[B] = (4/7)(4/7)(4/7) = 64/343.$$

The probability function of X is

Values of $X : x$	0	1	2	3
$p(x)$	27/343	108/343	144/343	64/343

(i) $P[\text{exactly two boys}] = P[X = 2] = p(2) = 108/343$.

(ii) $P[\text{at most two boys}] = P[X \leq 2] = P[X = 0] + P[X = 1] + P[X = 2]$

$$= p(0) + p(1) + p(2) = 27/343 + 108/343 + 144/343 = 279/343.$$

$$\text{Or, } P[X \leq 2] = 1 - P[X = 3] = 1 - 64/343 = 279/343.$$

(iii) $P[\text{at least two boys}] = P[X \geq 2] = P[X = 3] = p(3) = 64/343$.

(iv) $P[\text{no boys}] = P[X = 0] = p(0) = 27/343$.

A matched problem

Consider three children in which the probability of a boy is 0.50. Let X be the number of girls. Write probability function of the number of girls. What is the probability that a randomly selected family contains (i) exactly two girls; (ii) at least two girls, (iii) at most two girls and (iv) no girls.

Values of $X : x$	0	1	2	3
$p(x)$	1/8	1/8	3/8	1/8

Ans: (i) 3/8; (ii) 1/2; (iii) 7/8; (iv) 1/8.

9.4. Continuous Random Variable

A random variable obtained from a continuous sample space is called a continuous random variable. In this case a random variable cannot take any isolated value. It can take any value in a certain range. Age of a person is a good example of a continuous variable.

Other examples of continuous variable may be the length of life of a bulb, height of a student, weight of a student etc.

9.4.1. Probability density function. A function $f(x)$ of a continuous random variable X is called a probability density function if it satisfies the following two conditions

$$(i) f(x) \geq 0$$

$$(ii) \int_{-\infty}^{\infty} f(x) dx = 1.$$

The probability that a continuous random variable X takes a particular value x is zero. That is $P[X = x] = 0$.

Example 9.4.1. A continuous random variable has the following probability density function

$$f(x) = kx^2, 0 \leq x \leq 1.$$

(i) Find the value of k .

Find the probability of (ii) $P[0.2 \leq X \leq 0.5]$, (iii) $P[X < 0.3]$, (iv) $P[0.25 < X < 0.5]$, (v) $P[X > 0.75]$.

Solution. (i) Since the total probability is one,

$$\int_{-\infty}^{\infty} f(x) dx = 1 \Rightarrow \int_0^1 kx^2 dx = 1$$

$$\Rightarrow \left[k \frac{x^3}{3} \right]_0^1 = 1 \Rightarrow \frac{k}{3} = 1$$

Hence $k = 3$

$$(ii) P[0.2 \leq X \leq 0.5] = \int_{0.2}^{0.5} 3x^2 dx = \left[3 \frac{x^3}{3} \right]_{0.2}^{0.5} = (0.5)^3 - (0.2)^3 \\ = 0.125 - 0.008 = 0.117.$$

$$(iii) P[X < 0.3] = \int_0^{0.3} 3x^2 dx = \left[3 \frac{x^3}{3} \right]_0^{0.3} = (0.3)^3 = 0.027.$$

$$(iv) P[0.25 < X < 0.5] = \int_{0.25}^{0.5} 3x^2 dx = \left[3 \frac{x^3}{3} \right]_{0.25}^{0.5} = (0.5)^3 - (0.25)^3 \\ = 0.125 - 0.016 = 0.109.$$

$$(v) P[X > 0.75] = \int_{0.75}^1 3x^2 dx = \left[3 \frac{x^3}{3} \right]_{0.75}^1 = (1)^3 - (0.75)^3 = 1 - 0.423 = 0.578.$$

Example 9.4.2. Suppose that in a certain region of a country the daily rainfall (in inches) is a continuous random variable X with probability density function $f(x)$ given by

$$f(x) = \frac{3}{4}(2x - x^2), 0 < x < 2.$$

Find the probability that at a given day in this region the rainfall is (i) not more than 1 inch, (ii) more than 1.5 inches, (iii) between 0.5 and 1.5 inches (iv) equal to one inch and (v) less than one inch.

$$\text{Solution} \quad \text{(i)} \quad P[X < 1] = \int_0^1 \frac{3}{4}(2x - x^2) dx \\ = \frac{3}{4} \left[2 \frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = \frac{3}{4} \left(1 - \frac{1}{3} \right) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

$$\text{(ii)} \quad P[X > 1.5] = \int_{3/2}^2 \frac{3}{4}(2x - x^2) dx = \frac{3}{4} \left[2 \frac{x^2}{2} - \frac{x^3}{3} \right]_{3/2}^2 \\ = \frac{3}{4} \left(4 - \frac{8}{3} - \frac{9}{4} + \frac{27}{24} \right) = \frac{3}{4} \left(\frac{96 - 64 - 54 + 27}{24} \right) = \frac{3}{4} \times \frac{5}{24} = \frac{5}{32}.$$

$$\text{(iii)} \quad P[0.5 < X < 1.5] = \int_{1/2}^{3/2} \frac{3}{4}(2x - x^2) dx = \frac{3}{4} \left[2 \frac{x^2}{2} - \frac{x^3}{3} \right]_{1/2}^{3/2} \\ = \frac{3}{4} \left(\frac{4}{9} - \frac{8}{81} - \frac{1}{4} + \frac{1}{24} \right) = \frac{3}{4} \left(\frac{96 - 64 - 54 + 27}{24} \right) = \frac{3}{4} \times \frac{89}{648} = \frac{89}{864}.$$

(iv) $P[X = 1] = 0$, since the probability that a continuous variable takes a particular value is zero.

(v) It is the same as (i). Hence

$$P[X < 1] = \int_0^1 \frac{3}{4}(2x - x^2) dx \\ = \frac{3}{4} \left[2 \frac{x^2}{2} - \frac{x^3}{3} \right]_0^1 = \frac{3}{4} \left(1 - \frac{1}{3} \right) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

9.5. Mean and Variance of a Random Variable

The probability distribution provides a model for the theoretical frequency distribution of a random variable and hence must possess a mean, median, mode, variance, standard deviation and other descriptive measures associated with the theoretical population, which it represents. The mean of a random variable is called expected value or mathematical expectation of the random variable.

9.5.1. Mathematical expectation or expected value or mean of a discrete random variable

Definition. Let X be a discrete random variable which can take values x_1, x_2, \dots, x_n with associate probabilities $p(x_1), p(x_2), \dots, p(x_n)$, then mathematical expectation or mean of X , denoted by $E[X]$ is defined by

$$E[X] = x_1 p(x_1) + x_2 p(x_2) + \dots + x_n p(x_n) = \sum_{i=1}^n x_i p(x_i).$$

In other words, each value of the random variable is multiplied by the probability of the occurrence of the value and then all these products are summed up.

Actually mathematical expectation of a random variable is the mean of the population. It is denoted by a Greek symbol μ (mu).

9.5.2. Properties of expectation. The following are the important properties of an expected value of a random variable:

1. The expected value of a constant c is constant, i.e., $E(c) = c$ for every constant c
2. The expected value of the product of a constant c and a random variable X is equal to c times the expected value of the random variable, i.e., $E(cX) = c E(X)$
3. The expected value of a linear function of a random variable X is same as the linear function of its expectation i.e., $E(a + bX) = a + b E(X)$
4. The variance of the linear function of a random variable X is equal to the constant squared times the variance of the random variable X , i.e., $\text{Var}(a + bX) = b^2 \text{Var}(X)$

Example 9.5.1. A fair coin is tossed twice. Then the number of heads is a random variable that takes values 0, 1 and 2 with the following probability function:

Values of $X : x$	0	1	2	Total
$p(x)$	$1/4$	$1/2$	$1/4$	1

Find the expected number of heads.

Solution. The expected number of heads is

$$\begin{aligned} E[X] &= x_1 p(x_1) + x_2 p(x_2) + x_3 p(x_3) \\ &= 0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 0.5 + 0.5 = 1. \end{aligned}$$

This means on an average, we can expect one head if we toss one fair coin twice.

Example 9.5.2. Suppose a lot contains 4 items. Let the random variable X denotes the number of defective items in the lot. Suppose the probability function of the random variable X is

Values of $X : x$	0	1	2	3
$p(x)$	0.50	0.3	0.15	0.05

Find the expected number of defective items in the lot.

Solution. The expected number of defective items in the lot is

$$\begin{aligned} E[X] &= x_1 p(x_1) + x_2 p(x_2) + \dots + x_n p(x_n) \\ &= 0 \times 0.5 + 1 \times 0.3 + 2 \times 0.15 + 3 \times 0.05 \\ &= 0.3 + 0.3 + 0.15 = 0.75. \end{aligned}$$

This means that on an average 0.75 defective items can expected from the lot.

9.5.3. Variance of a discrete random variable. If X is a discrete random variable with mean μ which can take values x_1, x_2, \dots, x_n with associate probabilities $p(x_1), p(x_2), \dots, p(x_n)$, then the variance of X denoted by σ^2 is defined as

$$\sigma^2 = E[X - E(X)]^2 = E[X - \mu]^2 = E[X^2] - [E(X)]^2$$

$$\text{Here, } E[X^2] = x_1^2 p(x_1) + x_2^2 p(x_2) + \dots + x_n^2 p(x_n)$$

Standard deviation: The positive square root of variance is called standard deviation and it is denoted by σ .

Example 9.5.3. A company introduces a new product in the market and expects to make a profit of Tk. 2.5 lakh during the first year if the demand is good; Tk. 1.5 if the demand is moderate; and a loss of Tk. 1 lakh if the demand is poor. Market research studies indicate that the probabilities for the demand to be good, moderate and poor are 0.2, 0.5 and 0.3 respectively. Find the company's expected profit and the standard deviation.

Solution. Let X be a random variable representing the profit in three types of demand. Thus X may assume the values:

$x_1 = \text{Tk. 2.5 lakh when demand is good,}$

$x_2 = \text{Tk. 1.5 lakh when demand is moderate, and}$

$x_3 = \text{Tk. 1 lakh when demand is poor}$

The probability distribution of X is given by

Values of $X : x$	-1	1.5	2.5
$p(x)$	0.3	0.5	0.2

Hence, the expected profit is given by

$$E(X) = (-1) \times 0.3 + 1.5 \times 0.5 + 2.5 \times 0.2 = Tk. 0.95 \text{ lakh.}$$

On an average, the company can expect a profit of Tk. 0.95 lakh.

$$\begin{aligned} E(X^2) &= x_1^2 p(x_1) + x_2^2 p(x_2) + x_3^2 p(x_3) \\ &= (-1)^2 \times 0.3 + (1.5)^2 \times 0.5 + (2.5)^2 \times 0.2 = 2.675 \text{ lakh.} \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E(X^2) - [E(X)]^2 = 2.675 - (0.95)^2 \\ &= 2.675 - 0.9025 = 1.7725. \end{aligned}$$

$$S.D.(X) = \sqrt{1.7725} = Tk. 1.331 \text{ lakh.}$$

Example 9.5.4. A bakery has the following probability function of daily demand for marriage day cake:

No. of cakes demanded X : x	1	2	3	4	5	6	7	8
Probability p(x)	0.02	0.07	0.09	0.12	k	0.2	0.18	0.02

- (i) Find the value of k.
- (ii) Find the expected number of marriage day cakes demanded per day.

Solution. Since $\sum p(x) = 1$.

$$\text{We have, } 0.02 + 0.07 + 0.09 + 0.12 + k + 0.2 + 0.18 + 0.02 = 1$$

$$0.7 + k = 1 \Rightarrow k = 1 - 0.7 = 0.3.$$

The expected number of marriage day cakes is

$$\begin{aligned} E[X] &= x_1 p(x_1) + x_2 p(x_2) + \dots + x_n p(x_n) \\ &= 1 \times 0.02 + 2 \times 0.07 + 3 \times 0.09 + 4 \times 0.12 + 5 \times 0.3 + 6 \times 0.2 + 7 \times 0.18 + 8 \times 0.02 \\ &= 0.02 + 0.14 + 0.27 + 0.48 + 1.50 + 1.20 + 1.26 + 0.16 = 5.03 \approx 5. \end{aligned}$$

That is on an average, the expected number of cakes demanded per day is 5.

Some Matched Problems

Example 9.5.5. The following table shows the probability distribution of a number of long-distance calls made in a month by the residents of urban households in an area:

No. of calls	0	1	2	3	4	5
Probability	0.05	0.21	0.56	0.06	0.08	0.04

If X denotes the number of calls, find the expected value and the variance of X.

Solution. Mean = Expected value of the long – distance call is

$$\begin{aligned}\mu &= 0 \times (0.05) + 1 \times (0.21) + 2 \times (0.56) + 3 \times (0.06) + 4 \times (0.08) + 5 \times (0.04) \\ &= 0 + 0.21 + 1.12 + 0.18 + 0.32 + 0.20 = 2.03\end{aligned}$$

$$\text{Varianc} = \sigma^2 = E[X^2] - (E[X])^2$$

$$\begin{aligned}E[X^2] &= 0 \times (0.05) + 1 \times (0.21) + 4 \times (0.56) + 9 \times (0.06) + 16 \times (0.08) + 25 \times (0.04) \\ &= 0 + 0.21 + 2.24 + 0.54 + 1.28 + 1.00 = 5.27\end{aligned}$$

$$\sigma^2 = E[X^2] - (E[X])^2 = 5.27 - (2.03)^2 = 1.15.$$

Example 9.5.6. Bangladesh Bank has six tellers available to serve customers. The number of tellers busy with customers at peak time, say, 2:00 p.m. varies from day to day. So it is a random variable denoted as X . It is known from the past records that the probability distribution of X is as follows:

Values of $X : x$	0	1	2	3	4	5	6
$p(x)$	0.03	0.05	0.08	0.15	0.21	0.26	0.22

- a) Find the mean number of tellers busy with the customers at 2:00 p.m.
- b) Also find the variance and standard deviation of the number of tellers busy with the customers at 2:00 p.m.

Solution. (a) Mean = $\mu = E[X] = \Sigma xp(x)$

$$\begin{aligned}&= 0(0.03) + 1(0.05) + 2(0.08) + 3(0.15) + 4(0.21) + 5(0.26) + 6(0.22) \\ &= 0 + 0.05 + 0.16 + 0.45 + 0.84 + 1.30 + 1.32 = 4.12\end{aligned}$$

Hence the mean number of tellers busy with the customer is approximately 4.

(b) Variance = $\sigma^2 = E[X^2] - (E[X])^2$

$$\begin{aligned}E[X^2] &= 0 \times (0.03) + 1 \times (0.05) + 4 \times (0.08) + 9 \times (0.15) + 16 \times (0.21) + 25 \times (0.26) + 36 \times (0.22) \\ &= 0 + 0.05 + 0.32 + 0.54 + 1.35 + 3.36 + 6.50 + 7.92 = 20.04\end{aligned}$$

$$\sigma^2 = E[X^2] - (E[X])^2 = 20.04 - (4.12)^2 = 3.07$$

Standard deviation = $\sigma = 1.75$.

Example 9.5.7. Suppose you are interested in insuring a car stereo system for Tk. 500 against theft. An insurance company charges a premium of Tk. 60 for coverage for 1 year, claiming an empirically determined probability 0.1 that the stereo will be stolen some time during the year. What is your expected return from the insurance company if you take out this insurance?

Solution. This is actually a game of chance in which your stake is Tk.60. You have a 0.1 chance of receiving Tk.440 from the insurance company (Tk.500 minus your stake Tk.60) and a 0.9 chance of losing your stake of Tk.60. What is the expected value of the game? Here X be the amount payoff. The probability distribution of X is

Value of $X : x$	440	-60
$p(x)$	0.1	0.9

The expected value of the game is

$$E(X) = (440)(0.1) + (-60)(0.9) = 44 - 54 = -10.$$

This means that if you insure with this company over many years and the circumstances remain the same, you would have an average net loss to the insurance company of Tk. 10 per year.

Matched problem

Insurance. The annual premium for a Tk. 5000 insurance policy against the theft of a painting is Tk.150. If the probability that the painting will be stolen during the year is .01, what is your expected return from the insurance company if you take out this insurance?

Pay off table

$X : x$	4850	-150
$p(x)$	0.01	0.99 ; $E(X) = \text{Tk. } 100$

9.5.4. Mean and variance of a continuous random variable. Suppose X is a continuous random variable with probability density function $f(x)$, then the expected value or mean of X is defined as

$$\mu = E[X] = \int_{-\infty}^{\infty} f(x) dx ; -\infty < x < \infty$$

$$\text{Variance} = E[X - \mu]^2 = E[X^2] - \mu^2.$$

$$\text{Here } E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx.$$

Example 9.5.8. Suppose that in a certain region of a country the daily rainfall (in inches) is a continuous random variable X with probability density function $f(x)$ given by $f(x) = \frac{3}{4}(2x - x^2)$, $0 < x < 2$. Find the expected daily rainfall (in inches) in that region. Also find variance and standard deviation.

Solution. The expected daily rainfall is

$$\mu = E[X] = \int_0^2 x f(x) dx$$

$$= \int_0^2 x \cdot \frac{3}{4} (2x - x^2) dx = \frac{3}{4} \left[2 \frac{x^3}{3} - \frac{x^4}{4} \right]_0^2 = \frac{3}{4} \left(\frac{16}{3} - \frac{16}{4} \right) = \frac{3}{4} \times \frac{4}{3} = 1.$$

This means, on an average the daily rainfall of that region is one inch.

$$\text{Variance} = \sigma^2 = E[X - \mu]^2 = E[X^2] - \mu^2$$

$$E[X^2] = \frac{3}{4} \int_0^2 x^2 (2x - x^2) dx$$

$$= \frac{3}{4} \left[2 \frac{x^4}{4} - \frac{x^5}{5} \right]_0^2 = \frac{3}{4} \left(\frac{32}{4} - \frac{32}{5} \right) = \frac{3}{4} \times \frac{32}{20} = \frac{6}{5} = 1.2$$

$$\text{Variance} = 1.2 - 1 = 0.2.$$

$$\text{Standard deviation} = \sigma = \sqrt{0.2} = 0.45 \text{ inches.}$$

Matched problem.

Let X be a random variable with probability function $f(x) = 2x; 0 < x < 1$. Find mean, variance and standard deviation of X .

Ans. Mean=2/3, variance=5/9 and $\sigma = 0.745$.

Questions

- What is a random variable? Define a discrete random variable with examples.
- Define a discrete random variable. Explain its probability function with an example.
- Define a continuous random variable with example. Distinguish between probability function and probability density function.
- Define discrete and continuous random variable. Cite one example for each.
- Define a random variable. What is the mathematical expectation of a random variable?
- Define variance and standard deviation of a random variable.

Exercise

- The probability function of a discrete random variable X is as follows:

$X:x$	-5	-3	-1	1	3	5
$p(x)$	a	2a	3a	4a	5a	6a

- (i) Find the value of a , (ii) Find $P[X = 1]$, (iii) $P[-1 < X < 3]$, (iv) $P[-1 \leq X \leq 3]$, (v) $P[X > -1]$, (vi) $P[X \geq 4]$, (vii) $P[X \leq 1]$.
8. Let X be a discrete random variable with the following probability function

Values of $X : x$	0	1	2	3	4
$p(x)$	0.12	0.18	k	0.30	0.16

- (i) Find the value of k , Compute (ii) $P[X > 3]$, (iii) $P[1 < X < 4]$, (iv) $P[X < 1]$.
9. Let X be a random variable with probability function given below;

x	0	1	2	3
$p(x)$	$1/6$	$1/2$	$3/10$	$1/30$

- Find $P[X \leq 1]$; $P[X < 1]$ and $P[0, X, 2]$. Ans. $2/3, 5/6, 1/2$.
10. A continuous random variable X has the following probability distribution: $f(x) = kx^2$; $0 \leq x \leq 1$. (i) Find the value of k , (ii) Calculate the probability that X lies between 0.2 and 0.5, (iii) X less than 0.3, (iv) $1/4 < X < 1/2$. Ans. (i) 0.3, (ii) 0.117, (iii) 0.027
11. Find mean, variance and standard deviation of the following probability function:

Values of $X : x$	0	1	2	3
$p(x)$	$1/8$	$3/8$	$3/8$	$1/8$

Ans. 1.5; 0.74; 0.86.

12. A random variable X can assume five values: 0, 1, 2, 3, 4, A portion of the probability distribution is shown here:

Values of $X : x$	0	1	2	3	4
$p(x)$.1	.3	.3	?	.1

- (i) Find $p(3)$.
(ii) Calculate mean, variance and standard deviation
(iii) What is the probability that X is greater than 2?
(iv) What is the probability that X is 3 or less?

Ans. (i) 0.2; (ii) $\mu = 1.9$; $\sigma^2 = 1.29$; $\sigma = 1.136$; (iii) 0.3; (iv) 0.9

Application

13. A company has five applicants for two positions: two women and three men. Suppose that the five applicants are equally qualified and that no preference is given for choosing either gender. Let X be the number of women chosen to fill the two positions (i) Find the probability function of X . (ii) What is the probability that exactly 1 woman will be chosen? (iii) Also find the probability that at least one woman will be chosen.

Ans. $p(0) = 3/10$; $p(1) = 6/10$; $p(2) = 1/10$. (ii) $6/10$; (iii) $7/10$.

14. Past experience has shown that, on the average, only one in ten wells drilled hits oil. Let X be the number of drillings until the first success. Assume that the drillings represent independent events. (i) Find $p(1)$; $p(2)$; $p(3)$; (ii) Give a formula for $p(x)$.

Ans (i) .0.1; 0.09; 0.081 (ii) $p(x) = (0.9)^{x-1} (0.1)$

15. Let X be the number of times Mrs. Karim visits a grocery store in a 1-week period. The probability distribution of X is as follows:

$X : x :$	0	1	2	3
$p(x) :$.1	.4	.4	.1

Find the average number of times Mrs. Karim visits the store.

Ans. 1.5

16. In a lottery conducted to benefit the local company, 8000 tickets are to be sold at Tk. 5 each. The prize is a Tk. 12,000. If you purchase two tickets, what is your expected gain?

Ans. $x : -10 \quad 11990 \quad E[X] = -Tk. 7$
 $p(x) : 7998/8000 \quad 2/8000$

17. Mr. Rahman just has bought a VCR from Jalani's Videotape Service at a cost of \$300. He now has the option of buying an extended service warranty offering 5 years of coverage for \$100. After talking to friends and reading reports, Mr. Rahman believes the following maintenance could be incurred during the next five years

Expense (x)	0	50	100	150	200	250	300
$p(x)$	0.35	0.25	0.15	0.10	0.08	0.05	0.02

Find the expected value of the anticipated maintenance costs. Should Mr. Rahman pay \$100 for the warranty?

Ans. The expected value of the maintenance cost is \$77, so Mr. Rahman should not pay \$100 for the insurance.

18. A bakery has the following probability function of daily demand for birth day cake:

No. of cakes Demanded: x	1	2	3	4	5	6
Probability $p(x)$	0.05	0.15	0.30	0.25	0.17	0.08

- (i) Find the expected number of birth day cakes demanded per day; (ii) Find variance and standard deviation of the number of cakes demanded.

Ans. (i) 3.58; (ii) 1.67 and 1.29.

CHAPTER - 10

PROBABILITY DISTRIBUTIONS

10.1. Introduction

In this chapter, we shall discuss some important probability distributions, which are found to have wide applications in real life. There are a large number of probability distributions, but we shall limit ourselves only in three important probability distributions, two of which are discrete and the other is continuous. The two discrete distributions are

- (i) Binomial distribution and
- (ii) Poisson distribution

Moreover, we shall discuss the most important continuous distribution, known as normal distribution.

Some concepts related to Binomial distribution are discussed below.

10.2. Bernoulli Trial

Binomial distribution is related with Bernoulli trial. A trial is a unit experiment. An experiment is called Bernoulli trial if it has two possible outcomes namely success and failure. It is possible to categorize all possible outcomes of an experiment into two categories. For example, in a die throwing experiment, if our desired number is a six, then six is our success and other numbers are failure. The probability of success as well as the probability of failure of a Bernoulli trial remains the same from trial to trial. Suppose p is the probability of success, then $q = 1 - p$ is the probability of failure.

Example 10.1.1. A product of a factory may be a defective or non-defective. We may call a non-defective product as a success and a defective product as a failure. The probability of a non-defective product or success is p and the probability of a defective product or failure is $q = 1 - p$.

Example 10.1.2. A coin is tossed in which the outcome head is a success and the probability of success is p . Then $q = 1 - p$ is the probability of failure or tail.

A newly born baby may be a boy or a girl is a Bernoulli trial. A student may pass or fail in an examination, an applicant may or may not get a job etc are the examples of Bernoulli trial.

10.2.1. Binomial experiment. An experiment is called binomial experiment when it has the following properties :

1. The experiment consists of n Bernoulli trials.
2. Each trial has two possible outcomes namely success and failure.
3. The probability of success, p remains the same from trial to trial.
4. The repeated trials are independent.

10.2.2. Combinations. The number of combinations of selecting x objects out of n objects is given by ${}^n C_x = \frac{n!}{x!(n-x)!}$

where $n! = n(n-1)(n-2) \dots (2)(1)$ is called n factorial and $0! = 1$.

$$\text{For example, } {}^5 C_3 = \frac{5!}{3!(5-3)!} = \frac{5!}{3!(2!)} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1(2 \times 1)} = 10.$$

Example 10.2.3. Suppose there are 5 football teams in a tournament. How many games are possible with these five teams?

Solution. We know two teams are required to play a game. The two teams can be selected from the 5 teams in ${}^5 C_2$ ways. Hence

$${}^5 C_2 = \frac{5!}{2!(5-2)!} = \frac{5!}{2!(3!)} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1(3 \times 2 \times 1)} = 10.$$

That is 10 games are possible.

10.3. Binomial Distribution

Binomial distribution was first derived by Swiss Mathematician James Bernoulli (1654 - 1705) in 1700 and was published by his nephew Nicholas Bernoulli in 1713 eight years after his death. It is one of the most frequently used discrete probability distributions, and is very useful in many practical situations from rolling dice to quality control on a manufacturing production line.

The number x of successes in n Bernoulli trials of a binomial experiment is called binomial random variable, and the probability distribution of this discrete variable is called the binomial distribution. The probability function of the binomial variate X is defined by $p(x) = {}^n C_x p^x q^{n-x} : x = 0, 1, \dots, n$. We can also define binomial distribution in the following way:

Again, if p is the probability of success in a Bernoulli trial and if p remains the same from trial to trial, then probability of x successes in n independent trial is given by $p(x) = {}^n C_x p^x q^{n-x} ; x = 0, 1, 2, \dots, n$.

Definition. Binomial Distribution. A discrete random variable X is said to have a binomial distribution if its probability function is defined by

$$p(x) = {}^n C_x p^x q^{n-x}; x = 0, 1, 2, \dots, n.$$

where $p + q = 1$.

This distribution is called binomial distribution, since the probability for different values of X are the various terms of the binomial expansion

$$(q+p)^n = q^n + {}^n C_1 p^1 q^{n-1} + {}^n C_2 p^2 q^{n-2} + \dots + p^n.$$

Here p is the probability of success in a single Bernoulli trial and q is the probability of failure. It is to be noted that n and p are the two unknown constants of the distribution. They are known as the parameters of the distribution. Here n is the number of Bernoulli trials and p is the probability of successes:

The binomial variable X with parameter n and p is symbolically written as $X \sim B(n, p)$

Parameter

The unknown constants required to define a distribution are called the parameters of the distribution. A distribution is completely known if the parameters are known. It is to be noted that any numerical value calculated from the population is also called parameter.

10.3.1. Mean and Variance of the distribution. If n and p are two parameters of the binomial distribution, then mean and variance of the distribution are

$$\text{Mean} = \mu = E[X] = np \quad \text{and} \quad \text{Variance} = \sigma^2 = npq.$$

$$\text{Standard deviation} = \sigma = \sqrt{npq}.$$

It is to be noted that mean is greater than variance of the distribution since $q < 1$.

10.3.2. Some important properties of the distribution

1. It is a discrete probability distribution with parameters n and p .
2. The mean of the distribution is np and its variance is npq . The mean of the distribution is greater than variance since $q < 1$.
3. The distribution is positively skewed if $p < 1/2$ and negatively skewed if $p > 1/2$.
4. The distribution is symmetric if $p = q = 1/2$.
5. The distribution tends to Poisson distribution if the number of trials, n tends to infinity.

6. The distribution tends to normal distribution if n tends to infinity and p or q is not so small.
7. $P(X = n) = p^n$ and $P(X = 0) = q^n$

10.3.3. Some practical applications of Binomial distribution

1. Number of defective items in a randomly selected sample of 12 products.
2. Number of days of increasing price index in share market in a randomly selected 15 days.
3. Number of correct answers in a multiple choice test if a student answers all the questions randomly.
4. Number of workers suffer from occupational disease in a randomly selected sample of 10 workers.
5. Number of female babies in society.
6. Number of successful hits in a target out of fixed number of hits.
7. Number of customers buy a particular commodity.

Example 10.3.1. Suppose X is a binomial variate with parameters $n = 4$ and $p = 0.45$. Find the following probabilities

- (i) $P[X = 2]$; (ii) $P[X > 3]$ and (iii) $P[X \leq 2]$.

Solution. We have $n = 4$ and $p = 0.45$, then $q = 1 - p = 1 - 0.45 = 0.55$, i.e. $X \sim B(4, 0.45)$.

Then the probability function of the binomial variable X is

$$\begin{aligned} P[X = x] &= p(x) = {}^n C_x p^x q^{n-x} \quad x = 0, 1, 2, 3, 4 \\ &= {}^4 C_x (0.45)^x (0.55)^{4-x}; x = 0, 1, 2, 3, 4 \end{aligned}$$

$$\begin{aligned} (i) P[X = 2] &= p(2) = {}^4 C_2 p^2 q^{4-2} \\ &= {}^4 C_2 (0.45)^2 (0.55)^2 = 10(0.2025)(0.3025) = 0.3675 \\ (ii) P[X > 3] &= P[X = 4] = p(4) = (0.45)^4 = 0.0410 \\ (iii) P[X \leq 2] &= P[X = 0] + P[X = 1] + P[X = 2] = p(0) + p(1) + p(2) \\ &= 0.0915 + 0.2995 + 0.3675 = 0.7585. \end{aligned}$$

Matched problem

Problem. For a binomial distribution with $n = 7$ and $p = 0.2$, find

- (i) $P[X = 5]$; (ii) $P[X < 8]$ and (iii) $P[X \geq 4]$

Ans. (i) 0.0043; (ii) 1 (iii) 0.0333.

Example 10.3.2. The probability that a patient recovers from a rare blood disease is 0.4. If 5 people are known to have contracted this disease, what is

the probability that (i) exactly 3 survive, (ii) at least two survive (iii) at most two survive and (iv) none survive.

Solution. Here, we define survive as success and not survive as failure. Let X be the number of people that survive. Then X is a binomial variate with probability of success $p = 0.4$ and probability of failure $q = 1 - p = 1 - 0.4 = 0.6$ and $n = 5$, i.e., $X \sim B(5, 0.4)$. Then according to binomial distribution, the probability function of X is

$$P[X = x] = p(x) = {}^n C_x p^x q^{n-x}; \quad x = 0, 1, 2, \dots, n.$$

$$= {}^5 C_x (0.4)^x (0.6)^{5-x}; \quad x = 0, 1, 2, 3, 4, 5.$$

The probabilities for different values of X are

$$P[X = 0] = p(0) = (0.6)^5 = 0.07776.$$

$$P[X = 1] = p(1) = {}^5 C_1 (0.4)(0.6)^4 = \frac{5!}{4!1!} (0.4)(0.1296) = 5 \times 0.0518 = 0.2592.$$

$$P[X = 2] = p(2) = {}^5 C_2 (0.4)^2 (0.6)^3 = \frac{5!}{2!3!} (0.16)(0.216) = 10 \times 0.03456 = 0.3456.$$

$$P[X = 3] = p(3) = {}^5 C_3 (0.4)^3 (0.6)^{5-3} = \frac{5!}{3!2!} (0.4)^3 (0.6)^2 = 10 \times 0.064 \times 0.36 = 0.2304.$$

$$P[X = 4] = p(4) = {}^5 C_4 (0.4)^4 (0.6) = \frac{5!}{4!1!} (0.0256)(0.6) = 5 \times 0.01536 = 0.0768.$$

$$P[X = 5] = (0.4)^5 = 0.01024.$$

$$(i) \quad P[\text{Exactly 3 survive}] = P[X = 3] = 0.2304.$$

$$(ii) \quad P[\text{at least 2 survive}] = P[X \geq 2]$$

$$= P[X = 2] + P[X = 3] + P[X = 4] + P[X = 5]$$

$$= 0.3456 + 0.2304 + 0.0768 + 0.01024 = 0.66304.$$

$$(iii) \quad P[X \leq 2] = P[X = 2] + P[X = 1] + P[X = 0]$$

$$= 0.3456 + 0.2592 + 0.07776 = 0.65376.$$

$$(iv) \quad P[X = 0] = p(0) = 0.07776.$$

Example 10.3.3. The probability that a MBA graduate from a private University gets an executive job is 0.6. Four MBA graduates applied for executive jobs. What is the probability that (i) exactly 2 will get the job, (ii) all will get the job, (iii) at least 3 will get the job.

Solution. We define a MBA gets a job as success and does not get a job as failure. Let X be the number of MBA graduates get the job. Then X is a binomial variate with $p = 0.6$, $q = 0.4$ and $n = 4$, i.e., $X \sim B(4, 0.6)$.

Then the probability function of X is

$$\begin{aligned} P[X=x] = p(x) &= {}^n C_x p^x q^{n-x}; \quad x=0, 1, 2, \dots, n. \\ &= {}^4 C_x (0.6)^x (0.4)^{4-x}; \quad x=0, 1, 2, 3, 4. \end{aligned}$$

The probabilities for different values of X are

$$P[X=0] = p(0) = (0.4)^4 = 0.0256.$$

$$P[X=1] = p(1) = {}^4 C_1 (0.4)^3 (0.6) = \frac{4!}{3!1!} (0.4)^3 (0.6) = 4 \times 0.0384 = 0.1536.$$

$$P[X=2] = p(2) = {}^4 C_2 (0.6)^2 (0.4)^2 = \frac{4!}{2!2!} (0.36)(0.16) = 6 \times 0.0576 = 0.3456.$$

$$P[X=3] = p(3) = {}^4 C_3 (0.6)^3 (0.4) = \frac{4!}{3!1!} (0.6)^3 (0.4) = 4 \times 0.216 \times 0.4 = 0.3456.$$

$$P[X=4] = p(4) = (0.6)^4 = 0.1296.$$

$$(i) \quad P[X=2] = 0.3456.$$

$$(ii) \quad P[X=4] = 0.1296.$$

$$(iii) \quad P[X \geq 3] = P[X=3] + P[X=4] = 0.3456 + 0.1296 = 0.4752.$$

Example 10.3.4. Suppose X is a binomial variate with mean 4 and variance 3. Find the probability distribution of X .

Solution. Let n and p be the two parameters of the distribution. Here, we have

$$\text{Mean} = np = 4 \text{ and variance} = npq = 3$$

$$\text{Hence, } q = \frac{npq}{np} = \frac{3}{4}. \text{ Then } p = 1 - \frac{3}{4} = \frac{1}{4}.$$

$$n = 4/p = 4 \times 4 = 16.$$

Hence the required binomial distribution is

$$f(x; n, p) = f(x; 16, \frac{1}{4}) = \binom{16}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{16-x}; \quad x = 0, 1, 2, \dots, 16$$

Matched problem

Problem. Suppose X is binomial variate with mean 9 and variance 3.6. Find the binomial distribution.

$$\text{Ans } p = 3/4, n = 15$$

$$f(x; n, p) = f(x; 15, \frac{3}{4}) = \binom{15}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{15-x}; \quad x = 0, 1, 2, \dots, 15.$$

Exercise 10.3.5. Warranty records show that the probability that a new car needs a warranty repair in the first 90 days is 0.05. If a sample of three new cars is selected, what are the probabilities that in the first 90 days

- (i) None needs a warranty repair?
- (ii) At least one needs a warranty repair?
- (iii) More than one needs a warranty repair?
- (iv) What are the mean and standard deviation of the number of cars need warranty repair.

Solution. Here X is the number of cars need warranty repair. X is a binomial variate with parameters $n = 3$ and $p = 0.05$. Then $q = 1 - 0.05 = 0.95$. Then the probability function of X is

$$\begin{aligned} P[X = x] &= p(x) = {}^n C_x p^x q^{n-x}; \quad x = 0, 1, 2, \dots, n. \\ &= {}^3 C_x (0.05)^x (0.95)^{3-x}; \quad x = 0, 1, 2, 3. \end{aligned}$$

(i) The probability that non-needs warranty repair is

$$P[X = 0] = p(0) = (0.95)^3 = 0.8574$$

(ii) The probability that at least one needs repair is

$$P[X > 0] = 1 - P[X = 0] = 1 - 0.8574 = 0.1426.$$

(iii) The probability that more than one needs repair is

$$\begin{aligned} P[X > 1] &= 1 - P[X \leq 1] = 1 - (P[X = 0] + P[X = 1]) \\ &= 1 - (0.8574 + 0.1354) = 0.0072. \end{aligned}$$

(iv) The mean of the distribution is $\mu = np = 3 \times 0.05 = 0.15$.

The standard deviation of the distribution is

$$\sigma = \sqrt{npq} = \sqrt{3 \times 0.05 \times 0.95} = 0.38.$$

Example 10.3.6. Admission tests in different faculties of Chittagong University are held in MCQ method. There are ten MCQs in Bangla, each carrying one mark, are provided for each subject. There are five options in each question only one of which is correct answer. Suppose, a candidate has appeared in the test in a faculty without any preparation in Bangla. S/he has to answer all the questions randomly (i) Write down the distribution of correct answers, (ii) find the probability that only three of such answers are correct, (iii) if the pass mark for that subject is 5, find the probability of passing in the subject.

Solution: (i) Let X be the number of correct answers. Here, the number of correct answers can take the values 0, 1, ..., 10, and one answer out of 5 is correct, so the probability of a correct answer is $p = 1/5 = 0.20$, the probability of an incorrect answer is $q = 1 - p = 0.80$, which remain constant

from question to question. Thus, X follows binomial distribution with $n = 10$ and $p = 0.20$, i.e., $X \sim B(10, 0.20)$. The probability function of X is given by

$$P[X = x] = {}^n C_x p^x q^{n-x} ; x = 0, 1, 2, \dots, 10$$

$$\text{Or, } P[X = x] = {}^{10} C_x (0.2)^x (0.8)^{10-x} ; x = 0, 1, 2, \dots, 10$$

$$(ii) P[X = 3] = {}^{10} C_3 (0.20)^3 (0.80)^{10-3} = 120(0.20)^3 (0.80)^7 = 0.20113$$

(iii) To pass in that subject, the candidate must answer at least 5 questions correctly, so the required probability is

$$\begin{aligned} P[X \geq 5] &= P[X = 5] + P[X = 6] + \dots + P[X = 10] \\ &= 0.026424 + 0.005505 + \dots + 0.000000102 = 0.03793 \end{aligned}$$

10.4. Poisson Distribution

Poisson distribution was discovered by the French mathematician and physicist Simeon Denis Poisson [1781-1840], who published it in 1837. It has many applications in modern day life such as modeling the radioactive decay, predicting the number of telephone calls arriving at an exchange, etc. Poisson distribution is a limiting case of the binomial distribution under the following conditions:

- (i) The probability of success or failure in Bernoulli trial is very small.
That is $p \rightarrow 0$ or $q \rightarrow 0$,
- (ii) n , the number of trials is very large and
- (iii) $np = \lambda$ (say) is a finite constant.

Definition. Poisson distribution. A discrete random variable X is said to have a Poisson distribution if its probability function is defined by

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} ; x = 0, 1, 2, \dots, \infty$$

where $e = 2.71828$ and λ is the only parameter of the distribution which is the mean of the distribution.

It can be easily shown that

- i) $f(x; \lambda) \geq 0$ for each x
- ii) $\sum f(x; \lambda) = 1$.

It can be shown that the both mean and variance of the distribution are equal to λ . The distribution is positively skewed and leptokurtic. But if the mean of the distribution tends to infinity, Poisson distribution tends to normal distribution.

The Poisson variable X with parameter λ is written as $X \sim Po(\lambda)$

10.4.1. Some important properties of the distribution.

- 1) The mean of the distribution is λ
- 2) The variance of the distribution is also λ . So the mean and variance of the distribution are equal.
- 3) When λ tends to ∞ , Poisson distribution tends to normal distribution.
- 4) The distribution is positively skewed and leptokurtic.

10.4.2. Conditions under which a Poisson distribution is likely to exist

- 1) The events occur independently of each other
- 2) The events occur singly and at random in continuous space or time.
- 3) The events occur at a constant rate in the sense that the mean number of occurrences in the given space or interval of time is proportional to the length of space or time. For example, if the average number of cars passing through a certain street in time t is λ , then average number of cars passing through that street in time $2t$ is 2λ , if the average number of printing mistakes per page of a book is m , then the average number of printing mistakes in every 3-page of that book is $3m$, and so on.

10.4.3. Some practical applications of Poisson distribution

- 1) The number of cars passing through a certain street in time t .
- 2) The number of suicides reported per day of a certain area.
- 3) The number of faulty blades in a packet of 100.
- 4) The number of printing mistakes per page of a book.
- 5) The number of telephone calls received at a particular telephone exchange in some unit time.
- 6) The number of defective materials in packing manufactured by a goods concern.
- 7) The number of letters lost in a mail per day in a certain city.
- 8) The number of robbers caught per day in a certain city
- 9) Number of deaths from a rare disease in a locality.
- 10) The number of customers arriving in a departmental store per minute.
- 11) The number of accidents occurring in a factory over a month

Example 10.4.1. Given $\lambda = 4.2$, for a Poisson distribution, find (i) $P[X \leq 2]$; (ii) $P[X \geq 5]$; $P[x = 8]$.

Solution. Here $\lambda = 4.2$, then $e^{-4.2} = 0.0150$

$$(i) P[X \leq 2] = p(0) + p(1) + p(2) = \frac{(4.2)^0 e^{-4.2}}{0!} + \frac{(4.2)^1 e^{-4.2}}{1!} + \frac{(4.2)^2 e^{-4.2}}{2!}$$

$$= 0.0150 + 0.0630 + 0.1323 = 0.2103.$$

$$(ii) P[X \geq 5] = 1 - P[X \leq 4] = 1 - p(4) - p(3) - p[X \leq 2]$$

$$= 1 - \frac{(4.2)^4 e^{-4.2}}{4!} - \frac{(4.2)^3 e^{-3}}{3!} - 0.2103$$

$$= 1 - 0.1844 - 0.1852 - 0.2103 = 0.4101.$$

$$(iii) P[X=8] = \frac{(4.2)^8 e^{-4.2}}{8!} = 0.0360.$$

10.4.4. Poisson approximation to Binomial. When n is large ($n > 29$) and p is small such that $\lambda = np < 7$, then binomial distribution can be approximated by a Poisson distribution.

Example 10.4.2. Suppose a life insurance company insures the lives of 5000 men aged 42. If actual studies show the probability that any 42-year old man will die in a given year to be .001, find the exact probability that the company will have to pay $x = 4$ claims during a given year.

Solution. The exact probability is given by binomial distribution as

$$P(x; .001, 5000) = \binom{5000}{4} (.001)^4 (999)^{4996}$$

For which binomial table are not available. Here $np = (5000)(.001) = 5 < 7$ and n is very large. So we can safely use Poisson distribution for finding the probability.

$$P[X=4] = p(4) \frac{e^{-5} 5^4}{4!} = \frac{(.006738)(625)}{24} = 0.175.$$

Example 10.4.3. For a binomial distribution with $n = 30$ and $p = 0.04$, Find the following probabilities by using Poisson approximation to the binomial:
(i) $P(x = 25)$; (ii) $P[x = 3]$ and (iii) $P[x = 5]$.

Solution. For binomial $n = 30$ and $p = 0.04$

Since $n = 30$, we can use Poisson approximation to the binomial. For Poisson

$$\lambda = n = 30 \times 0.04 = 1.2; e^{-1.2} = 0.30119.$$

Here $\lambda < 7$ and n is 30. So we can safely use Poisson approximation to binomial.

$$(i) P[X=25] = \frac{(1.2)^{25} e^{-1.2}}{25!} = 0.0000$$

$$(ii) P[X=3] = p(3) \frac{(1.2)^3 e^{-1.2}}{3!} = 0.0867$$

$$(iii) P[X=5] = p(5) = \frac{(1.2)^5 e^{-1.2}}{5!} = 0.0062$$

Example 10.4.4. Suppose that the number of emergency patients X in a given day at a certain hospital follows Poisson distribution with parameter $\lambda = 20$. What is the probability that in a given day there will be (a) 15 emergency patients, (b) at least 3 emergency patients and (c) more than 20 but less than 25 patients.

Solution. (a) Here, we have $\lambda = 20$ and $x = 15$.

Since X follows Poisson distribution, then

$$P[X=15] = p(15) = \frac{(20)^{15} e^{-20}}{20!} = 0.0516.$$

(From the table of Poisson distribution)

$$(b) P[\text{at least 3 patients}] = P[X \geq 3] = 1 - P[X \leq 2]$$

$$= 1 - p(0) - p(1) - p(2) = 1 - 0.0000 - 0.0000 - 0.0000 = 1.$$

$$\begin{aligned} (c) P[20 < X < 25] &= p(21) + p(22) + p(23) + p(24) \\ &= 0.0846 + 0.0769 + 0.0669 + 0.0537 \\ &= 0.2441. \end{aligned}$$

Example 10.4.5. In a production process, on an average, one in 400 items is defective. Suppose the items are packed in boxes of 100.

a) A box is selected from a day's output, assuming appropriate probability model find the probability that the box will contain

- (i) no defective;
- (ii) less than two defectives
- (iii) one or more defectives

b) If the items are packed in boxes of 500, what is the probability that a randomly selected box will contain no defective?

Solution. (a) Here, number of defective items in a box is a random variable.

Let X be the number of defective items in a box,

The probability that an item will be defective is

$p = 1/400$, probability is very low.

$n = 100$, number of items packed in the box is quite large.

Hence the appropriate probability model is Poisson with mean,

$$\lambda = np = 100 \times \frac{1}{400} = 0.25;$$

which is the average number of defective items in a box of 100.

Thus, $X \sim Po(0.25)$ and $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$, so,

(i) Probability of no defective

$$= P[X=0] = p(0) = e^{-\lambda} = e^{-0.25} = 0.7788$$

(ii) Probability of less than two defectives

$$= P[X \leq 1] = P[X=0] + P[X=1] = p(0) + p(1)$$

$$= e^{-\lambda} + \lambda e^{-\lambda} = e^{-0.25}(1 + \lambda) = 0.7788(1 + 0.25) = 0.9735.$$

(iii) Probability of one or more defectives

$$= P[X \geq 1] = 1 - P[X=0] = 1 - p(0) = 1 - e^{-\lambda} = 1 - e^{0.25}$$

$$= 1 - 0.7788 = 0.2212.$$

b) Let Y be the number of defective items in a box of 500, then the average number of defectives in the box is $\lambda = np = 500 \times \frac{1}{400} = 1.25$,

Hence $Y \sim Po(1.25)$

The required probability is given by

$$P(Y=0) = e^{-\lambda} = e^{-1.25} = 0.2865$$

Example 10.4.6. A shop sells AC machines at the rate of 2.5 pieces per week, find the probability that in a 2-week period the shop sells at least 7 AC machines.

Solution. Let X be the number of ACs sold in a 2-week period, then according to question $X \sim Po(2 \times 2.5)$ or $X \sim Po(5)$.

$$\text{We know, } p(x) = \frac{e^{-\lambda} \lambda^x}{x!}; x = 0, 1, 2, \dots, \infty$$

$$\text{Then, } p(x \geq 7) = 1 - p(x \leq 6)$$

$$= 1 - 0.7622 \text{ (using Poisson distribution table)}$$

$$= 0.2378$$

10.5. Normal Distribution

So far we have discussed two discrete probability distributions. Now, we shall discuss a continuous probability distribution, which is known as normal distribution. It is a very important distribution in statistics. The normal distribution was discovered by De-Moivre as the limiting case of binomial distribution in 1733. The other eminent mathematicians were Laplace and Gauss who played important role in its development. In honour of the important contribution of Gauss, the normal distribution is often called the Gaussian distribution. The distribution of heights, weights, and errors made in measuring certain physical quantities, I.Q's are just a

few among the countless measurements whose distributions are normal. Two basic reasons why normal distribution occupies such a prominent place in statistics. First, it has an important property that most of the distributions can be converted into normal distribution under certain conditions that is known as central limit theorem. Second, it has wide range of practical applications in statistics.

Definition of Normal distribution: A continuous random variable X is said to have a normal distribution if its probability density function is defined by

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}; -\infty < x < \infty$$

where $\pi = 3.1416$; $e = 2.7183$. Here π and e are mathematical constants. μ and σ^2 are the two parameters of the distribution. Actually μ is the mean of the distribution and σ^2 is the variance of the distribution. It is symbolically expressed as $X \sim N(\mu, \sigma^2)$.

There are many ways we can get normal distributions.

10.5.2. Some important properties of normal distribution.

- 1) The distribution is symmetrical about μ .
- 2) Mean, median and mode of the distribution are equal.
- 3) The mean of the distribution is μ and the variance is σ^2 .
- 4) The curve has a single peak, i.e. it is unimodal.
- 5) $\mu \pm \sigma, \mu \pm 2\sigma, \mu \pm 3\sigma$, covers 68.27%, 95.45% and 99.73% area respectively.
- 6) All odd central moments of the distribution are zero.
- 7) For large sample most of the distributions tend to normal distribution.
- 8) Skewness of the distribution is zero. That is $\beta_1 = 0$.
- 9) The distribution is mesokurtic and the value of $\beta_2 = 3$.
- 10) $\mu \pm \sigma$ are the points of inflection of the curve.

10.5.3. Some important uses of normal distributions.

Normal distribution plays central role in the theory of statistics. Now, we shall cite some of its important applications both in theory and practice.

- 1) According to central limit theorem, if mean and variance of a distribution exist, then the distribution converted to normal distribution. This is the most important application of normal distribution.
- 2) Normal distribution is the basis of all sampling distributions. Without the assumption of normality, sampling distributions have no existence.
- 3) Assumption of normality is the basis of all parametric test of significance.

- 4) Normal distribution finds its application in industrial statistics such as quality control.

10.5.4. Standard normal distribution. A continuous random variable Z is said to have a standard normal distribution if its probability density function is defined by

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} ; -\infty < z < \infty$$

The variable Z is called standard normal variate. The mean of the distribution is zero and the variance is one. In symbols, it can be expressed as

$$Z \sim N(0, 1)$$

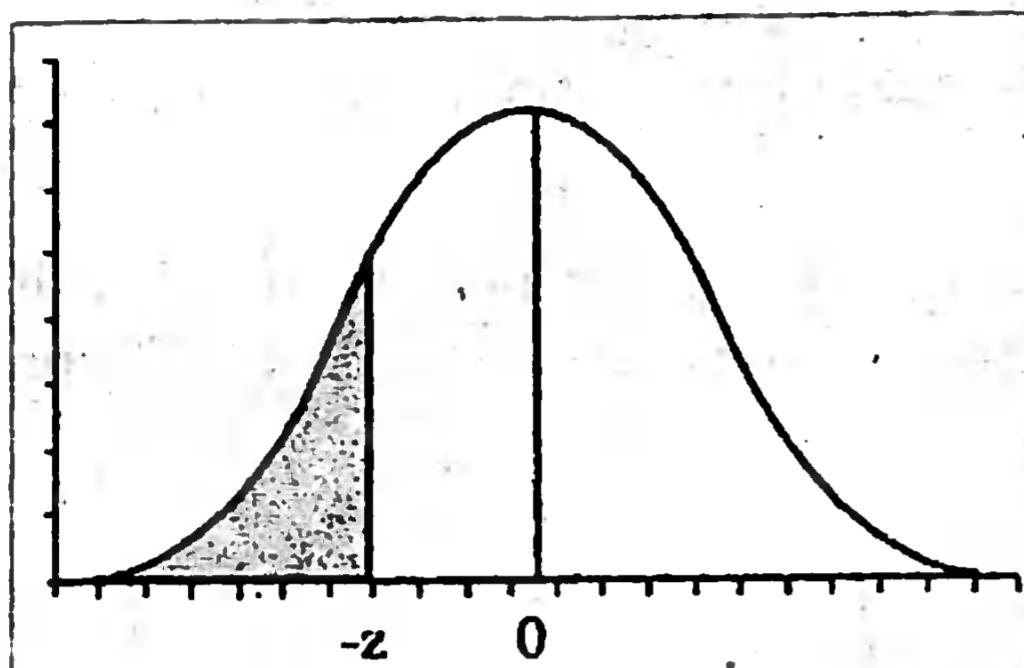
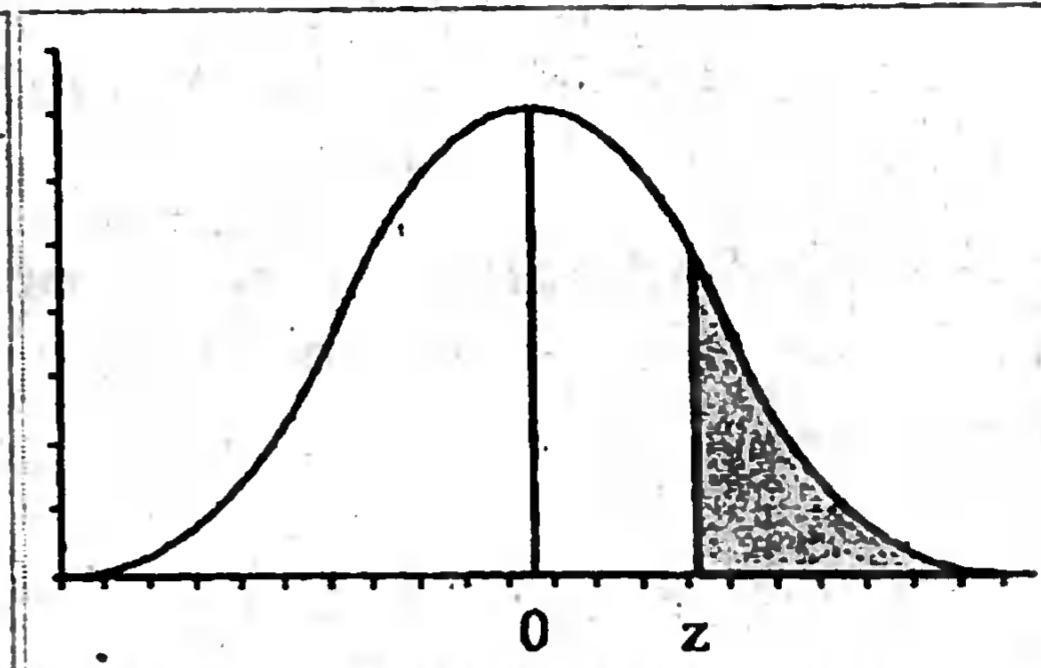
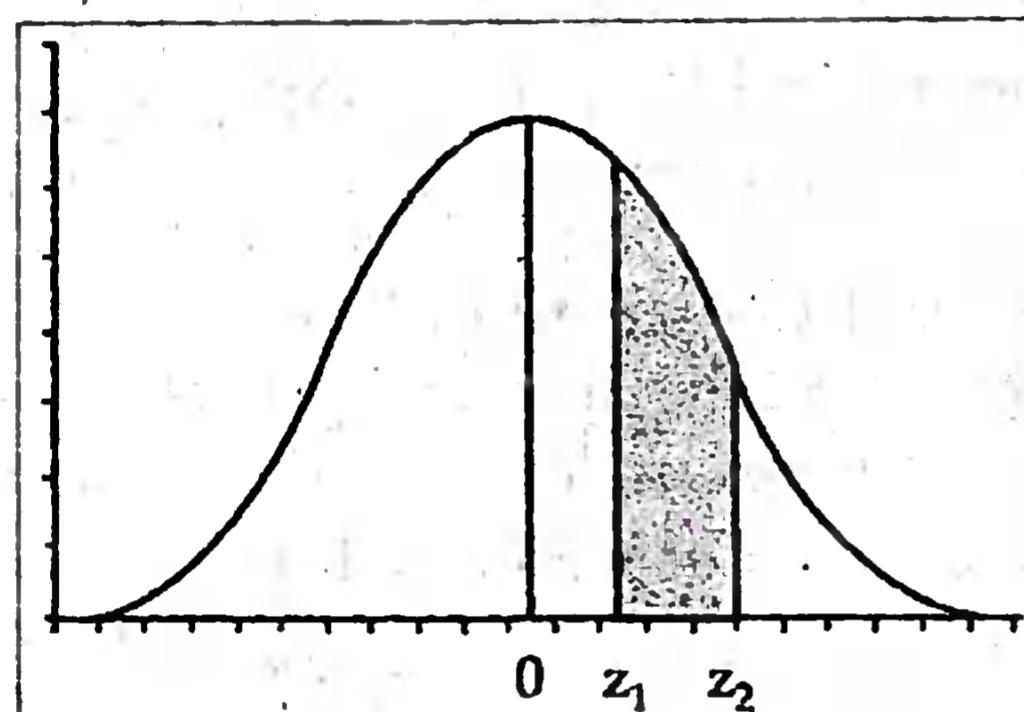
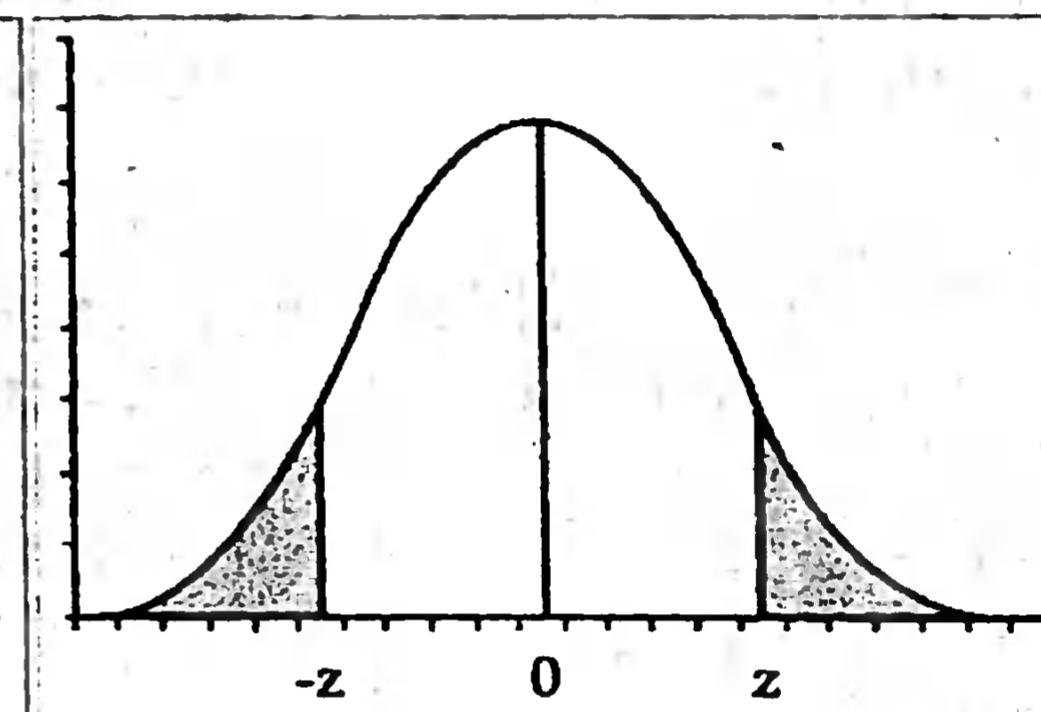
10.5.5. Finding probabilities for a normal distribution. Normal curve depends on mean and variance. Once mean and variance are specified, the normal curve is completely determined. The area under the normal curve between two ordinates depends upon the values of μ and σ^2 . It is difficult task to make normal integral tables for different values of μ and σ^2 . Fortunately, we are able to transfer any normal random variable to standard normal variate. This can be done by means of following transformation

$$Z = \frac{X - \mu}{\sigma}$$

It can be easily shown that $E[Z] = 0$ and $\text{Var}[Z] = 1$.

That is Z is normally distributed with mean zero and variance one. So normal variates with different means and variances can be converted into a standard normal variate. Hence a single table for a standard normal integral can serve to find the probability of normal distributions with different means and standard deviations. Table 1 gives the area under the standard normal curve corresponding to $P[Z \leq z]$ for values of Z from -3.4 to $+3.4$.

However, area for right tail such as $P[Z \geq z_1]$ is computed using the relation $P[Z \geq z_1] = 1 - P[Z \leq z_1]$, area between two values of z such as $P[z_1 \leq Z \leq z_2]$ is computed using the relation $P[z_1 \leq Z \leq z_2] = P[Z \leq z_2] - P[Z \leq z_1]$. Again, in some Tables area only for positive values of z are given, in that case, area corresponding to negative values can be computed using the relation $P[Z \leq -z_1] = 1 - P[Z \leq z_1]$.

Fig. 10.1. Area for $Z \leq z = P[Z \leq z]$ Fig. 10.2. Area for $Z \geq z = P[Z \geq z]$ Fig. 10.3. Area for $z_1 \leq Z \leq z_2$
= $P[z_1 \leq Z \leq z_2]$ Fig. 10.4. Area for $Z \leq -z = P[Z \leq -z]$
(where the two shaded areas are same)

Example 10.5.1. Let Z be standard normal variable. Find (i) $P[Z < 1.5]$; (ii) $P[Z > 2.4]$; (iii) $P[1.5 < Z < 2.14]$; $P[Z = 2]$

Solution. Here $Z \sim N(0, 1)$. Using table 1, we obtain

$$(i) P[Z < 1.5] = 0.9332;$$

$$(ii) P[Z > 2.4] = 1 - P[Z \leq 2.4] = 1 - 0.9918 = 0.0082;$$

$$(iii) P[1.5 < Z < 2.14] = P[Z < 2] - P[Z < 1.5] = 0.9838 - 0.9332 = 0.0506;$$

$P[Z = 2] = 0$; since the probability that continuous variable takes a particular value is zero.

Example 10.5.2. Given a normal curve with mean $\mu = 50$ and standard deviation $\sigma = 10$. Find the probability that X assumes a value between 45 and 62.

Solution. The value of Z corresponding to $x_1 = 45$ and $x_2 = 62$ are

$$z_1 = \frac{45 - 50}{10} = -0.5, \quad \text{and} \quad z_2 = \frac{62 - 50}{10} = 1.2.$$

$$\begin{aligned} \text{Therefore, } P[45 < X < 62] &= P[-0.5 < Z < 1.2] = P[Z < 1.2] - P[Z < -0.5] \\ &= 0.8849 - 0.3085 = 0.5764. \end{aligned}$$

Example 10.5.3. Suppose X is a normal variable with mean 8 and variance 4. Find (i) $P[X \geq 12]$; (ii) $P[X \leq 12]$; (iii) $P[0 < X < 8]$; and (iv) $P[x = 13]$

Solution. Here, we have a normal distribution with mean 8 and variance 4. Hence standard deviation is $\sigma = 2$.

$$\begin{aligned} \text{(i)} \quad P[X \geq 12] &= P\left[\frac{X-\mu}{\sigma} \geq \frac{12-\mu}{\sigma}\right] = P\left[\frac{X-8}{2} \geq \frac{12-8}{2}\right] \\ &= P[Z \geq 2] = 1 - P[Z < 2] = 1 - 0.9772 = 0.0228. \end{aligned}$$

(From the table)

$$\text{(ii)} \quad P[X \leq 12] = P\left[\frac{X-\mu}{\sigma} \leq \frac{12-\mu}{\sigma}\right] = P[Z \leq \frac{12-8}{2}] = P[Z \leq 2] = 0.9772$$

$$\text{(iii)} \quad P[0 < X < 8] = P\left[\frac{0-8}{2} < \frac{X-8}{2} < \frac{8-8}{2}\right] = P[-4 < Z < 0]$$

$$= P[Z < 0] - P[Z < -4] = 0.5000 - 0.0000 = 0.5000; \text{ and}$$

(iv) $P[x = 13] = 0$; since the probability that a continuous random variable takes a particular value is zero.

Example 10.5.4. An electric firm manufactures light bulbs that have a length of life that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a bulb burns between 778 and 834 hours.

Solution. Here the length of life of bulb follows normal distribution with mean, $\mu = 800$ hrs and a standard deviation, $\sigma = 40$ hrs. That is $X \sim N(\mu, \sigma^2)$. The value of Z corresponding to $x_1 = 778$ and $x_2 = 834$ are

$$z_1 = \frac{778 - 800}{40} = -0.55; z_2 = \frac{834 - 800}{40} = 0.85$$

$$\begin{aligned} \text{Hence, } P[778 < X < 834] &= P[-0.55 < Z < 0.85] \\ &= P[Z < 0.85] - P[Z < -0.55] = 0.8023 - 0.2912 = 0.5111. \end{aligned}$$

Example 10.5.5. Suppose the daily wages of workers of a factory follows normal distribution with mean wage of Tk. 500 and standard deviation of Tk. 100. (i) What is the percentage of workers getting daily wage between Tk. 400 and Tk. 650? (ii) If the number of workers in that factory is 1500, how many workers get daily wage between Tk. 400 and Tk. 650.

Solutions. (i) Here $X \sim N(400, 10000)$

First, we have to find $P[400 < X < 650]$.

$$\begin{aligned} P[400 < X < 650] &= P\left[\frac{400 - 500}{100} < \frac{X - 500}{100} < \frac{650 - 500}{100}\right] = P[-1 < Z < 1.5] \\ &= P[Z < 1.5] - P[Z < -1] = 0.9332 - 0.1587 = 0.7745 = 77.45\% \end{aligned}$$

Hence the percentage of workers getting daily wage between Tk.400 to Tk.650 is 77.45%

(ii) Number of workers getting daily wage between Tk.400 and Tk.650 is

$$0.7745 \times 15,000 = 11,618$$

Example 10.5.6. A manufacturing process produces light bulbs with life expectancies that are normally distributed with a mean of 500 hours and standard deviation of 100 hours. What percentage of the light bulbs can be expected to last between (i) 500 and 670 hours: (ii) 380 and 500 hours?

Solution. (i) Here $X \sim N(500, 10000)$

First, we have to find $P[500 < X < 670]$.

$$\begin{aligned} P[500 < X < 670] &= P\left[\frac{500 - 500}{100} < \frac{X - 500}{100} < \frac{670 - 500}{100}\right] = P[0 < Z < 1.7] \\ &= P[Z < 1.7] - P[Z < 0] = 0.9554 - 0.5000 \\ &= 0.4554 = 45.54\%. \end{aligned}$$

Hence 45.54% of the light bulbs will be expected to last between 500 and 670 hours.

(ii) The probability that light bulb can be expected to last between 380 and 500 hours is

$$\begin{aligned} P[380 < X < 500] &= P\left[\frac{380 - 500}{100} < \frac{X - 500}{100} < \frac{500 - 500}{100}\right] \\ &= P[-1.2 < Z < 0] \\ &= P[Z < 0] - P[Z < -1.2] \\ &= 0.5000 - 0.1151 = 0.3849 = 38.49\%. \end{aligned}$$

Thus, we conclude that the percentage of light bulbs expected to last between 380 and 500 hours is 38.49%.

10.5.6. Normal distributions from binomial. Normal distribution is obtained from binomial distribution under the following conditions:

- i) The probability of success p or the probability of failure is not so small.
- ii) n , the number of trials is very large i.e. n tends to infinity.

In fact standard binomial variate converted to standard normal variate under the above two conditions. Actually, it follows from the central limit theorem.

Normal approximation to the Binomial

The normal approximation to the binomial is fairly good if n is large ($n > 29$), p is not so small, $np > 5$, $nq > 5$ and $x \pm 3\sqrt{npq}$ lies in the interval $(0, n)$. In

this case X is normally distributed with mean $\mu = np$ and variance $\sigma^2 = npq$. When p or q is close to 0.5 and n is very large, then binomial distribution reduces normal distribution. The result can be stated in the following theorem:

Theorem. If X is a binomial random variable with parameters a and p and mean $\mu = np$ and variance $\sigma^2 = npq$, then the limiting form of the distribution of

$$Z = \frac{(X - 0.5) - np}{\sqrt{npq}}$$

as $n \rightarrow \infty$, is the standard normal distribution $N(Z; 0, 1)$.

10.5.7. Normal distribution from Poisson distributions. If the mean of the Poisson distribution tends to infinity then Poisson distribution tends to normal distribution. Similarly, in fact standard Poisson variate converted to standard normal variate under the above condition. Actually, the above two results follow from the central limit theorem.

Example 10.5.7. The probability that a patient recovers from a rare blood disease is 0.4. If 100 people are known to have contracted this disease, what is the probability that less than 30 survive?

Solution. Let the binomial variable X represent the number of patients that survive. Since $n=100$, we should obtain fairly accurate results using normal approximation with

$$\mu = np = (100)(0.4) = 40 \text{ and } \sigma = \sqrt{npq} = \sqrt{(100)(0.4)(0.6)} = 4.899.$$

Here mean is greater than 5, so we can safely use normal approximation to the binomial.

The value of Z corresponding to 29.5 is $z = \frac{29.5 - 40}{4.899} = -2.14$.

Hence, $P(X < 30) \approx P(Z < -2.14) = 0.0162$.

Example 10.5.8. An electric firm manufactures light bulbs that have a length of life that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. If 10 such bulbs are randomly selected from firm's production, find the probability that exactly 5 of them will have the life length between 778 and 834 hours.

Solution. To find the required probability, at first we have to compute the probability that a randomly selected bulb will have length of life between 778 and 834 hours, which is 0.5111 (from example 10.5.4).

Now, let Y be the number of bulbs which will burn between 778 and 834 hours, then according to the given information $Y \sim B(10, 0.5111)$ and we have to find $P(Y = 2)$.

We know the probability function of binomial distribution is given by

$$P[Y = y] = {}^n C_y p^y q^{n-y} ; y = 0, 1, 2 \dots \dots 10$$

Here, $n = 10$, $p = 0.5111$ and $q = (1 - 0.5111) = 0.4889$

$$\begin{aligned} \text{So, the required probability is } P[Y = 2] &= {}^{10} C_2 (0.5111)^2 (0.4889)^{10-2} \\ &= 0.038 \end{aligned}$$

Example 10.5.9. Suppose bolts of different weights are packed in the same type of boxes each containing a single bolt of particular size. The bolts are sold in auction and the buyer would have to collect boxes randomly. The weights of bolts in boxes follow normal distribution with mean 500 gm and standard deviation 100 gm. A buyer is interested to select the boxes with heavier bolts and s/he wants to take decision on the basis of probability that if 15 boxes are randomly selected, bolts of at least 9 boxes will weigh more than 500 gms. Find the probability for the buyer.

Solutions. (i) Let X be the weight of bolts. Here $X \sim N(400, 100^2)$

First, we have to find $P[X > 500]$.

$$P[X > 500] = P\left[\frac{X - 500}{100} > \frac{500 - 500}{100}\right] = P[Z > 0] = 0.5$$

Again, let Y be the number of boxes with weight of bolts more than 500 gms, thus $Y \sim B(15, 0.5)$

We know, $P[Y = y] = {}^n C_y p^y q^{n-y} ; y = 0, 1, 2 \dots \dots 15$

So, the probability for the buyer is given by

$$P(Y \geq 9) = P(Y = 9) + \dots + P(Y = 15) = 0.3056$$

The buyer will have to take decision on the basis of this probability.

Questions

1. What is a Bernoulli trial? Define binomial distribution. State some of its important properties.
2. What is a binomial distribution? Under what conditions will binomial distribution tend to normal distribution?

3. What is a Poisson distribution? State some of its important properties. Cite some practical applications of this distribution.
4. Define a normal distribution. Under what conditions binomial distribution tends to normal distribution. State some of its important properties.
5. Define a Poisson distribution. Under what conditions it tends to normal distribution. Cite some practical examples of Poisson distributions.
6. What is a binomial experiment? Define a binomial distribution. How it is related with Poisson distribution.

Exercise

7. The mean and standard deviation of a binomial distribution are 20 and 4 respectively. Find the values n and p, and hence write the form of the probability distribution.
Ans. $p = 0.2$ and $n = 100$. $P[X=x] = \binom{n}{x} p^x (q)^{n-x} = \binom{100}{x} p^x (q)^{100-x}$; $x = 0, \dots, 100$
8. For a binomial distribution with $n = 10$ and $p = 0.30$. Find (i) $P[X=0]$; (ii) $P[X=4]$; (iii) $P[X=2]$; (iv) $P[X \geq 8]$
Ans. (i) 0.0003; (ii) 0.2001; (iii) 0.2759; (iv) 0.0016.
9. Find mean, variance and standard deviation of the following binomial distributions: (i) $n = 8$, $p = 0.5$; (ii) $n = 15$, $p = 0.20$.
Ans. (i) 4, 2 and approx 1.4; (ii) 3, 2.4 and 1.549
10. For a Poisson distribution with mean $\lambda = 4$, find (i) $P[X = 6]$; (ii) $P[X < 3]$; (iii) $P[X > 2]$
Ans. (i) 0.1042; (ii)
11. For a binomial distribution with $n = 8000$ and $p = 0.001$, use the Poisson approximation to the binomial, find the mean of the Poisson distribution and then find (i) $P[X=3]$; (ii) $P[X = 5]$ and $P[X \leq 7]$
Ans. $\lambda = 8$, (i) 0.0286; (ii) 0.0916 (iii) 0.4530
12. Given $\lambda = 6.1$ for a Poisson distribution, find (i) $P[X \leq 3]$; (ii) $P[X \geq 2]$; (iii) $P[X=6]$; (iv) $P[1 \leq X \leq 4]$. Ans. 0.1424; (ii) 0.9841; (iii) 0.1605; (iv) 0.2696.

Application

13. The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease, what is the probability that (i) at least 10 survive; (ii) from 3 to 8 survive and (iii) exactly 5 survive?
Ans. (i).0338; (ii) 0.8779; (iii) 0.1859.
14. Suppose the probability that a graduate from public university gets an executive job is 0.55, a graduate from a private university gets the same type of job is 0.30 and a graduate from the national university gets a job is 0.15. 10 graduates have been working in a company, what is the probability that (i) four of them are from private University; (ii) none of

them are from public university; (ii) two are from National University; (iii) at least 8 are from private university.

Ans. (i) 0.2001, (ii) 0.0003; (iii) 0.2759; (iv) 0.0016

15. A manufacturer produces some parts of car., finds that 0.1% of the parts are defective. The parts are packed in boxes containing 500. A repairing company buys 100 boxes from the producer.

(i) What is the appropriate probability model to find the following probabilities:
(ii) $P[X=0]$; (iii) $P[X < 2]$ and also find the number of boxes with (iv) $P[X = 0]$ and (v) $P[X < 2]$.

Ans. Here $p = 0.001$ and $n = 500$. since p is very small and n is very large, then the appropriate probability model is Poisson with mean $n \times p = 500 \times 0.001 = 0.5$. (ii) 0.6065; (iii) 0.0025; (iv) 61 and (v) 10.

16. In a town 10 accidents took place in a span of 50 days Assuming that the number of accidents per day follows Poisson distribution, find the probability that there will be three or more accidents in a particular day. Ans. 0.0012

17. Suppose 20% bolts produced by a machine are defective. Find the probability that out of 4 bolts (i) 0; (ii) 1 and (iii) at most 2 bolts will be defective. Ans. (i) 0.4096 ; (ii) 0.4096 (iii) 0.9728

18. The mean and standard deviation for the life times of a population of light bulbs are 1200 and 50 hours respectively. Assuming these life times are normally distributed, what is the probability that a light bulb will last over 1500 hours? Ans. 0.0228

19. The time required by a bank cashier to deal with a customer has been observed to be normally distributed with mean 25 seconds, and a standard deviation 10 seconds. Find the probability that a customer arriving at random will have to wait
(i) between 20 and 28 seconds;
(ii) less than 23 seconds. Ans. (i) 0.3094; (ii) 0.4207.

20. The chances of a bomber hitting the target and missing it are 3:2. Compute the probability that in five sorties, the target will be hit (i) exactly once, (ii) at least once Ans. (i) (ii) 0.98976

21. A multiple choice test consists of 8 questions with 3 options to each question, of which one is correct. A student was not at all prepared for such test and answers each question randomly. To qualify the test, the student must secure at least 75 percent correct answers. Suppose there is no negative marking, what is the probability that the student secures pass mark? Ans. 0.019

22. Suppose a completely unprepared candidate has appeared in a true/false type of test with 10 questions each carrying 2 mark. Assuming that the student answers all the questions at random. (i) Find the probability that the student gets all the answers correct, (ii) If pass

mark in the test is 16, find the probability that the student qualifies the test

Ans. i) 0.00098; ii) 0.0547

(hints: here the probability of answering a question correctly is 0.5 and to qualify the test, the student must answer at least 8 question correctly).

23. A shop sells television at a rate of 2.5 per week. (i) Find the probability that in a two-week period the shop sells at least 7 TVs. (ii) If the delivery of the TVs is made monthly, find the probability of selling fewer than 12 TVs in a month.
24. Suppose the weights of apples in a shop are normally distributed with mean 100 gm and variance 144 gm^2 . A buyer wants to buy 12 apples of medium size and with almost homogeneous weights for a particular purpose. Find the probability that all the apples will be between 112 gm and 120 gm.

Ans. i) 0.2378; ii) 0.6968

CHAPTER - 11

SIMPLE CORRELATION

11.1. Introduction

So far we have discussed different characteristics of a single variable. For example, demand of a commodity over a period of time, income of a number of families, volume of sales by a number of salesmen, etc. The analysis with a single variable is termed as uni-variate analysis. In the real field two or more variables may be interrelated. There are many situations in business where we are interested to measure the relationship between two variables such as the income and expenditure of a certain class of people, price of a commodity and amount demanded, volume of sales and the experience of the salesman of a departmental store, deposit in a bank and number of clients, family income and expenditure on luxury items, the fertilizer used and production of certain crop, etc. Pairs of observations of two such variables produce a bi-variate distribution. It is often required to know how stronger the relationship between two such variables is or it might be required to know the impact of change in one variable on another variable, or it may be required to forecast the value of one variable for a particular given value of another. The study of these types of relationship can be performed through two types of statistical tools, viz the correlation analysis and regression analysis. In this chapter, we will discuss different aspects of correlation analysis for two variables.

11.2. Correlation Analysis

In this section we shall consider the problem of measuring the relationship between two quantitative variables. In business we come across a large number of problems involving the use of two or more related variables. For example, there exists some relationship between family income and expenditure on luxury items, price of a commodity and amount demanded, price of a commodity and the amount supplied, advertisement expenditure of a commodity and amount sold etc. The statistical tool with the help of which these relationships between two or more than two variables can be studied is called correlation.

A.M. Tuttle gives a very simple definition of correlation. According to A.M. Tuttle "an analysis of covariation of two more variables is usually called correlation."

Covariation or relationship between two variables is measured by covariance. We shall define covariance both for population and sample.

Population covariance. Suppose $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ are N pairs of values of two variables X and Y with respective means μ_x and μ_y , then population covariance between X and Y, denoted by μ_{11} is defined as

$$\text{Cov}(X, Y) = \mu_{11} = \frac{\sum(X - \mu_x)(Y - \mu_y)}{N}$$

Population covariance can also be defined for two random variables.

Population covariance. Suppose X and Y are two random variables with means $E[X]$ and $E[Y]$, then population covariance is defined by

$$\mu_{11} = E[X - E(X)][Y - E(Y)]$$

Sample covariance. Suppose $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are n pairs of values of two variables x and y with respective means \bar{x} and \bar{y} , then sample covariance between x and y is defined by

$$\text{Cov}(x, y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{n}$$

Working formula for finding sample covariance is

$$\text{Cov}(x, y) = \frac{1}{n} \left\{ \sum xy - \frac{(\sum x)(\sum y)}{n} \right\}$$

Some important properties of covariance

- (i) The range of covariance is $-\infty$ to ∞ .
- (ii) It depends on the units of measurements on which the variables are measured.
- (iii) It gives the magnitude and direction of the statistical relationship between two variables.
- (iv) The value of covariance will be positive if the increase or decrease of one variable is associated with the increase or decrease of the other variable.
- (v) The value of covariance will be negative if the increase or decrease of one variable is associated with the decrease or increase of the other variable.
- (vi) The value of covariance is zero if the two variables are linearly independent.

Generally, there are three types of correlation. They are:

- (i) Simple correlation;
- (ii) Partial correlation; and
- (iii) Multiple correlations.

Here we shall discuss only simple correlation.

11.3. Simple Correlation

If only two variables are chosen to study the correlation between them, then such a correlation is referred to as simple correlation. The most widely used measure of linear relationship between two variables is called Karl Pearson product-moment correlation coefficient or simply the correlation coefficient, which is better than covariance as a measure of relationship between two variables.

Simple Correlation coefficient. Simple correlation coefficient is a quantitative measure of the strength and direction of linear relationship between two numerically measured variables.

We shall define correlation coefficient both for population and sample.

Population simple correlation coefficient. Population simple correlation coefficient measures the strength of linear relationship between two variables of a bivariate population. Suppose $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ are N pairs of values of two variables X and Y of a bi-variate population with means μ_x and μ_y , then the population correlation coefficient denoted by ρ is defined by Karl Pearson as

$$\rho = \frac{\sum (X - \mu_x)(Y - \mu_y)}{\sqrt{\sum (X - \mu_x)^2 \sum (Y - \mu_y)^2}}$$

Simple population correlation coefficient can also be defined for two random variables of a bi-variate population.

Population correlation coefficient. Simple Population correlation coefficient also measures the strength of linear relationship between two random variables. Suppose X and Y are two random variables of a bi-variate population, and then population correlation denoted by ρ is defined as

$$\rho = \frac{E[X - E(X)][Y - E(Y)]}{\sqrt{E[X - E(X)]^2 E[Y - E(Y)]^2}} = \frac{\mu_{11}}{\sigma_x \sigma_y}$$

where $\sigma_x^2 = E[X - E(X)]^2$ and $\sigma_y^2 = E[(Y - E(Y))^2]$.

Actually, $E(X)$ and $E(Y)$ are the population means of X and Y, and σ_x^2 and σ_y^2 are population variances of the random variables X and Y which were defined in chapter 9.

Simple sample correlation coefficient. Simple sample correlation coefficient measures the strength of linear relationship between two variables when a bi-variate sample is taken from a bi-variate population. Suppose $(x_1, y_1), (x_2,$

$y_2), \dots, (x_n, y_n)$ are n pairs of sample values of two variables x and y from a bi-variate population. Let \bar{x} and \bar{y} be the sample means of x and y . Then the sample correlation between x and y , denoted by r is defined as

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Actually, it is defined with the help of the sample covariance and variances of x and y as

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})/n}{\sqrt{\sum(x - \bar{x})^2/n \sum(y - \bar{y})^2/n}} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)}}$$

The sample correlation coefficient r is used to estimate the population correlation coefficient ρ . So sample correlation coefficient is important in correlation analysis. Usually, correlation coefficient is computed with any one of the following formulae

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}} = \frac{n \sum xy - \sum x \sum y}{\sqrt{\{n \sum x^2 - (\sum x)^2\} \{n \sum y^2 - (\sum y)^2\}}}$$

These two formulae are used to compute the value of r . They are also known as the working formula for finding the value of r . The value of correlation coefficient r or ρ lies between -1 to $+1$.

We shall use sample notation r for the further correlation analysis. The value of correlation coefficient may be positive, negative or zero.

11.3.1. Examples of positive correlation coefficient. The value of correlation coefficient between two variables will be positive if the increase or decrease of one variable is associated with the increase or decrease of the other variable. Some examples where correlation coefficient give positive value:

- (i) The heights and weights of a group of persons.
- (ii) The income and expenditure of a certain class of people.
- (iii) The fertilizer used and the production of certain crop.
- (iv) The amount of sales and the experience of the salesman of a departmental store.
- (v) The deposit in a bank and the number of clients.
- (vi) The ages of husbands and the ages of wives etc.

11.3.2. Examples of negative correlation coefficient. The value of correlation coefficient will be negative if the increase or decrease of one variable is associated with the decrease or increase of the other variable.

11.3.3. Some examples of negative correlation coefficients are:

- (i) The price and demand of a commodity, as the price of a product increases the demand for that product decreases. The demand of mobile set increases as the price decreases.
- (ii) The volume of a perfect gas increases as the pressure decreases.
- (iii) The price of a commodity decreases as the supply increases.

11.3.4. Some Examples of independence of two variables are:

- (i) The rainfall of Bangladesh and the production of rice in Vietnam.
- (ii) The demand of some commodities does not depend on the increase or decrease of the prices of the commodities. For example, salt, oil, rice etc are such commodities. They are known as perishable goods.
- (iii) The heights and ages of University students.
- (iv) The price of gasoline and the rainfall etc.

Remarks. When two variables are linearly independent, then the value of correlation coefficient is zero. But $r = 0$ does not mean that the two variables x and y are not related. For example, the correlation coefficient between x and y is zero when $y = x^2$. Here x and y are not linearly related. Actually, this is the equation of a parabola.

11.4. Assumption Underlying Karl Pearson's Correlation Coefficient or Simple Correlation Coefficient

The simple correlation coefficient r is based on the following assumptions:

- (i) The relationship between the variables is linear.
- (ii) Both the variables are measured on interval or ratio scales
- (iii) The two variables follow bi-variate normal distribution
- (iv) The sample is of adequate size to assume normality.

11.5. Some Important Properties of Correlation Coefficient

1. The value of r lies between -1 to $+1$.
2. It measures the magnitude and direction of statistical relationship between two variables.
3. It is a pure number. That is, it is independent of the units of measurements of the variables.
4. It is a symmetrical function of x and y . That is $r_{xy} = r_{yx}$.
5. $r = 0$ indicates no linear relationship between x and y .
6. $r = +1$ indicates a perfect positive relationship between x and y .
7. $r = -1$ indicates a perfect negative relationship between x and y .
8. The geometric mean of the two regression coefficients is equal to the correlation coefficient.
9. The correlation coefficient is independent of the shift of origin and change of scale.

Now, we shall prove some important properties of correlation coefficient.

Theorem 11.5.1. The value of correlation coefficient lies between -1 to +1.

Proof. Suppose $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are n pairs of values of a bivariate sample. The correlation coefficient between x and y is

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2} \sqrt{\sum(y - \bar{y})^2}} = \frac{\sum XY}{\sqrt{\sum X^2} \sqrt{\sum Y^2}}$$

Here $X = (x - \bar{x})$ and $Y = (y - \bar{y})$ and \bar{x} and \bar{y} are means x and y .

Let us consider the following expression which is always positive.

$$\begin{aligned} & \sum \left(\frac{X}{\sqrt{X^2}} + \frac{Y}{\sqrt{Y^2}} \right)^2 \geq 0 \\ \text{or, } & \sum \left(\frac{X^2}{\sum X^2} + \frac{Y^2}{\sum Y^2} + \frac{2XY}{\sqrt{\sum X^2} \sqrt{\sum Y^2}} \right) \geq 0 \\ \text{or, } & \frac{\sum X^2}{\sum X^2} + \frac{\sum Y^2}{\sum Y^2} + 2r \geq 0 \\ \text{or, } & 1 + 1 + 2r \geq 0 \\ \text{or, } & r \geq -1 \end{aligned} \quad \dots \quad (11.5.1)$$

Again

$$\begin{aligned} & \sum \left(\frac{X}{\sqrt{X^2}} - \frac{Y}{\sqrt{Y^2}} \right)^2 \geq 0 \\ \text{or, } & \sum \left(\frac{X^2}{\sum X^2} + \frac{Y^2}{\sum Y^2} - \frac{2XY}{\sqrt{\sum X^2} \sqrt{\sum Y^2}} \right) \geq 0 \\ \text{or, } & \frac{\sum X^2}{\sum X^2} + \frac{\sum Y^2}{\sum Y^2} - 2r \geq 0 \\ \text{or, } & 1 + 1 - 2r \geq 0 \\ \text{or, } & r \leq 1 \end{aligned} \quad \dots \quad (11.5.2)$$

From (11.5.1) and (11.5.2), we have

$$-1 \leq r \leq 1$$

This is the proof of the theorem.

Theorem 11.5.2. The value of the correlation coefficient is 1 when $y = a + bx$

Proof. The correlation coefficient between x and y is

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}}, \quad \dots \dots \dots (11.5.3)$$

Here $y = a + bx$ and $\bar{y} = a + b\bar{x}$

Now, put the values of y and \bar{y} in (11.5.3), we have

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} = \frac{\Sigma(x - \bar{x})(a + bx - a - b\bar{x})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y + bx - a - b\bar{x})^2}} = \frac{b\Sigma(x - \bar{x})^2}{b\Sigma(x - \bar{x})^2} = 1$$

This proof the theorem.

Theorem 11.5.3. The value of the correlation coefficient is -1 when $y = a - bx$

Proof. The correlation coefficient between x and y is

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} \quad \dots \dots \dots (11.5.4)$$

Here $y = a - bx$ and $\bar{y} = a - b\bar{x}$

Now, put the values of y and \bar{y} in (11.5.4), we have

$$R = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} = \frac{\Sigma(x - \bar{x})(a - bx - a + b\bar{x})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - bx - a + b\bar{x})^2}} = \frac{-b\Sigma(x - \bar{x})^2}{b\Sigma(x - \bar{x})^2} = -1$$

This proof the theorem.

Theorem 11.5.4. Correlation coefficient is independent of the shift of origin and change of scale.

Proof. Suppose x and y are two variables. Now we shall define two new variables

$$u = \frac{x - A}{h} \quad \text{and} \quad v = \frac{y - B}{k}$$

This means we have shift the origin of x to A and y to B . Also we have change the scale of x by dividing it by h and y by k .

$$\text{That is } x = A + hu \Rightarrow \bar{x} = A + h\bar{u}$$

$$y = B + kv \Rightarrow \bar{y} = B + k\bar{v}$$

The sample correlation coefficient is defined by

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2(y - \bar{y})^2}}.$$

Now by putting the values of x , \bar{x} , y , \bar{y} in the above theorem, we have

$$\begin{aligned} r_{xy} &= \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2(y - \bar{y})^2}} = \frac{\sum(A + hu - A - h\bar{u})(B + kv - B - k\bar{v})}{\sqrt{\sum(A + hu - A - h\bar{u})^2 \sum(B + kv - B - k\bar{v})^2}} \\ &= \frac{hk \sum(u - \bar{u})(v - \bar{v})}{hk \sqrt{(u - \bar{u})^2(v - \bar{v})^2}} = r_{uv}. \end{aligned}$$

It is one of the important properties of correlation coefficient. We shall now give some of its applications.

Remarks. (i) The property says if a constant value A and a constant value B are subtracted from x and y respectively then the correlation coefficient of the new variables u and v is the same as the correlation coefficient between the original variables x and y . This means correlation coefficient is independent of the shift of origin.

(ii) If the values of x and the values of y are very large, we can divide each value of x by a constant value h and each of y by k , then the correlation coefficient between the two new variables is the same as the original variables. This means correlation coefficient is independent of the change of scale.

Very often we shift the origins and change the scales of both the variables to get the maximum computations benefit. Now we shall cite some examples.

The above formula is very useful to find the correlation coefficient from a bi-variate frequency distribution which is not given in this book.

11.6. Scatter Diagram:

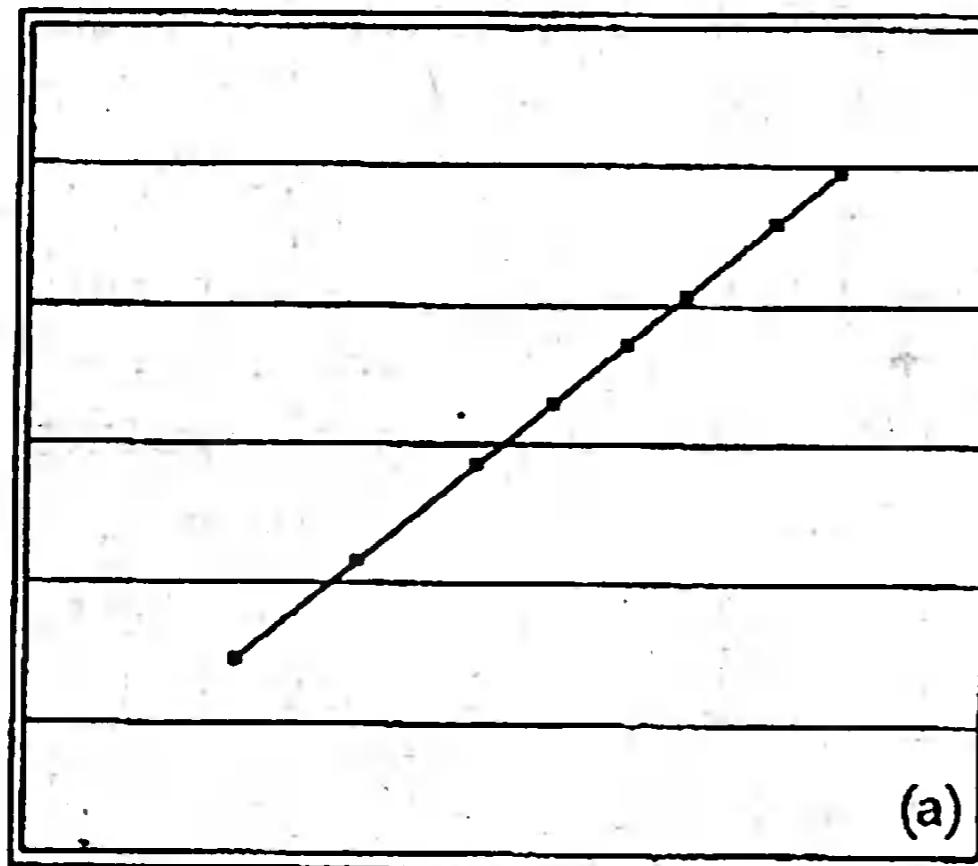
The nature of association between two variables may be observed by plotting pairs of observations of the variables on a graph, such a graph is known as a scatter diagram or scatter plot, since it depicts how two variables are related or the pairs of points are scattered. The scatter plot is an essential and important step in studying the relationship between two variables.

Definition. Scatter Diagram. The simplest device for showing the relationship between two variables on a graph paper in the form of dots is called scatter diagram or scatter plot.

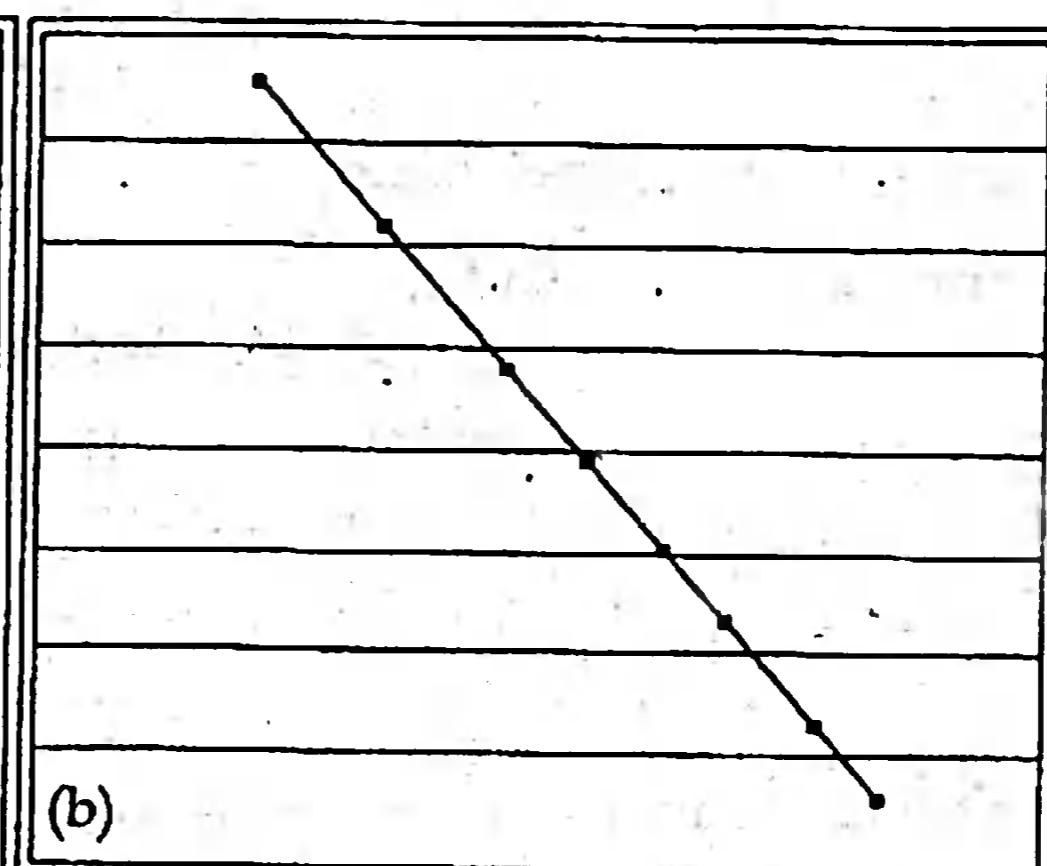
Actually it gives a rough idea about the relationship between two quantitative variables. Suppose $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are n pairs of sample values of two variables x and y . If the values of the variables x and y be plotted along the x -axis and y -axis respectively in the xy -plane, the

diagram of dots so obtained is known as scatter diagram. Actually it portrays the relationship between these two variables graphically. By looking at the scatter of the various points on the chart, it is possible to determine the extent of association between these two variables. The wider the scatter on the chart, the less close is the relationship. On the other hand, the closer the points and the closer they come to falling on a line passing through them, the higher the degree of relationship. If all the points fall on a line, the relationship is perfect. If this line goes up from the lower left hand corner to the upper right hand corner, i.e., if the slope of the line is positive, then the correlation between the two variables is considered to be perfect positive and the value of r is +1. Similarly, if the line starts at upper left hand corner and comes down to the lower right hand corner of the diagram, i.e., if the slope is negative, and also all the points fall on the line, then their correlation is said to be perfect negative and the value of r is -1.

Scatter diagrams for different values of r are as follows:



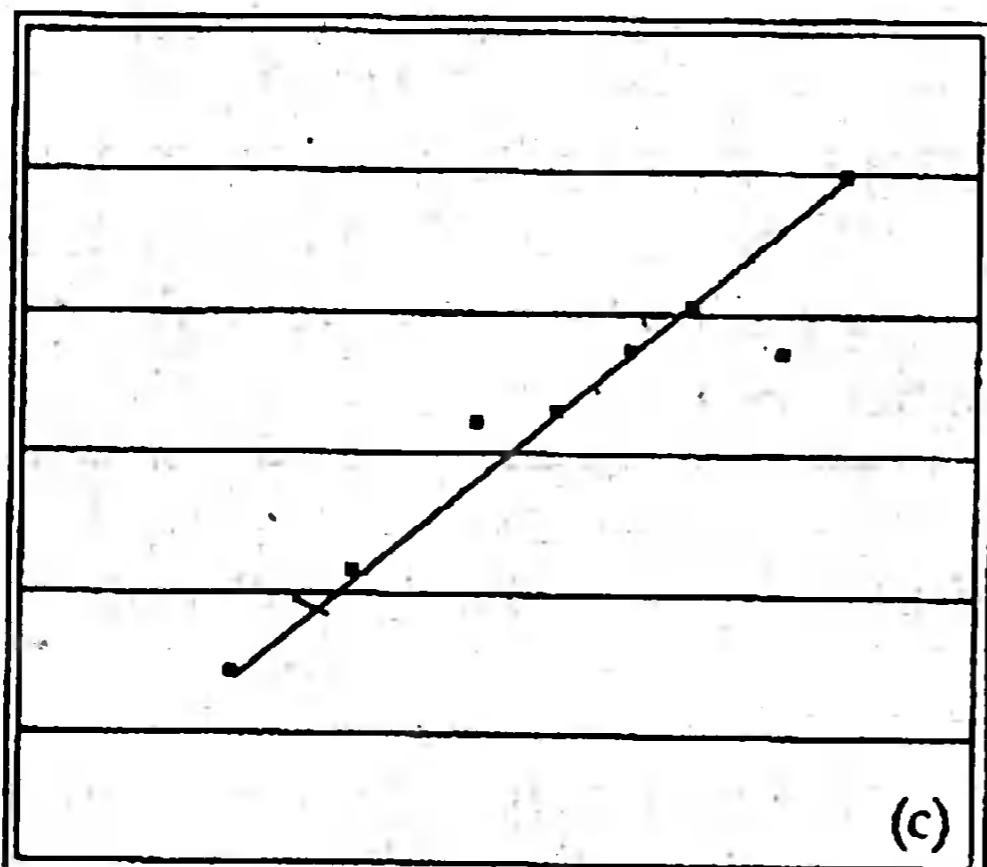
(a)



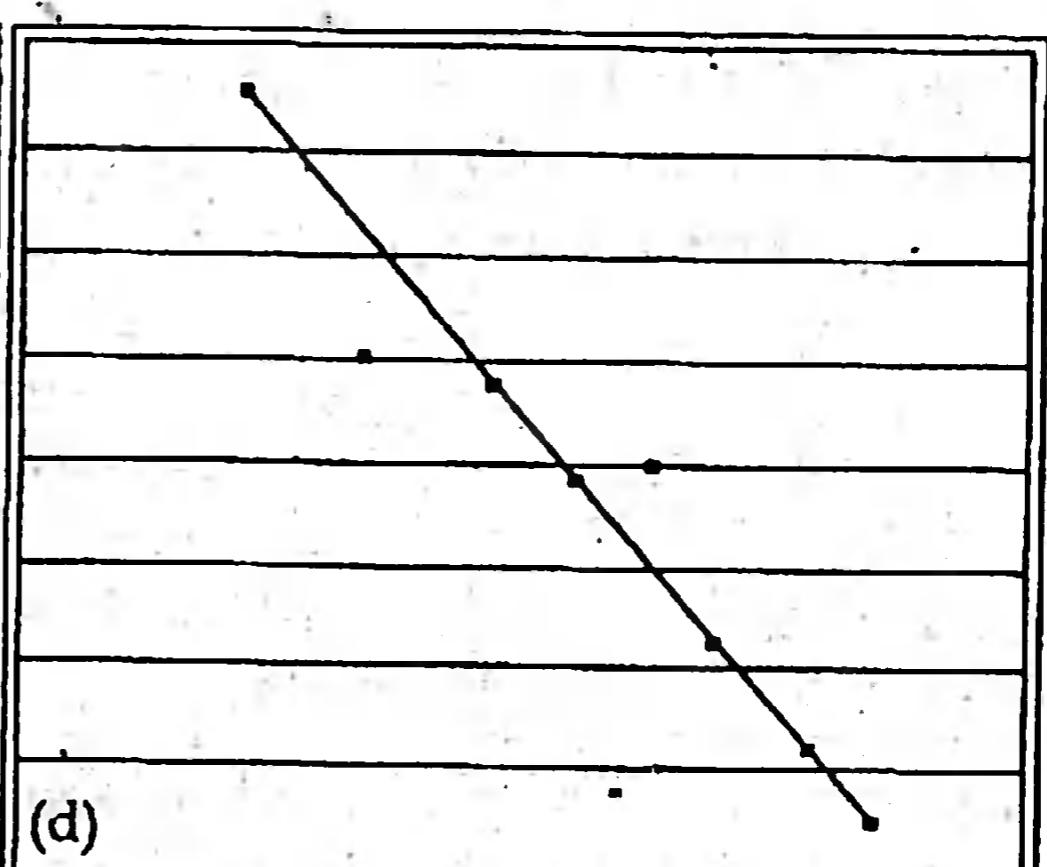
(b)

Fig 11.6.1a. Scatter Diagram showing $r = +1$

Fig 11.6.1b. Scatter Diagram showing $r = -1$



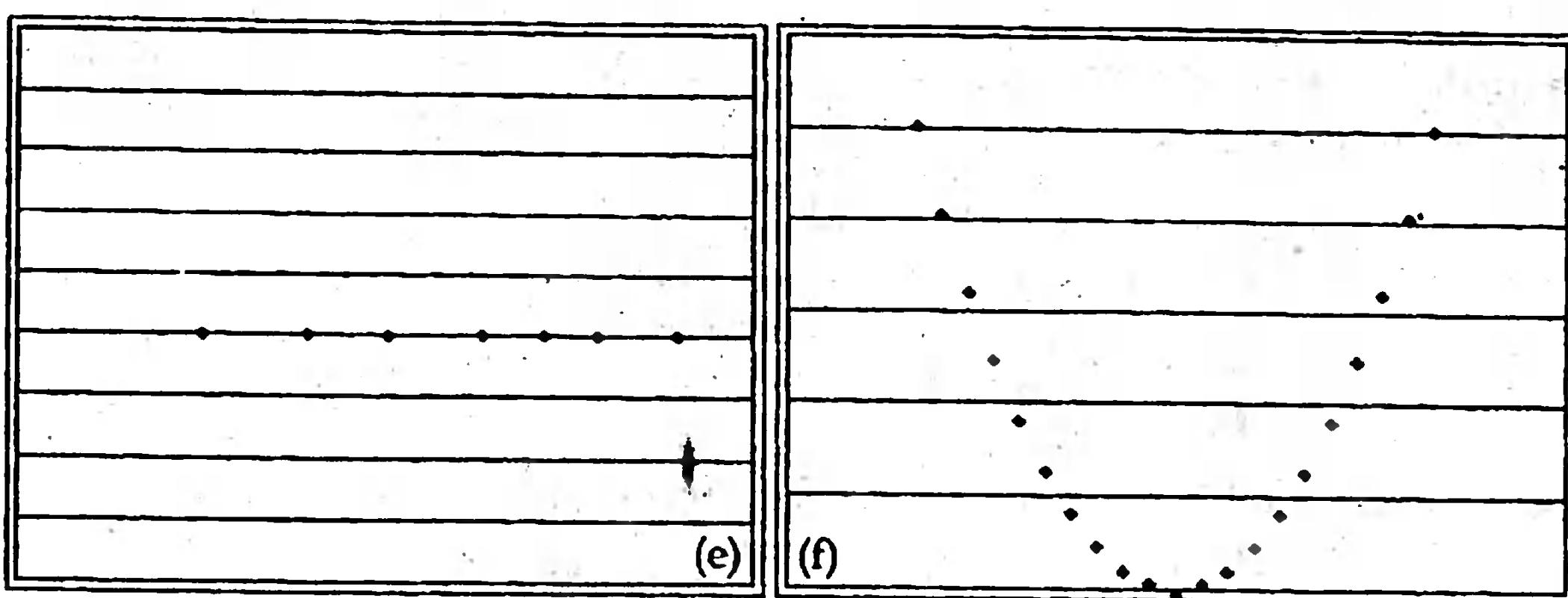
(c)



(d)

Fig 11.6.1c. Scatter Diagram showing $0 < r < 1$

Fig 11.6.1d. Scatter Diagram showing $-1 < r < 0$

Fig 11.6.1e. Scatter Diagram showing $r = 0$ Fig 11.6.1f. Scatter Diagram showing $y = x^2$

Interpreting the values of r

1. $r = +1$ indicates a perfect positive relationship between x and y . In this case all the values of x and y fall in a straight line and the scatter diagram will be as in fig. 11.6.1a .The mathematical relationship between x and y is $y = a + bx$.
2. $r = -1$ indicates a perfect negative relationship between x and y . In this case all the values of x and y fall in a straight line and the scatter diagram will be as in fig.11.6.1b. The mathematical relationship between x and y is $y = a - bx$.
3. $r = 0$ means there is no linear relationship between the variables x and y . In this case the two variables are linearly independent. The scatter diagram will be as fig.11.6.1.e and Fig.11.6.1f.
4. The closer r to $+1$ or -1 , the closer the relationship between the variables x and y . The closer r to zero, the less close the relationship. In these cases the scatter diagrams will be as fig.11.6.1.c and 11.6.1d.

Now we shall cite some examples to show the different values of r .

Example 11.6.1. Compute r for the following paired sets of values:

a	x	1	2	3	4	5
	y	1	3	5	7	9
b	x	1	2	3	4	5
	y	2	3	5	4	7
c	x	1	2	3	4	5
	y	10	8	6	4	2
d	x	2	3	4	5	6
	y	9	5	6	2	1
e	x	1	2	3	4	5
	y	3	2	2	2	3
f	x	-2	-1	0	1	2
	y	4	1	0	1	4

Solution. (a) The computing formula for finding Karl Pearson's correlation coefficient is

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}}$$

Let us make a table to calculate correlation coefficient

X	y	x ²	y ²	xy
1	1	1	1	1
2	3	4	9	6
3	5	9	25	15
4	7	16	49	28
5	9	25	81	45
15	25	55	165	95

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}} = \frac{95 - \frac{15 \times 25}{5}}{\sqrt{55 - \frac{(15)^2}{5}} \sqrt{165 - \frac{(25)^2}{5}}} \\ = \frac{20}{\sqrt{10 \times 40}} = \frac{20}{20} = 1$$

Conclusion. Here there exists a perfect and positive relationship between x and y. In this case the increment of the value of y for unit change of x is the same. All the points of the scatter diagram will fall on a straight line. The mathematical relationship between x and y is $y = a + bx$ or $y = 5 + 3x$.

(b) The formula for finding correlation coefficient is

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}}$$

Let us make a table to calculate correlation coefficient

X	y	x ²	y ²	xy
1	2	1	4	2
2	3	4	9	6
3	5	9	25	15
4	4	16	16	16
5	7	25	49	35
15	21	55	103	74

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}} = \frac{74 - \frac{15 \times 21}{5}}{\sqrt{\left\{ 55 - \frac{(15)^2}{5} \right\} \left\{ 103 - \frac{(21)^2}{5} \right\}}}$$

$$= \frac{11}{\sqrt{10 \times 14.8}} = \frac{11}{12.17} = 0.90$$

Conclusion. There exists a strong positive relationship between x and y.

(c) The computing formula for finding Karl Pearson's correlation coefficient is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}}$$

Let us make a table to calculate correlation coefficient

X	y	x^2	y^2	xy
1	10	1	100	10
2	8	4	64	16
3	6	9	36	18
4	4	16	16	16
5	2	25	4	10
15	30	55	220	70

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}} = \frac{70 - \frac{15 \times 30}{5}}{\sqrt{\left\{ 55 - \frac{(15)^2}{5} \right\} \left\{ 220 - \frac{(30)^2}{5} \right\}}}$$

$$= \frac{-20}{\sqrt{10 \times 40}} = \frac{-20}{20} = -1$$

Conclusion. Here there exists a perfect negative relationship between x and y. The decrease of the value of y for unit change of x is the same. All the points of the scatter diagram will fall on a straight line, which starts at upper left hand corner and comes down to the lower right hand corner of diagram. The mathematical relationship between x and y is $y = a - bx$ or $y = 3 - 2x$.

(d) The working formula for finding correlation coefficient is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}}$$

Let us make a table to calculate correlation coefficient

X	y	x^2	y^2	xy
2	9	4	81	18
3	5	9	25	15
4	6	16	36	24
5	2	25	4	10
6	1	36	1	6
20	23	90	147	73

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}} = \frac{73 - \frac{20 \times 23}{5}}{\sqrt{\left\{ 90 - \frac{(20)^2}{5} \right\} \left\{ 147 - \frac{(23)^2}{5} \right\}}}$$

$$= \frac{20}{\sqrt{10 \times 41.2}} = \frac{-19}{20.3} = -0.94$$

Conclusion. There exists a strong negative relationship between x and y.

(e) The computing formula for finding r is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}}$$

Let us make a table to calculate correlation coefficient

x	Y	x^2	y^2	xy
1	3	1	9	3
2	2	4	4	4
3	2	9	4	6
4	2	16	4	8
5	3	25	9	15
15	12	55	30	36

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}} = \frac{36 - \frac{15 \times 12}{5}}{\sqrt{\left\{ 55 - \frac{(15)^2}{5} \right\} \left\{ 36 - \frac{(12)^2}{5} \right\}}}$$

$$= \frac{0}{\sqrt{10 \times 7.2}} = 0.00.$$

Conclusion. Here there exists no linear relationship between the variables x and y . That is the variables x and y are linearly independent.

(f) The formula for finding correlation r is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}}$$

Let us make a table to calculate correlation coefficient

X	y	x^2	y^2	xy
-2	4	4	16	-8
-1	1	1	1	-1
0	0	0	0	0
1	1	1	1	1
2	4	4	16	8
0	10	10	34	0

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}} = \frac{0 - \frac{0 \times 10}{5}}{\sqrt{\left\{ 10 - \frac{(0)^2}{5} \right\} \left\{ 34 - \frac{(10)^2}{5} \right\}}}$$

$$= \frac{0}{\sqrt{10 \times 14}} = 0$$

Conclusion. Here the value of r is zero. But there exists a perfect non-linear relationship between x and y . Actually, the relationship between x and y is $y = x^2$. Hence simple correlation coefficient cannot measure the strength of non-linear relationship between x and y .

We can measure the non-linear relationship between x and y by correlation ratio, which is, beyond the scope of this book.

11.7. Probable Error of Correlation Coefficient

If r is the correlation coefficient in a sample of n pairs of observations, then its standard error is given by

$$S.E (r) = \frac{1-r^2}{\sqrt{n}}$$

The probable error (P.E.) of the coefficient of correlation helps in interpreting its value.

Probable of error (P.E.) of coefficient of correlation is defined as

$$P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

where r is the sample correlation coefficient and n is the number of pairs of observations.

The probable error (P.E.) of the coefficient of correlation helps in interpreting its value.

Probable error is an old measure for testing the reliability of an observed correlation coefficient. The reason for taking the factor 0.6745 is that in a normal distribution, the range $\mu \pm 0.6745\sigma$ covers 50% of the total area.

Conclusion

1. If the value of r is less than probable error, there is no evidence of correlation between the variables i.e., the value of r is not at all significant.
2. If the value of r is more than six times the value of probable error, the existence of correlation is practically certain, i.e., the value of r is significant.
3. The population correlation coefficient ρ is expected to lie in the interval $r \pm P.E.r$

Remarks. Now a day t-test is used to test the significance of an observed correlation coefficient will be discussed in chapter 16.

Example 11.7.1. The following data relate to advertising expenditure (in lakhs of taka) and sales (in crores of taka) of a firm:

Advertising expenditure (in lakhs of Tk.)	10	12	15	20	23
Sales (in crores of Tk.)	14	17	23	21	25

Compute the coefficient correlation between advertising expenditures and sales and comment on the value of r .

Solution. Table for calculation of correlation coefficient

Advertising expenditure	Sale : y	x^2	y^2	xy
10	14	100	196	140
12	17	144	289	204
15	23	225	529	345
20	21	400	441	420
23	25	529	525	575
80	100	1396	2080	1684

Coefficient of correlation

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\{n\sum x^2 - (\sum x)^2\}\{n\sum y^2 - (\sum y)^2\}}} = \frac{(5 \times 1684) - (80 \times 100)}{\sqrt{(5 \times 13980) - (80)^2} \{(5 \times 2080) - (100)^2\}}$$

$$= \frac{8420 - 8000}{\sqrt{(6990 - 6400)(10400 - 10000)}} = \frac{420}{\sqrt{(590 \times 400)}} = \frac{420}{\sqrt{23600}} = \frac{420}{485.5} = 0.864$$

Here probable error of r is

$$P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}} = 0.6745 \times \frac{1-(.864)^2}{\sqrt{5}} = \frac{0.254}{2.236} = 0.077.$$

Six times of the $P.E.(r) = 6 \times 0.077 = 0.462$ which is less than the value of correlation coefficient. Hence the value of r is significant. Moreover,

Example 11.7.2. A researcher wants to find out if there is any relationship between the ages of husbands and the ages of wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below:

Age of husbands (x)	25	27	29	32	35	37	39
Ages of wives (y)	18	20	20	25	25	30	37

- (i) Draw a scatter diagram with the above data.
- (ii) Compute coefficient of correlation between the ages of husbands and the ages of wives.

Solution. The computing formula for finding the correlation coefficient is

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}}$$

Let us make a table to find the value of r

X	Y	x^2	y^2	xy
25	18	625	324	450
27	20	729	400	540
29	20	841	400	580
32	25	1024	625	800
35	25	1225	625	875
37	30	1369	900	1110
39	37	1521	1369	1443
$\Sigma x = 224$	$\Sigma y = 175$	$\Sigma x^2 = 7334$	$\Sigma y^2 = 4643$	$\Sigma xy = 5798$

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{ \Sigma x^2 - \frac{(\Sigma x)^2}{n} \right\} \left\{ \Sigma y^2 - \frac{(\Sigma y)^2}{n} \right\}}}$$

$$= \frac{5798 - \frac{224 \times 175}{7}}{\sqrt{7334 - \frac{(224)^2}{7}}} = \frac{5798 - 5600}{\sqrt{166 \times 268}} = \frac{198}{210.92} = 0.94.$$

Conclusion. Here probable error of r is

$$P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}} = 0.6745 \times \frac{1-(0.94)^2}{\sqrt{7}} = \frac{0.1164}{2.646} = 0.044$$

Six times of the $P.E.(r) = 6 \times 0.044 = 0.2639$ which is less than the value of correlation coefficient. Hence the value of r is significant.

Example 11.7.3. The following table gives the prices and consumption of salts of a family for the last 5 months

Price per kg (in taka)	5	6	7	8	9
Consumption in kg	4	4	4	4	4

Find covariance between the prices and consumption of salt and comment.

Solution.

Price : x	Consumption : y	xy
5	4	20
6	4	24
7	4	28
8	4	32
9	4	36
35	20	140

The computing formula for finding covariance is

$$\text{Covariance} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n} = \frac{1}{n} \left\{ \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} \right\}$$

$$= \frac{1}{5} \left\{ 140 - \frac{35 \times 20}{5} \right\} = \frac{1}{5} (140 - 140) = 0$$

Comment. Here prices of the salt and consumptions of salt are independent since the covariance between them is zero. Correlation coefficient between them will also be zero.

When the mean of the variables are whole number, then it is quite easy to find the value of the correlation coefficient by the original formula. Now we shall cite some examples.

Calculation of correlation coefficient using the original formula.

The sample correlation coefficient is defined by

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

where $X = (x - \bar{x})$ and $Y = (y - \bar{y})$

This formula is very useful when the values of x and y are large and the means of x and y are whole number.

(i) When the means of x and y are both integer or whole number

Example 11.7.4. The following data refer to the sales and expenses of 10 firms in Lakh taka for 2011:

Firm	1	2	3	4	5	6	7	8	9	10
Sales (in Lakh Tk.)	65	65	65	60	60	50	60	55	50	50
Expenses (in Lakh Tk.)	16	15	15	14	13	13	16	14	13	11

Compute correlation coefficient and comment.

Solution. $\Sigma x = 580, \Sigma y = 140; n = 10$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{580}{10} = 58 ; \bar{y} = \frac{\Sigma y}{n} = \frac{140}{10} = 14$$

Here both the mean of x and y are integer. So we can comfortably use the above formula for finding correlation coefficient.

Table for computation of correlation coefficient

Firm	Sales x	$X = (x - \bar{x})$ $= x - 58$	X^2	Expenses y	$Y = (y - \bar{y})$ $= y - 14$	Y^2	XY
1	65	7	49	16	2	4	14
2	65	7	49	15	1	1	7
3	65	7	49	15	1	1	7
4	60	2	4	14	0	0	0
5	60	2	4	13	-1	1	-2
6	50	-8	64	13	-1	1	8
7	60	2	4	16	2	4	4
8	55	-3	9	14	0	0	0
9	50	-8	64	13	-1	1	8
10	50	-8	64	11	-3	9	24
	$\Sigma x = 580$	$\Sigma X = 0$	$\Sigma X^2 = 360$	$\Sigma y = 140$	$\Sigma Y = 0$	$\Sigma Y^2 = 22$	$\Sigma XY = 70$

Correlation coefficient,

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X \Sigma Y}} = \frac{70}{\sqrt{360 \times 22}} = \frac{70}{88.994} = 0.787.$$

There is a high degree of positive correlation between the variables. That is, as the sales go up expenses also go up.

(i) When the means of x and y are not integer

When the values of x and y are not so large and the actual means are in fractions, say the actual means of x and y are 24.23 and 12.57 respectively, the calculation of coefficient of correlation by the original formula would involve too many calculations and would take a lot of time. In such cases we can only shift the origins of x and y to some constants A and B which are very near to the original means of x and y to get the maximum benefit of calculations. The formula for finding correlation coefficient is

$$r = \frac{\Sigma uv - \frac{(\Sigma u)(\Sigma v)}{n}}{\sqrt{\left\{ \sum u^2 - \frac{(\Sigma u)^2}{n} \right\} \left\{ \sum v^2 - \frac{(\Sigma v)^2}{n} \right\}}}.$$

$$u = x - A; \quad v = y - B$$

Here A is the assumed mean of x which is usually taken as a whole number and very near to the actual mean of x. Similarly, B is the assumed mean of y which is usually taken as a whole number and very near to the actual mean of y. Let us take consider the following example.

Example 11.7.5. Calculate Karl Pearson's coefficient of correlation for the following data and interpret its value:

x	112	116	103	116	98	118	112	104	111	105
y	65	69	60	68	56	72	60	53	64	62

Solution. Here $\Sigma x = 1095$, $\Sigma y = 629$ and $n = 10$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1095}{10} = 109.5; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{629}{10} = 62.9.$$

Here the means of x and y are in fractions. We can take 105 and 60 as assumed means for x and y and construct the following table.

X	$u = x - 105$	u^2	y	$v = y - 60$	v^2	uv
112	7	49	65	5	25	35
116	11	121	69	9	81	99
103	-2	4	60	0	0	0
116	11	121	68	8	64	88
98	-7	49	56	-4	16	28
118	13	169	72	12	144	156
112	7	49	60	0	0	0
104	-1	1	53	-7	49	7
111	6	36	64	4	16	24
105	0	0	62	2	4	0
$\Sigma x = 1095$	$\Sigma u = 45$	$\Sigma u^2 = 599$	$\Sigma y = 629$	$\Sigma v = 29$	$\Sigma v^2 = 399$	$\Sigma uv = 437$

$$\begin{aligned}
 r &= \frac{\Sigma uv - \frac{(\Sigma u)(\Sigma v)}{n}}{\sqrt{\left\{ \Sigma u^2 - \frac{(\Sigma u)^2}{n} \right\} \left\{ \Sigma v^2 - \frac{(\Sigma v)^2}{n} \right\}}} \\
 &= \frac{437 - \frac{(45)(29)}{10}}{\sqrt{\left\{ 599 - \frac{(45)^2}{10} \right\} \left\{ 399 - \frac{(29)^2}{10} \right\}}} s \\
 &= \frac{437 - 130.5}{\sqrt{(599 - 202.5)(399 - 84.1)}} = \frac{306.5}{353.35} = 0.867
 \end{aligned}$$

Example 11.7.6. The following data give the advertising expenditure and sales of a firm for last ten months:

Month	Expenditure : x	Sales : y	Month	Expenditure : x	Sales : y
Jan.	50	1,600	June	150	2600
Feb.	60	2000	July	140	2800
March	70	2200	Aug	160	2900
April	90	2500	Sept.	170	3100
May	120	2400	Oct.	190	3900

Compute the correlation coefficient.

Solution. Here, $\bar{x} = \frac{1200}{10} = 120$ and $\bar{y} = \frac{26000}{10} = 2600$.

Both the means of x and y are integers. Moreover, the values of x are some multiple of 10 and the values of y are some multiple of 100. So we can define the new variables u and v as

$$u = \frac{x - 120}{10}; \quad v = \frac{y - 2600}{100}$$

Here A = 120, B = 2600, h = 10 and k = 100.

Computation of correlation Coefficient

x	$u = \frac{x - 120}{10}$	u^2	y	$v = \frac{y - 2600}{100}$	v^2	uv
50	-7	49	1600	-10	100	70
60	-6	36	2000	-6	36	36
70	-5	25	2200	-4	16	20
90	-3	9	2500	-1	1	3
120	0	0	2400	-2	4	0
150	3	9	2600	0	0	0
140	2	4	2800	2	4	4
160	4	16	2900	3	9	12
170	5	25	3100	5	25	25
190	7	49	3900	13	169	91
$\Sigma x = 1200$	$\Sigma u = 0$	$\Sigma u^2 = 222$	$\Sigma Y = 26,000$	$\Sigma v = 0$	$\Sigma v^2 = 364$	$\Sigma uv = 261$

$$r = \frac{\Sigma uv}{\sqrt{u^2 \times v^2}} = \frac{261}{\sqrt{222 \times 364}} = \frac{261}{284.27} = 0.918.$$

There is a very high degree of positive correlation coefficient between advertising expenditure and sales.

11.8. Rank Correlation

The British Psychologist Edward Spearman developed Spearman's rank correlation coefficient in 1904. This method is applied in a situation in which quantitative measure of certain qualitative factors such as beauty, intelligent, judgment, leadership, colour, taste cannot be fixed but individual observations can be arranged in a definite order called rank.

Definition. Suppose 1, 2, ..., n are assigned to the x observations in order of magnitude and similarly to the y observations. Then the simple correlation coefficient between the two sets of ranks is called Spearman's rank correlation coefficient. It is denoted by R..

When there are no ties among either set of observations, then the formula for computing the Spearman's rank correlation coefficient

$$R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Here, R = rank correlation coefficients

R_1 = Rank of observations with respect to first variable x

R_2 = Rank of observations with respect to second variable y

$d = R_1 - R_2$, difference in a pair of ranks

n = number of pairs of observations being ranked.

Actually, Spearman's rank correlation coefficient is a nonparametric counterpart of the Karl Pearson's simple correlation coefficient r .

If there are no ties among either set of observations, the value of R will usually be very close to the value of r based on numerically observations and is interpreted in much the same way. The value of R will range from -1 to $+1$. A value of $+1$ or -1 indicates perfect association between x and y , the plus sign occurring for identical rankings and the minus sign occurring reverse rankings. When R is close to zero, we would conclude that the variables are uncorrelated.

Remarks. We always have

$$\sum d_i = \sum (R_1 - R_2) = \sum R_1 - \sum R_2 = 0$$

This serves as a check on the calculation.

Properties of rank correlation coefficient:

1. Like simple correlation coefficient, rank correlation coefficient lies between -1 to $+1$.
2. $R = 1$, when the ranks of x completely agree with the ranks of y , i.e. $(R_1, R_2) = (1, 1), (2, 2), \dots, (n, n)$.
3. $R = -1$; when there is complete disagreement in the ranks, in this case $(R_1, R_2) = (1, n), (2, n-1), \dots, (n, 1)$.
4. This is the only method for finding relationship between two qualitative variables like beauty, honesty, intelligence efficiency and so on.
5. This is the only method for finding relationship between two variables when ranks are given.

Limitations. This method cannot be used for finding correlation in a grouped frequency distribution.

For finding rank correlation coefficient, we may have two types of data:

- (i) Actual observations are given
- (ii) Actual ranks are given

Now we shall cite some examples to show how the different values of the R change.

Example 11.8.1. Suppose that two managers I and II are to rank 5 employees A, B, C, D and E on the basis of their performance in a test and the results are recorded as follows:

	Examiner/ Employees	A	B	C	D	E
(a)	I	1	2	3	4	5
	II	1	2	3	4	5
(b)	Examiner/ Employees	A	B	C	D	E
	I	1	2	3	4	5
	II	2	4	1	5	4
(c)	Examiner/ Employees	A	B	C	D	E
	I	1	2	3	4	5
	II	5	4	3	2	1
(d)	Examiner/ Employees	B	C	D	A	E
	I	2	3	4	1	5
	II	3	4	1	5	2
(e)	Examiner/ Employees	B	C	D	A	E
	I	2	3	4	1	5
	II	3	2	1	4	5

Solution. (a) The formula for computing rank correlation coefficient is

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Ranking by manager I : R ₁	Ranking by manager II : R ₂	d ² = (R ₁ - R ₂) ²
1	1	0
2	2	0
3	3	0
4	4	0
5	5	0
		$\Sigma d^2 = 0$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{0}{5 \times 24} = 1.$$

Conclusion. There is a positive perfect ranking between the managers. That is, there is a full agreement between the two managers regarding the ranking of the employees.

(b) The formula for computing rank correlation coefficient is

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Ranking by manager I : R ₁	Ranking by manager II : R ₂	d ² = (R ₁ - R ₂) ²
1	2	1
2	3	1
3	1	4
4	5	1
5	4	1
		$\Sigma d^2 = 8$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 8}{5 \times 24} = 1 - \frac{48}{120} = 1 - 0.4 = 0.6.$$

Conclusion. There is a positive rank correlation coefficient between the rankings of two managers.

(c) The formula for computing rank correlation coefficient is

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Ranking by manager I : R ₁	Ranking by manager II : R ₂	d ² = (R ₁ - R ₂) ²
1	5	16
2	4	4
3	3	0
4	2	4
5	1	16
		$\sum d^2 = 40$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 40}{5 \times 24} = 1 - \frac{240}{120} = 1 - 2 = -1.$$

Conclusion. There is a perfect negative relationship between the rankings of the two managers. That is, there is a full disagreement between rankings of the two managers.

(d) The formula for computing rank correlation coefficient is

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Ranking by manager I : R ₁	Ranking by manager II : R ₂	d ² = (R ₁ - R ₂) ²
1	5	16
2	3	1
3	4	1
4	1	9
5	2	9
		$\sum d^2 = 36$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 36}{5 \times 24} = 1 - \frac{216}{120} = 1 - 1.8 = -0.8.$$

Conclusion. There is a negative strong rank correlation coefficient between the rankings of two managers.

(e) The formula for computing rank correlation coefficient is

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Ranking by manager I : R ₁	Ranking by manager II : R ₂	D ² =(R ₁ -R ₂) ²
2	3	1
3	2	1
4	1	9
1	4	9
5	5	0
		$\sum d^2 = 20$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 20}{5 \times 24} = 1 - \frac{120}{120} = 1 - 1 = 0.$$

Comment. In this case, the ranking of the two managers are independent.

When ranks are not given. When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of all the variables. Now we shall cite some examples.

Example 11.8.2. A Social Scientist wants to see whether there is any association between the intelligence and beauty among the female students. To verify this he randomly selected 6 female students from a class. The scores on intelligence and beauty are found as follows:

Student	A	B	C	D	E	F
Scores on intelligence	80	75	90	70	65	60
Scores on beauty	65	70	60	75	85	80

Compute rank correlation coefficient and comment

Solution. The formula for computing R is $R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$

Here ranks of the scores are not given. Let us start ranking from the highest value for both the variables as shown in the table given below:

Student	Scores on intelligence : x	Scores on beauty : y	Ranks on intelligence : R ₁	Ranks on beauty : R ₂	d ² = (R ₁ -R ₂) ²
A	80	65	2	5	9
B	75	70	3	4	1
C	90	60	1	6	25
D	70	75	4	3	1
E	65	85	5	1	16
F	60	80	6	2	16
Total					$\sum d^2 = 68$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 68}{6 \times 35} = 1 - \frac{68}{35} = 1 - 1.94 = -0.94.$$

Conclusion. There exists almost perfect negative relationship between the intelligence and beauty.

Example 11.8.3. Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers. The scores of the two managers are as follows:

Employee	A	B	C	D	E
Scores by Manager I	80	72	70	75	65
Scores by Manager II	75	69	71	78	67

Solution. The formula for computing R is $R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$

Here ranks of the scores are not given. Let us start ranking from the highest value for both the variables as shown in the table given below:

Employees	Scores on manager I : x	Scores on Manager II : Y	Ranks by manager I : R ₁	Ranks by manager II : R ₂	d ² = (R ₁ - R ₂) ²
A	80	75	1	2	1
B	72	69	3	4	1
C	70	71	4	3	1
D	75	78	2	1	1
E	65	67	5	5	0
Total					$\Sigma d^2 = 4$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 4}{5 \times 24} = 1 - \frac{1}{5} = 1 - 0.2 = 0.8$$

Conclusion. There exists strong positive relationship between the rankings of the two managers.

Example 11.8.4. Two housewives, Rahima and Maksuda were asked to express their preference for different kinds of detergents, gave the following replies:

Detergent	A	B	C	D	E	F	G	H	I	J
Rahima	10	9	5	6	8	7	3	1	2	4
Maksuda	10	9	6	5	7	8	3	2	1	4

Compute rank correlation and compare how far their preferences go together.

Solution. Here we will compute rank correlation coefficient.

Table for computing rank correlation coefficient.

Detergent	R_1	R_2	d^2
A	10	10	0
B	9	9	0
C	5	6	1
D	6	5	1
E	8	7	1
F	7	8	1
G	3	3	0
H	1	2	1
I	2	1	1
J	4	4	0
$n = 10$			$\Sigma d^2 = 6$

Rank correlation coefficient,

$$R = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 6}{10 \times 99} = 1 - \frac{1}{990} = 1 - 0.036 = 0.964.$$

Thus the preferences of these two ladies agree very closely as far as their opinion on detergents is concerned.

Remarks. Normal test is used to test the significance of an observed rank correlation coefficient will be discussed later in test of significance.

11.8.1. Repeated ranks or ties observations. If there is more than one observations either x or y are same, then the Spearman's formula for calculating the rank correlation coefficient breaks down. In this case, common ranks are given to the repeated observations. This common rank is the average of the ranks, which these observations would have assumed, and the observation will get the rank next to the ranks already assumed. As results a correction term is added in the rank correlation formula. In the formula, we add the factor $\frac{m(m^2 - 1)}{12}$ to $\sum d^2$, where m is the number of times an observation is repeated. This correction factor is to be added for each repeated value.

Now, we shall cite an example.

When ranks are repeated the following formula is used for finding rank correlation coefficient:

$$R = 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} (m_1^2 - m_1) + \frac{1}{12} (m_2^2 - m_2) + \dots \right\}}{n(n^2 - 1)}$$

$$R = 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} \sum (m_i^2 - m_i) \right\}}{n(n^2 - 1)}$$

Where m_i refers to the number of times i th value repeated.

Example 11.8.5. The following data refer to the marks obtained by 8 students in mathematics and statistics:

Marks in Mathematics	20	80	40	12	28	20	15	60
Marks in Statistics	30	60	20	30	50	30	40	20

Compute rank correlation coefficient and comment.

Solution. Let the marks obtained by mathematics be X and the marks obtained by Statistics be Y . Here let us start ranking from the lowest values for both the variables.

Table for computation of rank correlation.

X	R _x	Y	R _y	d ²
20	3.5	30	4	0.25
80	8	60	8	0.00
40	6	20	2	16.00
12	1	30	4	9.00
28	5	50	7	4.00
20	3.5	30	4	0.25
15	2	40	6	16.00
60	1	10	1	36.00

Here $\sum d^2 = 81.5$

In series X , 20 have come two times and in series Y , 30 have come three times. The necessary adjustment for this repeated rank has to be made. Hence adjusted formula for rank correlation coefficient is

$$R = 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} (m_1^2 - m_1) + \frac{1}{12} (m_2^2 - m_2) + \dots \right\}}{n(n^2 - 1)}$$

Here $\sum d^2 = 81.5, m_1 = 2, m_2 = 3, n = 8$

$$R = 1 - \frac{6 \left\{ 81.5 + \frac{1}{12} (2^2 - 2) + \frac{1}{12} (3^2 - 3) \right\}}{8(8^2 - 1)} = 1 - \frac{6 \times 84}{504} = 1 - \frac{504}{504} = 0.$$

Example 11.8.6. The data refer to the marks obtained by a student in ten subjects by two examiners:

Examiner I : 68 64 75 50 64 80 75 40 55 64
 Examiner II : 62 58 68 45 81 60 68 48 50 70

Find rank correlation coefficient.

Solution. Suppose the marks by examiner I is X and the examiner II is Y. Let us start ranking from the highest value for both the variables.

Table for computing rank correlation coefficient R

Marks of Examiner I : x	Marks of Examiner II : y	Ranks of marks of Examiner I : R ₁	Ranks of marks of Examiner II : R ₂	d=R ₁ - R ₂	D ²
68	62	4	5	-1	1
64	54	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	$\Sigma d^2 = 72$

In the observations for x we see that the value 75 occurs 2 times. The common rank given to these values is 2.5, which is the average of 2 and 3, the ranks that these values would have taken if they were different. The next value 68 then gets the next rank, which are 4. Again we see that value 64 occurs thrice. The common rank given to it is 6, which is the average of 5, 6 and 7. Similarly, in the observations for y we see that the value 68 occurs twice and its common rank is 3.5, which is the average of 3 and 4. As a result of these common rankings, the formula for R has to be corrected. To

Σd^2 we add $\frac{m(m^2 - 1)}{12}$ for each value repeated, where m is the number of

times a value occurs. Correction for x is to be applied twice, once for the value 75 that occurs twice ($m_1=2$) and then for the value 64, which occurs thrice ($m_2 = 3$). The total correction for x is

$$\frac{2(2^2 - 1)}{12} + \frac{3(3^2 - 1)}{12} = \frac{6}{12} + \frac{24}{12} = \frac{5}{2} = 2.5$$

Similarly, the correction for y is $\frac{2(2^2 - 1)}{12} = \frac{1}{2} = 0.5$

$$\text{Thus, } R = 1 - \frac{6[\Sigma d^2 + 2.5 + 0.5]}{n(n^2 - 1)} = 1 - \frac{6(72 + 3)}{10 \times 99} = 1 - \frac{450}{990} = 1 - 0.455 = 0.545.$$

11.8.2. Advantages of rank correlation coefficient over simple correlation coefficient.

1. Rank correlation coefficient can be safely used in case of linear and curvilinear relationship between two variables x and y . But simple correlation coefficient measures only the strength of linear relationship between the variables.
2. No assumption of normality is required for testing the significance of R whereas the assumption of normality is required to test the significance of the sample correlation coefficient.
3. Rank correlation coefficient can be used for finding the association between two qualitative as well as two quantitative variables, whereas simple correlation coefficient is used for finding linear relationship between two quantitative variables only.
4. Rank correlation coefficient is easy to understand and apply as compared with simple correlation coefficient.
5. If there are no ties among the either set of observations or no many ties exist, the rank correlation coefficient is slightly over than the simple correlation coefficient.

Example 11.8.7. The scores of eight students in Statistics and English in an examination are given below:

Students Number	1	2	3	4	5	6	7	8
Marks in Statistics	52	60	50	54	55	58	48	70
Marks in English	48	51	68	55	60	53	47	62

Compute

- (i) Simple correlation coefficient and
- (ii) Rank correlation coefficient

Solution. (i) Let the marks in statistics be denoted by x and English by y .

Here $\Sigma x = 447$, $\Sigma y = 444$ and $n=8$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{447}{8} = 55.88 \text{ and } \bar{y} = \frac{\Sigma y}{n} = \frac{444}{8} = 55.5$$

Here both the means of x and y are fractions. Let us take the assumed mean for x as 55 and for y as 55 which are very close to their actual means.

Computation table for correlation coefficient

X	$d_x = x - 55$	d_x^2	y	$d_y = y - 55$	d_y^2	$d_x d_y$
52	-3	9	48	-7	49	21
60	5	25	51	-4	16	-20
50	-5	25	68	13	169	-65
54	-1	1	55	0	0	0
55	0	0	60	5	25	0
58	3	9	53	-2	4	-6
48	-7	49	47	-8	64	56
70	15	225	62	7	49	105
$\Sigma x = 447$	$\Sigma d_x = 7$	$\Sigma d_x^2 = 343$	$\Sigma y = 444$	$\Sigma d_y = 4$	$\Sigma d_y^2 = 376$	$\Sigma d_x d_y = 91$

$$\begin{aligned}
 r &= \frac{\Sigma d_x d_y - \frac{(\Sigma d_x \times \Sigma d_y)}{n}}{\sqrt{\left\{ \Sigma d_x^2 - \frac{(\Sigma d_x)^2}{n} \right\} \left\{ \Sigma d_y^2 - \frac{(\Sigma d_y)^2}{n} \right\}}} \\
 &= \frac{91 - \frac{7 \times 4}{8}}{\sqrt{\left\{ 343 - \frac{(7)^2}{8} \right\} \left\{ 376 - \frac{(4)^2}{8} \right\}}} \\
 &= \frac{91 - 3.5}{\sqrt{(343 - 6.125)(376 - 2)}} = \frac{87.5}{\sqrt{336.875 \times 374}} = \frac{87.5}{354.95} = 0.25
 \end{aligned}$$

(ii) Calculation of rank correlation coefficient

X	R_x	y	R_y	d^2
52	6	48	7	1
60	2	51	6	16
50	7	68	1	36
54	5	55	4	1
55	4	60	3	1
58	3	53	5	4
48	8	47	8	0
70	1	62	2	1
				$\Sigma d^2 = 60$

$$R = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{8 \times 63} = 1 - \frac{360}{504} = 1 - 0.714 = 0.29$$

Comment. Here it is seen that rank correlation coefficient is slightly more than the simple correlation coefficient.

Limitations. This method cannot be used for finding out correlation in a bivariate frequency distribution. Usually for $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

Example 11.8.8. Two judges have ranked 12 students in order of their merits as follows:

Students	A	B	C	D	E	F	G	H	I	J	K	L
Rank by 1st Judge	5	2	4	1	8	9	10	6	3	11	7	12
Rank by 2nd judge	6	9	7	10	1	2	4	12	3	5	11	8

Calculate rank correlation coefficient to find out whether the judges are in agreement with each other or not.

Solution.

Student	R ₁	R ₂	d ²
A	5	6	1
B	2	9	49
C	4	7	9
D	1	10	81
E	8	3	25
F	9	2	49
G	10	4	36
H	6	12	36
I	3	3	0
J	11	5	36
K	7	11	16
L	12	8	16
			$\Sigma d^2 = 354$

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 354}{12(12^2 - 1)} = 1 - \frac{2124}{1716} = 1 - 1.237 = -0.237.$$

There is a negative rank correlation coefficient that indicates that the judges are not in agreement with each other.

Example 11.8.9. Two bank officers examined eleven loan applications and ranked them

Applicants	A	B	C	D	E	F	G	H	I	J	K
Officer I	1	7	4	2	3	6	5	9	10	8	11
Officer II	1	6	5	2	3	4	7	11	8	10	9

Compute rank correlation coefficient and comment.

Solution.

Applicants	R ₁	R ₂	d ²
A	1	1	0
B	7	6	1
C	4	5	1
D	2	2	0
E	3	3	0
F	6	4	4
G	5	7	4
H	9	11	4
I	10	8	4
J	8	10	4
K	11	9	4

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 26}{11(11^2 - 1)} = 1 - \frac{156}{1320} = 1 - 0.118 = 0.882.$$

Comment. There is close agreement between the officers since the value of the rank correlation coefficient is 0.88.

Questions

1. What is a scatter diagram? Interpret the different values of r with the help of scatter diagram.
2. Define Karl Pearson's correlation coefficient. State some of its important properties. Cite some practical examples.
3. What is covariance? State some of its important properties. What are advantages of correlation coefficient over covariance?
4. Define Spearman's rank correlation coefficient. State some of its important properties. Cite some of its advantages over simple correlation coefficient. What are the limitations of rank correlation coefficient?
5. What is correlation? Distinguish between positive and negative correlation with the help of scatter diagram. Cite some examples of positive and negative correlation.
6. Define correlation coefficient. Explain the value of r when it takes values $-1, 0, +1$.
7. What is probable error of r ? How can you explain the value of r with the help of probable error?
8. What is rank correlation coefficient? Explain the value of R when it takes values -1 and $+1$ with the help of examples.

Exercises

9. Find the simple correlation coefficients of the following sets of bivariate data:

(a)	X : x	1	2	3	4	5
	Y : y	4	6	8	10	12
(b)	X : x	1	2	3	4	
	Y : y	7	5	3	1	
(c)	X : x	1	2	3	4	
	Y : y	1	3	7	8	
(d)	X : x	2	3	4	5	6
	Y : y	9	5	6	2	1
(e)	X : x	1	2	3	4	5
	Y : y	3	2	2	2	3
(f)	X : x	-3	-2	-1	0	1
	Y : y	4	1	0	1	4

Ans. (a) $r = 1$; (b) $r = -1$; (c) $r = 0.976$; (d) $-r = 0.936$; (e) $r = 0$; (f) $r = 0$

10. Find the correlation between x and y from the following information:

$$n = 6, \Sigma x = 68; \Sigma y = 112; \Sigma xy = 1292; \Sigma x^2; \Sigma y^2.$$

Also find the value of the coefficient of determination and comment.

Ans. $r = 0.947$; $r^2 = 0.90$; we can say that 90% of the variation in the values of X is accounted for by a linear relationship with X.

Applications

11. The following table gives the ages and blood pressure of 10 women:

Age in years	x :	56	42	36	47	49	42	72	63	55	60
Blood Pressure	y :	147	125	118	128	145	140	155	160	149	150

- (i) Draw a scatter diagram.
(ii) Find correlation coefficient between x and y and comment.

12. The following data refer to the advertising expenditures and sales of a firm for the period 1995-2002:

Year	1995	1996	1997	1998	1999	2000	2001	2002
Advertising Expenditure :	12	15	18	23	24	38	42	48
('000Tk.)								
Sales (Lakh TK.):	5.0	5.6	5.8	7.0	7.2	8.8	9.2	9.5

Calculate Karl Pearson's coefficient of correlation between advertising expenditure and sales.

13. Find the coefficient of correlation between price and sales from the following data and interpret its value through probable error.

Price (Taka) : 103 98 85 92 90 84 88 90 94 95

Sales (Kg) : 500 610 700 630 670 800 800 570 700 680

Ans. $r = -0.856$ and $P.E. = 0.057$.

14. Ten competitors in a beauty contest are ranked by two judges in the following order:

Judge I : 1 8 6 7 5 9 10 4 3 2

Judge II : 6 7 4 5 9 10 8 3 1 2

Find the rank correlation coefficient and comment on the judgment.

15. Find the coefficient of correlation between the sale and expenses of the following ten firms:

Firm	1	2	3	4	5	6	7	8	9	10
Sales	50	52	55	60	65	65	63	67	68	70
Expenses	8	7	9	9	11	13	10	13	12	14

16. Two teachers have ranked 8 students in order of merits as under:

Students	A	B	C	D	E	F	G	H
Ranked by 1st Teacher	1	3	6	7	4	5	8	2
Ranked by 2nd Teacher	2	4	5	8	3	6	7	1

- Calculate rank correlation coefficient to find out the agreement of their assessment.
17. A company gives on the job training to its salesmen which are followed by a test. It is considered whether it should terminate the services of any salesman who does not do well in the test.

Test Scores	14	19	24	21	26	22	15	20	10
Sales in thousand taka	31	36	48	37	50	45	33	41	39

Compute the correlation coefficient between test scores and sales. Does it indicate that the termination of the service of salesman with low test scores is justified? [Hints. Here $\bar{x} = 20$ and $\bar{y} = 40$, one can use the original formula to find the value of r . Ans. $r = .947$, not justified since the value of r is very high.]

18. Following figures give the rainfall in inches for 7 districts and the production of Rabi crop in hundred kgs per acre of a particular year.

Rainfall : 20.22 24 26 28 30 32

Rabi production: 30 35 40 50 60 60 55

Compute coefficient of correlation between the rainfall and production and comment.

Ans. $R=0.917$, A very high degree of positive correlation between rainfall and agricultural production.

19. Find the coefficient of correlation between the price and sales of a commodity for last ten years:

Price in taka : 95 93 90 88 84 90 92 85 98 103
 Sales per unit: 68 70 75 80 80 67 63 70 61 50

Ans. 0.85

20. The following data relate to the prices and supplies of commodities during a periods of eight years:

Price per kg(in taka) : 17 18 19 15 16 18 12 10
 Supply (100kg) : 46 47 48 42 44 45 35 30

Compute correlation coefficient between the two series. Ans. 0.98.

21. The following data refer to the sales and purchase per unit of a commodity in different period of times:

Sales	51	57	67	73	91	97	108	111	121
Purchase	39	47	70	61	71	75	69	80	97

Compute correlation coefficient between the sales and purchase.

22. The marks obtained by 8 students in mathematics and statistics in an examination are as follows:

Student No.	1	2	3	4	5	6	7	8
Marks in Mathematics	72	80	70	74	75	78	68	90
Marks in Statistics	68	71	88	75	80	73	67	82

Compute

- (i) Correlation coefficient, and
 (ii) Rank correlation coefficient and, compare the two values.

CHAPTER - 12

SIMPLE REGRESSION ANALYSIS

12.1. Introduction

Correlation coefficient measures the strength of linear relationship between two variables. It does not measure the cause and effect relationship between two variables. By regression analysis we can measure the cause and effect relationship between two variables.

The concept of regression was first introduced by a British biometrician, Sir Francis Galton in 1877 while studying the relationship between the heights of fathers and sons. He found that tall fathers tend to have tall sons and short fathers short sons; but the average height of the sons of a group of tall fathers is less than that of the tall fathers and the average height of the sons of a group of short fathers is greater than that of the short fathers. But now it is widely used in statistics.

Today regression analysis is a very powerful tool in the field of statistical analysis in predicting the value of one variable on the basis of the given value of another variable, when these two variables are related to each other.

There are many situations in business and other fields where we are interested to measure the relationship between two variables.

Some examples of these related variables are:

- i) Fertilizer used and yield of various plots of land
- ii) Ages of husbands and ages of wives of a group of couples
- iii) Income and expenditure of a class of people.
- iv) The price of a commodity and amount demanded.
- v) The advertising expenditure and the volume of sales of a product.
- vi) The volume of sales and the experience of the salesman of a departmental store.
- vii) The deposit in a bank and the number of clients.
- viii) The performance of a student in a high school and the performance of the same student in the college.
- ix) The heights and weights of students in a class etc.

In all the above examples, there are two variables involved, the value of one variable depending upon the value of the other variable. For example, within the limits, the yield of a plot depends upon the kind and amount of fertilizer used. Hence the yield variable y is known as the dependent

variable and the fertilizer variable x is known as independent variable. The expenditure y of a person would depend upon the income x of the person. Similarly, expenditure of a family depends on his income, demand of a commodity depends on price, sale of a store depends on its experience of its salesman etc. Since the value of y is related to the value of x , knowing this relationship would help us predict the value of y for a given value of x . This general process of predicting the value of dependent variable y on the basis of known value of the independent variable x is known as the regression analysis.

12.1.1. Dependent and independent variables. In regression analysis there are two types of variables. They are known as (i) dependent variable and (ii) independent variable.

(i) **Dependent variable** : The variable whose value is influenced or is to be predicted is called dependent variable. It is usually denoted by y . Dependent variable is also known as explained variable, predictant, regressand, response or endogenous variable. In example (i) yield per plot y is called dependent variable.

(ii) **Independent variable** : The variable, which influences the values or is to be used for prediction, is called independent variable. It is usually denoted by x . In regression analysis independent variable is also known as explanatory variable, predictor, regressor, control or exogenous variable. In example (i) amount of fertilizer x per plot is known as independent variable.

12.1.2. Regression Analysis. Regression analysis is a statistical technique, which has developed to study and measure the statistical relationship among two or more variables with a vision to estimate or predict the value of dependent variable for some known value of the independent variable.

In this section, we shall study the relationship between two variables only. The process is called **Simple Regression**.

12.1.3. The purpose of regression analysis. The main purpose of simple regression analysis is to

- (i) Establish a functional or mathematical relationship between dependent and independent variables;
- (ii) Estimate or predict the value of the dependent variable for given value of the independent variable;
- (iii) Show the pattern or trend of the dependent variable for various value of the independent variable.

Broadly speaking, the mathematical or functional relationship between the dependent and independent variable may be (i) linear or (ii) non-linear.

Examples of linear relation:

- (i) $y = a + bx$; (ii) $y = a - bx$; (iii) $y = x$ etc.

Examples of non-linear relation:

- (i) $y = x^2$; (ii) $y = \sqrt{x}$; (iii) $y = a + bx + cx^2$ etc.

In simple regression analysis, we consider the linear or straight-line relationship between the variables. In simple linear regression, a mathematical regression equation is developed to describe the functional relationship that exists between the two variable x and y , and this association is exhibited by plotting the values of paired coordinated (XY) on a graph, with dependent variable y along the vertical Y-axis and the independent variable x along the horizontal X-axis.

12.2. Population Regression Line and Model

Suppose we have N individuals in a population. Suppose we want to study two characteristics say income and expenditure of the individuals by two variables X and Y respectively. If the expenditure variable Y functionally related with income variable X , then the population simple regression equation is a straight line of Y on X and is defined as

$$Y = \alpha + \beta X \quad (12.1)$$

where α is the intercept and β is the slope of the straight line. In regression analysis β is called the regression coefficient of Y on X .

In practice, expenditure of an individual does not depend only on income. There may be other factors such as number of family members and others. Hence in general we can assume a regression model of Y on X as

$$Y_m = \alpha + \beta X + \varepsilon \quad (12.2)$$

Here α and β have the same meaning as mentioned above. ε is the error or disturbance. Actually it is the distance between the observed value Y_m and the expected value of Y . The error term ε 's are normally and independently distributed with zero mean and constant variance σ^2 . Under the assumption, the mean value of Y_m is Y .

To find the best fit of the population regression model or equation we need N pairs of observations of the two variables on the N individuals of the population. Suppose $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$ are N pairs of observations. We can find the best fit of the population regression line by obtaining the value of α and β in such a way that the error sum of square i.e. $\sum \varepsilon^2$ is minimum. This is done by the method of least squares, which will be discussed, for sample data.

12.3. Sample Regression Equation and Model

In practice, it is not always possible to get population data. Suppose we have a sample of n pairs of observations say $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ from a bi-variate population of interest. The sample regression equation is the best-fitted straight line of y on x is

$$\hat{y} = a + bx \quad (12.3)$$

This sample regression line can be considered as an estimate of the population regression line

$$Y = \alpha + \beta X.$$

Like population regression model sample regression model can be defined as

$$y_m = a + bx + e \quad (12.4)$$

This is the regression model of y on x .

Here a is the intercept and b is the slope of the regression line. In regression analysis b is called the regression coefficient of y on x , giving the change in y per unit change of x . e 's are random error or disturbance term which are normally and independently distributed with zero mean and constant variance s^2 . The simple regression equation is the best fitted straight line in the least-squares sense with the sample data. It is defined as

$$\hat{y} = a + bx \quad (12.5)$$

But the observation y follows the model

$$y = a + bx + e \quad (12.6)$$

Then the error is

$$e = y - \hat{y} = (y - a - bx)$$

and the error sum of squares is

$$L = \sum e^2 = \sum (y - a - bx)^2 \quad (12.7)$$

To find the best fitted regression, we have to find the values of a and b in such a way that error sum of squares is minimum. This is done by the method of least squares developed by Legendre. This can be done by taking partial derivatives of L with respect to both a and b , and equate to zero. That is

$$\frac{\partial L}{\partial a} = -2 \sum (y - a - bx) = 0 \Rightarrow \sum y = na + \sum x \quad (12.8)$$

$$\frac{\partial L}{\partial b} = -2 \sum (y - a - bx)x \Rightarrow \sum xy = a \sum x + b \sum x^2 \quad (12.8)$$

Equations (3.8) and (3.9) are called normal equations for finding a and b. By solving the equations (3.8) and (3.9) we get the values of a and b as

$$a = \bar{y} - b\bar{x} \text{ and } b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Now by putting the calculated values of a and b in (3.3) we get the best fitted sample regression line of y on x as

$$\begin{aligned} \hat{y} &= a + bx = \bar{y} - b\bar{x} + bx \\ \hat{y} - \bar{y} &= b(x - \bar{x}) \end{aligned} \quad (12.10)$$

This regression line is used to estimate or predict the values of y for given values of x. The advantage of using this method is that we do not need the value of the intercept a.

The fitted regression line of y on x can also be obtained with the help of the correlation coefficient between x and y.

The correlation coefficient between x and y is

$$\begin{aligned} r &= \frac{\sum (x - \bar{x})(y - \bar{y})/n}{\sqrt{\sum (x - \bar{x})^2/n} \sqrt{\sum (y - \bar{y})^2/n}} = \frac{\text{Cov}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)}} = \frac{\text{Cov}(x, y)}{\sqrt{s_x^2 \times s_y^2}} = \frac{\text{Cov}(x, y)}{s_x s_y} \\ \Rightarrow \text{Cov}(x, y) &= r s_x s_y \end{aligned} \quad (12.11)$$

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})/n}{\sum (x - \bar{x})^2/n} = \frac{\text{Cov}(x, y)}{s_x^2}$$

$$\Rightarrow \text{Cov}(x, y) = b s_x^2 \quad (12.12)$$

From (12.1) and (12.2), we have

$$b s_x^2 = r s_x s_y$$

$$\text{Hence, } b = \frac{r s_y}{s_x}$$

Hence, the best fitted regression line of y on x can be written as

$$\hat{y} = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x})$$

So the best-fitted regression line of y on x can be written in three different ways as

(i) $\hat{y} = a + bx$; by finding the values of a and b by the least squares method.

(ii) $\hat{y} = \bar{y} + b(x - \bar{x})$; when the values of b , \bar{y} and \bar{x} are available.

(iii) $\hat{y} = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x})$; when the values of r , s_x and s_y are available.

It is to be noted that the formulae for finding a and b are quite easy when the values of x and y are measured from their respective means. In that case

$$a = 0 \text{ and } b = \frac{\sum uv}{\sum u^2}; \text{ where } u = x - \bar{x} \text{ and } v = y - \bar{y}$$

Then the regression equation of v on u is $\hat{v} = bu$ (3.13)

By putting the values of u and v in (3.13) we get the required regression equation of y on x as

$$\hat{y} - \bar{y} = b(x - \bar{x}).$$

In many cases the variables x and y are inter-dependent. For example, ages of husbands and ages of wives are inter-dependent. Here one can use any one as dependent variable. Here two regression lines are meaningful. Let us assume that the age of wife x depends on the age of husband y . Then the best fitted regression line of x on y in the least squares sense is

$$\hat{x} = c + dy (12.14)$$

Where c is the intercept and d is the slope of the line. d is also called the regression coefficient of x on y , giving the change in x per unit change of y . e 's are random error which are normally and independently distributed with zero mean and constant variance s^2 . The values of c and d are obtained by the following formulae

$$c = \bar{x} - d\bar{y} \text{ and } d = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2}.$$

By putting the value of c and d in (3.14), we get the best-fitted regression line of x on y as

$$\hat{x} - \bar{x} = d(y - \bar{y}) (12.15)$$

From this regression line, we can estimate or predict the values of x for different given values of y .

Similarly, the best-fitted regression line of x on y can be written in three different ways as

- (i) $\hat{x} = c + dy$; by finding the values of c and d by the least squares method.
- (ii) $\hat{x} = \bar{x} + d(y - \bar{y})$; when the values of d , \bar{y} and \bar{x} are available.
- (iii) $\hat{x} = \bar{x} + r \frac{s_x}{s_y} (y - \bar{y})$; when the values of r , s_x and s_y are available.

Remarks. It is noted that the numerator of r , b and d are same. They have the same sign when they are calculated from the same set of data. That is if b is negative, then d and r are also negative. If r is positive, then b and d are also positive.

12.4. Relationship Between Correlation Coefficient and Regression Coefficients

There is a relationship between the correlation coefficient and the regression coefficients. Actually correlation coefficient is the geometric mean of the regression coefficients.

Theorem 12.1. Correlation coefficient is the geometric mean of the regression coefficients.

Proof. Suppose r is the correlation coefficient between x and y , b is the regression coefficient of y on x and d is the regression coefficient of x on y , then we have to prove

$$r = \sqrt{b \times d}.$$

The coefficient of correlation between x and y

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}}$$

The regression coefficients of y on x and x on y are

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}; \quad d = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

$$\text{Then } b \times d = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \times \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2} = \left[\frac{\{\Sigma(x - \bar{x})(y - \bar{y})\}}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} \right]^2 = r^2$$

$$\text{Hence } r = \sqrt{b \times d}$$

Theorem 12.2. Regression coefficient is independent of the shift of origin but depends on the change of scale.

Proof. Suppose $b_{y/x}$ is the regression coefficient of y on x . That is

$$b_{y/x} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}.$$

Let us define two new variables u and v by shifting the origin and changing the scale as

$$u = \frac{x - A}{h} \Rightarrow x = A + hu$$

$$v = \frac{y - B}{k} \Rightarrow y = B + kv$$

The regression coefficient of v on u is $b_{v/u} = \frac{\Sigma(u - \bar{u})(v - \bar{v})}{\Sigma(u - \bar{u})^2}$

$$\begin{aligned} \text{Now, } b_{y/x} &= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \\ &= \frac{\Sigma(A + hu - A - h\bar{u})(B + kv - B - k\bar{v})}{\Sigma(A + hu - A - h\bar{u})^2} = \frac{k\Sigma(u - \bar{u})(v - \bar{v})}{h\Sigma(u - \bar{u})^2} = \frac{k}{h} b_{v/u}. \end{aligned}$$

This shows that regression coefficient depends on h and k but independent of A and B . This means regression coefficient is independent of the shift of origin but depends on change of scale. That is, the value of regression coefficient will remain same if we subtract a constant quantity from all the values of x and another constant quantity from the entire values y . But regression coefficient will be different if we divide all the values of x and y by two constant quantities.

12.5. Some Important Properties of Regression Coefficient

- (i) The regression co-efficient measures the average change in dependent variable for a unit change in independent variable.
- (ii) Regression coefficients are not symmetrical function of x and y . Suppose $b_{y/x}$ and $b_{x/y}$ are regression coefficients of y on x and x on y respectively, then $b_{y/x} \neq b_{x/y}$.
- (iii) Both the regression coefficients have the same sign.
- (iv) The correlation coefficient is the geometric mean of two regression coefficients, that is $r_{xy} = \sqrt{b_{x/y} \cdot b_{y/x}}$.

- (v) The arithmetic mean of the regression coefficients is equal to or greater than the Correlation coefficient, that is $\frac{b_{x/y} + b_{y/x}}{2} \geq r_{xy}$.
- (vi) If one of the regression coefficients is greater than one, then the other regression coefficient must be less than one since $r_{xy}^2 = b_{x/y} \cdot b_{y/x} \leq 1$.
- (vii) The sign of correlation coefficient and the regression coefficients are same, since all the measures depend on the sign of the covariance appearing in the numerator.

12.6. Difference between Simple Correlation and Simple Regression

Although correlation and regression both are used to analyze the relationship between two variables and there is an algebraic relationship between two coefficients, but there are differences between these two types of statistical tools with respect to their analysis, characteristics and interpretation. Some of the differences are listed below:

	Simple Correlation	Simple Regression
1	Simple correlation measures the direction and strength of linear relationship between two variables.	Regression measures the effect of independent variable on dependent variable
2	Correlation does not measure cause and effect relationship between the variables under study.	However, regression analysis measures the cause and effects relationship between the variables. Here, the variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
3	Question of dependent and independent variables do not arise in correlation analysis.	Dependent variable is regressed on the independent variable in regression analysis.
4	Correlation analysis is confined only to the study of linear relationship between the variables, and therefore has limited applications.	Regression analysis has much wider application as it studies linear as well as non-linear relationship between the variables.
5	Correlation coefficient is symmetrical about the variables, that means, $r_{xy} = r_{yx}$.	Regression coefficients are not symmetrical, that is $b_{y/x} \neq b_{x/y}$.
6	The value of the correlation coefficient lies between -1 and $+1$.	Regression coefficient can take any real value between $-\infty$ to $+\infty$.
7	Correlation coefficient is a pure number. It is a relative measurement.	Regression co-efficient is an absolute measurement. It depends on the units of measurement of the variables.
8	Correlation coefficient is independent of the shift of origin and change of scale.	Regression co-efficient is independent of shift of origin, but depends on change of scale.

12.7. The Coefficient of Determination r^2

The coefficient of determination r^2 is really the square of the correlation coefficient r . But it is a much more precise measure of the strength of the relationship between the two variables and is more precise interpretation because it can be presented as a proportion or as a percentage.

The coefficient of determination r^2 can be defined as the proportion of the variation in the dependent variable Y that is explained by the independent variable X , in the regression mode. In other words:

$$r^2 = \frac{\text{Explained Variation}}{\text{Total variation}} = \frac{\sum(Y_c - \bar{Y})^2}{\sum(y - \bar{y})^2}$$

For two variables it can also be calculated from the following formula

$$r^2 = \frac{[\sum(x - \bar{x})(y - \bar{y})]^2}{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}$$

It is the square of the simple correlation coefficient. It follows from the definition that the value varies from 0 to 1.

Interpretation of r^2 . The coefficient of determination is a summary measure that tells us how well the sample regression line fits the observed data. It is a measure of the goodness of fit of a regression.

Usually the closer the value of r^2 to 1, the better the model, that means, a regression line is better if its r^2 value is closed to 1. On the other hand, if it is closed to zero, it indicates that the regression line does not fit the data well and it fails to predict the future value of y for given value of x . Verbally, r^2 measures the proportion or percentage of the total variation in the dependent variable explained by the regression model.

For example, if for a model $y = a + bx$, $r^2 = 0.88$, it means that the independent variable x can explain 88% of the total variation in y or 88% of the total variation of dependent variable is due to the independent variable and the rest 12% is unexplained variation. In regression analysis r^2 is more meaningful than r .

12.8. Some Examples

Example 12.8.1. The following data give the test scores and sales made by nine salesmen during the last year of a big departmental store:

Test Scores : y	14	19	24	21	26	22	15	20	19
Sales (in lakh Taka) : x	31	36	48	37	50	45	33	41	39

- (i) Find the regression equation of test scores on sales.
- (ii) Find the test score when the sale is Tk. 40 lakh.
- (iii) Find the regression line of sales on test scores.
- (iv) Predict the value of sale if the test score is 30.
- (v) Compute the value of correlation coefficient with the help of regression coefficients and also with the original formula.

Solution. (i) The best fitted regression equation of test scores y on sales x is

$$\hat{y} = a + bx$$

Table for calculation of regression equation.

Sales (x)	Test Scores (y)	xy	x^2	y^2
31	14	434	961	196
36	19	684	1296	361
48	24	1152	2304	576
37	21	777	1369	441
50	26	1300	2500	676
45	2	990	2025	484
33	15	495	1089	225
41	20	820	1681	400
39	19	741	1521	361
$\Sigma x = 360$	$\Sigma y = 180$	$\Sigma xy = 7393$	$\Sigma x^2 = 14746$	$\Sigma y^2 = 3720$

Here $\bar{y} = \frac{180}{9} = 20$, $\bar{x} = \frac{360}{9} = 40$

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{7393 - \frac{180 \times 360}{9}}{14746 - \frac{(360)^2}{9}} = \frac{7393 - 7200}{14746 - 14400} = \frac{193}{346} = 0.56$$

$$a = \bar{y} - b\bar{x} = 20 - 0.56 \times 40 = 20 - 22.4 = -2.4$$

Hence the required regression equation of test scores on sales is

$$\hat{y} = a + bx = -2.4 + 0.56x$$

(ii) When $x = 40$, the value of y is

$$\hat{y} = -2.4 + 0.56 \times 40 = -2.4 + 22.4 = 20$$

The test score is 20 when the sale is Tk. 40 lakh.

(iii) The regression equation of sales x on test scores y is

$$x = c + dy$$

$$\text{Here } d = \frac{\frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{n}}{\frac{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}{n}} = \frac{7393 - \frac{180 \times 360}{9}}{3720 - \frac{(180)^2}{9}} = \frac{7393 - 7200}{3720 - 3600} = \frac{193}{120} = 1.61$$

$$c = \bar{x} - dy = 40 - 1.61 \times 20 = 40 - 32.2 = 7.8.$$

Hence the required regression equation of sales on test scores is

$$\hat{x} = c + dy = 7.8 + 1.61y$$

(iv) When $y = 30$, then x would be

$$x = 7.8 + 1.61 \times 30 = 7.8 + 48.3 = \text{TK.56.1 lakh}$$

The predicted sales would be Tk. 56.1 lakh if the test score is 30.

(v) The formula of correlation coefficient with the help of regression coefficients is $r = \sqrt{c \times d} = \sqrt{0.56 \times 1.61} = \sqrt{0.9016} = 0.95$

The value of correlation coefficient would be positive since the regression coefficients are positive.

The correlation coefficient from the original formula,

$$r = \frac{\frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{n}}{\sqrt{\left\{ \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n} \right\} \left\{ \frac{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}{n} \right\}}} = \frac{193}{\sqrt{346 \times 120}} = \frac{193}{203.96} = 0.95.$$

Here the coefficient of determination r^2 is 0.9016.

Comment. The test scores of the salesmen explain 90.16% sales of the store. Hence both the regression lines fit well. Any prediction or estimate made by the regression lines may be accepted.

When the means of x and y are whole numbers. The calculation of regression equation discussed above is quite difficult when the values of x and y are large. The work can be simplified if the all values of x and y are subtracted from their respective means \bar{x} and \bar{y} when they are whole number. Now we shall cite an example.

Example 12.8.2. The following data relate to advertising expenditure (in lakha of taka and sales (in crores of taka) of a firm:

Advertising Expenditure : x (in lakh Tks)	10	12	13	17	18
Sales (in crores of Tks) : y	5	6	7	9	13

- (i) Find the equation of the regression line of sales y on expenditure x .
- (ii) Predict the sales target for an advertising expenditure of Tk. 20 lakhs.
- (iii) Find the equation of the regression line of expenditure x on sales y .
- (iv) Predict the advertising expenditure for a sales target Tk 20 crores.
- (v) Find the correlation coefficient with the help of regression coefficients.

Solution. Here $\bar{x} = \frac{\Sigma x}{n} = \frac{70}{5} = 14$; $\bar{y} = \frac{\Sigma y}{n} = \frac{40}{5} = 8$. Since \bar{x} and \bar{y} are whole numbers, we can subtract all the values of x and y from their respective means to simplify the calculations. Now let us make a table for calculation of regression equation.

X	$u = (x - 14)$	u^2	y	$v = (y - 8)$	v^2	uv
10	-4	16	5	-3	9	12
12	-2	4	6	-2	4	4
13	-1	1	7	-1	1	1
17	3	9	9	1	1	3
18	4	16	13	5	25	20
$\Sigma x = 70$	$\Sigma u = 0$	$\Sigma u^2 = 46$	$\Sigma y = 40$	$\Sigma v = 0$	$\Sigma v^2 = 40$	$\Sigma uv = 40$

(i) The regression of y on x is $\hat{y} - \bar{y} = b(x - \bar{x})$

$$\text{Here } b = \frac{\Sigma uv}{\Sigma u^2} = \frac{40}{46} = 0.87$$

Hence the regression equation of y on x is

$$\begin{aligned} \hat{y} - \bar{y} &= b(x - \bar{x}) \\ \Rightarrow \hat{y} &= \bar{y} + b(x - \bar{x}) = 8 + 0.87(x - 14) \\ &= 8 + 0.87x - 12.18 = -4.18 + 0.87x \end{aligned}$$

(ii) When $x = 20$, the value of y is

$$\hat{y} = -4.18 + 0.87 \times 20 = \text{Tk. } 13.22 \text{ Crores.}$$

(iii) The best fitted regression line of expenditure x on sales y is

$$\hat{x} - \bar{x} = d(y - \bar{y})$$

Here d is the regression coefficient of x on y .

$$d = \frac{\Sigma uv}{\Sigma v^2} = \frac{40}{40} = 1.0.$$

Hence the required regression line of x on y is

$$\hat{x} - \bar{x} = d(y - \bar{y})$$

$$\begin{aligned}\Rightarrow \hat{x} &= \bar{x} + d(y - \bar{y}) \\ &= 14 + y - 8 \\ &= 6 + y\end{aligned}$$

(iv) When $y=20$, then the advertising expenditure of the firm is

$$\hat{x} = 6 + 20 = \text{Tk.} 26 \text{ lakhs.}$$

(v) The correlation coefficient between sales and expenditure is

$$r = \sqrt{b \times d} = \sqrt{0.87 \times 1} = 0.933.$$

When the means of x and y are not whole numbers. When the actual means of x and y are not whole numbers, the calculation can also be simplified by subtracting two suitable whole numbers from each value of x and y since the regression coefficients are independent of the shift of origin. In this case the regression equation of y on x is

$$\hat{y} - \bar{y} = b(x - \bar{x})$$

$$b = \frac{\sum uv - \frac{\sum u \sum v}{n}}{\sum u^2 - \frac{(\sum u)^2}{n}}$$

Example 12.8.3. The following data give the ages and blood pressure of 10 women.

Age (in years) (x)	56	42	36	47	49	42	60	72	63	55
Blood pressure (y)	147	125	118	128	145	140	155	160	149	150

- (i) Obtain the regression line of y on x .
- (ii) Estimate the blood pressure of a woman whose age is 50 years.

Solution. Here $\bar{x} = \frac{\sum x}{n} = \frac{522}{10} = 52.2$ and $\bar{y} = \frac{\sum y}{n} = \frac{1417}{10} = 141.7$

Here the means of x and y are not whole number. So, we can subtract two suitable whole numbers from the values of x and y to make the calculation easy. Here we can take $A = 49$ and $B = 145$ as they are the two values of x and y situated in the middle of the two series of x and y . Then the new variables are

$$u = x - 49 \quad \text{and} \quad v = y - 145$$

Then the table of calculation will be as follows:

Calculation of regression equation

Age : x	Blood Pressure : y	$u = x - 49$	$v = y - 145$	u^2	v^2	uv
56	147	7	2	49	4	14
42	125	-7	-20	49	400	140
36	118	-13	-27	169	729	351
47	128	-2	-17	4	289	34
49	145	0	0	0	0	0
42	140	-7	-5	49	25	35
60	155	11	10	121	100	110
72	160	23	15	529	225	345
63	149	14	4	196	16	56
55	150	6	5	36	25	30
522	1417	32	-33	1202	1813	1115

(a) The equation of the best-fitted regression line is

$$\hat{y} = a + bx$$

$$\Sigma uv - \frac{(\Sigma u)(\Sigma v)}{n}$$

$$\text{where } a = \bar{y} - b\bar{x} \text{ and } b = \frac{n}{\Sigma u^2 - \frac{(\Sigma u)^2}{n}}$$

Here we have to find b first

$$b = \frac{\frac{1115 - \frac{32 \times (-33)}{10}}{1202 - \frac{(32)^2}{10}}}{\frac{1115 + 105.6}{1202 - 102.4}} = \frac{1115 + 105.6}{1202 - 102.4} = \frac{1220.6}{1099.6} = 1.11$$

$$a = 141.7 - 1.11 \times 52.2 = 83.758$$

Hence the best-fitted regression line is

$$\hat{y} = 83.76 + 1.11x$$

The blood pressure of the woman whose age 50 is

$$\hat{y} = 83.76 + 1.11(50) = 139.26.$$

Example 12.8.4. A researcher wants to find out if there is any relationship between the heights of the sons and the heights of the fathers. He took a random sample of seven fathers and their seven sons. Their heights in inches are given below:

Height of father: (In inches)	68	63	66	67	65	67	66
Height of Son: (In inches)	70	66	65	69	68	67	64

- (a) Fit a regression line of the height of father y on the height of son x .
- (b) Predict the height of son if father's height is 70 inches.
- (c) Fit a regression line of the height of son on the height of father.
- (d) Predict the height of father if son's height is 65 inches.
- (e) Calculate the correlation coefficient with the help of regression coefficients and from the original formula.

Solution. Here $\bar{x} = \frac{\Sigma x}{n} = \frac{462}{7} = 66$ and $\bar{y} = \frac{\Sigma y}{n} = \frac{469}{7} = 67$

Here both the means are whole numbers. We define the new variables as

$$u = x - 66 \text{ and } v = y - 67$$

Calculation table

x	$u = x - 66$	u^2	y	$v = y - 67$	v^2	uv
68	2	4	70	3	9	6
63	-3	9	66	-1	1	3
66	0	0	65	-2	4	0
67	1	1	69	2	4	2
65	-1	1	68	1	1	-1
67	1	1	67	0	0	0
66	0	0	64	-3	9	0
$\Sigma x = 462$		$\Sigma u = 0$	$\Sigma u^2 = 16$	$\Sigma y = 469$	$\Sigma v = 0$	$\Sigma v^2 = 28$
						$\Sigma uv = 10$

(a) Hence the required regression line of y on x is

$$\begin{aligned}\hat{y} - \bar{y} &= b(x - \bar{x}) \\ \hat{y} &= \bar{y} + b(x - \bar{x}) \\ &= 67 + 0.625x - (0.625)(66) \\ &= 67 + 0.625 - 41.25 \\ &= 25.75 + 0.625x\end{aligned}$$

(b) Hence, if the height of the father is 69 inches or $x = 69$, the height of the son is

$$\begin{aligned}\hat{y} &= 25.75 + (0.625)(69) \\ &= 25.75 + 43.125 = 68.875 \text{ inches}\end{aligned}$$

(c) The regression line of son x on father y is

$$\hat{x} - \bar{x} = d(y - \bar{y})$$

Here d is the regression coefficient of x on y .

$$d = \frac{\Sigma uv}{\Sigma v^2} = \frac{10}{28} = 0.357$$

Hence the required regression line of x on y is

$$\begin{aligned}\hat{x} - \bar{x} &= d(y - \bar{y}) \\ \hat{x} &= \bar{x} + d(y - \bar{y}) \\ &= 66 + 0.357y - (0.357)(67) \\ &= 66 + 0.357y - 23.919 \\ &= 42.081 + 0.357y\end{aligned}$$

(d) Hence, if the height of son is 71 inches, the height of father would be

$$\hat{x} = 42.081 + (0.357)(71) = 42.081 + 25.357 = 67.438 \text{ inches.}$$

(e) Correlation coefficient by using regression coefficients is

$$r = \sqrt{b \times d} = \sqrt{(0.625)(0.357)} = \sqrt{0.223} = 0.472.$$

Correlation coefficient by the original formula

$$r = \frac{\sum uv}{\sqrt{(\sum u^2)(\sum v^2)}} = \frac{10}{\sqrt{16 \times 28}} = \frac{10}{21.166} = 0.472.$$

Example 12.8.5. A researcher wants to find out if there is any relationship between the ages of the husbands and the ages of the wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below:

Age of Husband (in years) : x	39	25	29	35	32	27	37
Age of Wife (in years) : y	37	18	20	25	25	20	30

- (a) Compute the regression line of y on x by direct method and the short cut method.
- (b) Predict the age of wife whose husband's age is 45 years.
- (c) Find the regression line of x on y and estimate the age of husband if the age of his wife is 28 years.

Solution. The equation of the best-fitted regression line of y on x is

$$\hat{y} = a + bx$$

$$\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

$$\text{where } b = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} ; \quad a = \bar{y} - b\bar{x}$$

Computation Table

x	y	x^2	y^2	xy
39	37	1521	1369	1443
25	18	625	324	450
29	20	841	400	580
35	25	1225	625	875
32	25	1024	625	800
27	20	729	400	540
37	30	1369	900	1110
$\Sigma x = 224$	$\Sigma y = 175$	$\Sigma x^2 = 7334$	$\Sigma y^2 = 4643$	$\Sigma xy = 5798$

$$b = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{5798 - (224)(175)/7}{7334 - \frac{(224)^2}{7}} = \frac{5798 - 5600}{7334 - 7168} = \frac{198}{166} = 1.193;$$

$$a = \bar{y} - b\bar{x} = 25 - (1.193)(32) = 25 - 38.176 = -13.176$$

Hence the required regression line is

$$\hat{y} = -13.176 + 1.193x$$

(b) Hence, if the age of husband is 45, the probable age of wife would be

$$\hat{y} = -13.176 + 1.193 \times 45 = 40.51 \text{ years.}$$

$$\text{Here, } \bar{x} = \frac{\Sigma x}{n} = \frac{224}{7} = 32 \text{ and } \bar{y} = \frac{\Sigma y}{n} = \frac{175}{7} = 25$$

Here both the means are whole numbers.

We define the new variables, $u = x - 32$ and $v = y - 25$

Short cut method for finding the regression lines

(a) The equation of the best fitted regression of y on x is

$$\hat{y} - \bar{y} = b(x - \bar{x})$$

Here b is the regression coefficient of y on x .

$$b = \frac{\Sigma uv}{\Sigma u^2}$$

(b) The equation of the best fitted regression of x on y is

$$\hat{x} - \bar{x} = d(y - \bar{y})$$

Here d is the regression coefficient of x on y .

$$d = \frac{\Sigma uv}{\Sigma v^2}$$

Computation Table

X	$u=x-32$	u^2	y	$v=y-25$	v^2	uv
39	7	49	37	12	144	84
25	-7	49	18	-7	49	49
29	-3	9	20	-5	25	15
35	3	9	25	0	0	0
32	0	0	25	0	0	0
27	-5	25	20	-5	25	25
37	5	25	30	5	25	25
$\Sigma x = 224$	$\Sigma u = 0$	$\Sigma u^2 = 166$	$\Sigma y = 175$	$\Sigma v = 0$	$\Sigma v^2 = 268$	$\Sigma uv = 198$

(c) The equation of the best fitted regression of y on x is

$$\hat{y} - \bar{y} = b(x - \bar{x})$$

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

The formula for b is

$$b = \frac{\Sigma uv}{\Sigma u^2} = \frac{198}{166} = 1.193$$

Hence the required regression line of y on x is

$$\begin{aligned}\hat{y} &= \bar{y} + b(x - \bar{x}) \\ &= 25 + 1.193(x - 32) \\ &= -13.176 + 1.193x\end{aligned}$$

(d) The equation of the best fitted regression of x on y is

$$\hat{x} - \bar{x} = d(y - \bar{y})$$

$$= \bar{x} + d(y - \bar{y})$$

The regression coefficient of x on y is

$$d = \frac{\Sigma uv}{\Sigma v^2} = \frac{198}{268} = 0.739$$

Hence the required regression line of x on y is

$$\begin{aligned}\hat{x} &= \bar{x} + d(y - \bar{y}) \\ &= 32 + 0.739y - (0.739)(25)\end{aligned}$$

$$\begin{aligned} &= 32 + 0.739y - 18.475 \\ &= 13.525 + 0.739y \end{aligned}$$

Hence, if the age of wife is 28 years, the estimate age of husband is

$$\hat{x} = 13.525 + (.739)(28) = 34.22 \text{ years.}$$

Example 12.8.6. A researcher got the following summary data from a community:

Correlation coefficient between the ages of husbands and wives is 0.9

- The average age of husbands = 30 years
- The average age of wives = 25 years
- The standard deviation of the ages of husbands = 5
- The standard deviation of the ages of wives = 6

On the basis of the above information, compute the regression lines of the ages of

- (i) Husband on the ages of wives and estimates the age of husband if the age of wife is 18;
- (ii) Wives on the ages of husbands and predict the age of wife if the age of husband is 40 years.

Solution. Suppose the age of wife is y and the age of husband is x . Here, we have $\bar{x} = 30$, $\bar{y} = 25$, $\sigma_x = 5$, $\sigma_y = 6$ and $r = 0.9$

(i) The required regression line will be x on y . On the basis of the above information, the regression equation of x on y is

$$\begin{aligned} \hat{x} &= \bar{x} + r \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \\ &= 30 + \frac{(0.9)(5)}{6} (y - 25) \\ &= 30 + 0.75y - 18.75 \\ &= 11.25 + 0.75y \end{aligned}$$

When $y = 18$, then the probable age of husband is

$$\hat{x}_{18} = 11.25 + 0.75 \times 18 = 24.75 \text{ years.}$$

(ii) The regression line of the age of wife y on age of husband x is

$$\begin{aligned} \hat{y} &= \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \\ &= 25 + \frac{(0.9)(6)}{5} (x - 30) \end{aligned}$$

$$= 25 + 1.08x - 32.4$$

$$= -7.4 + 1.08x$$

When $x = 40$, the probable age of wife is :

$$\hat{y}_{40} = -7.4 + 1.08 \times 40 = 38.8 \text{ years.}$$

Example 12.8.7. The following summary data refers to the rainfall and production of a rabi crop in a particular region :

	Rainfall (In inches)	Production (In quintals)
Mean	29	40
Standard deviation	3	6

Coefficient of correlation between rainfall and production is 0.8.

- (i) Find the regression line of production on rainfall.
- (ii) Find the most probable production corresponding to a rainfall of 40 inches

Solution. (i) Let rainfall be denoted by x and production by y . The equation of the best fitted regression line of production on rainfall is.

$$\hat{y} = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Here $\bar{x} = 29$, $\bar{y} = 40$, $\sigma_x = 3$, $\sigma_y = 6$ and $r = 0.8$

$$\begin{aligned}
 \hat{y} &= \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \\
 &= 40 + (0.4) \frac{6}{3} (x - 29) \\
 &= 40 + 0.8x - 23.2 \\
 &= 16.8 + 0.8x
 \end{aligned}$$

(ii) When $x = 40$ inches, the possible production the rabi crop is

$$\begin{aligned}
 \hat{y}_{40} &= 16.8 + (0.8)(40) \\
 &= 16.8 + 32 = 48.8 \text{ quintals}
 \end{aligned}$$

Questions

1. Distinguish between correlation and regression. State some important properties of regression coefficient.
2. What are regression lines? Define regression coefficients. Show that correlation coefficient is the geometric mean of the regression coefficients.
3. Distinguish between regression model and regression line. Write the equation of the best fitted regression line of y on x . Discuss its role in business.
4. Define dependent and independent variables. Is it reasonable to write two regression lines (i) y on x and (ii) x on y ? Justify in favour of your answer.
5. Distinguish between correlation and regression. State some uses of regression in business.

Exercise

6. Consider the following data set on two variables x and y :

$x : 1 \ 2 \ 3 \ 4 \ 5 \ 6$

$y : 6 \ 4 \ 3 \ 5 \ 4 \ 2$

- (a) Find the equation of the regression line y on x .
- (b) Graph the line on a scatter diagram.
- (c) Estimate the value of y when $x = 4.5$
- (d) Predict the value of y when $x = 8$.

Ans. (a) $\hat{y} = 5.799 - 0.514x$; (c) $\hat{y} = 3.486$; (d) $\hat{y} = 1.687$

7. Calculate the regression equations of

(a) y on x and

(b) x on y from the following pairs of data set:

$x : 1 \ 2 \ 3 \ 4 \ 5$

$y : 2 \ 5 \ 3 \ 8 \ 7$

(c) Estimate the value y when $x = 3.5$

(d) Predict the value x when $y = 10$

Ans. (a) $\hat{y} = 1.10 + 1.30x$; (b) $\hat{x} = 0.5 + 0.5x$; (c) $\hat{y} = 5.65$; (d) $\hat{x} = 5.5$.

8. Calculate the regression lines of

(i) y on x

(ii) x on y

(iii) Estimate the value of y when $x = 16$.

(iv) Find the probable value of x when $y = 16$

(v) Find the correlation coefficient with the help of the regression coefficients.

Ans. (i) $\hat{y} = 8.5 + 1.015x$, (ii) $\hat{y}_{16} = 5.04$ (iii) $\hat{x} = -7.05 + .91y$ (iv) $\hat{x}_{16} = 7.51$, $r = 0.96$.

Application

9. The following summary data were obtained for closing prices of twelve stocks x on Chittagong Stock Exchange on a certain day, along with the volume of sales in thousands of shares y :

$$\Sigma x = 580, \Sigma y = 370, \Sigma xy = 11494, \Sigma x^2 = 41,658, \Sigma y^2 = 17,206$$

Find

- (i) The regression line of y on x .
- (ii) The regression line of x on y .
- (iii) Correlation coefficient between x and y .

Ans. (i) $\hat{y} = 53.55 - 0.47x$, (ii) $\hat{x} = 79.16 - 1.1y$, (iii) $r = -0.517$

10. A study was made by a retail merchant to determine the relation between weekly advertising expenditure and sales. The following data were recorded:

Expenditure (\$)	40	20	25	20	30	50	40	20	50	40	25	50
Sales (\$)	385	400	395	365	475	440	490	420	560	525	480	510

- (i) Plot a Scatter diagram.
- (ii) Find the equation of the regression line to predict weekly sales from advertising expenditures.
- (iii) Estimate the weekly sales when advertising costs are 350.

Ans. (b) $\hat{y} = 343.699 + 3.221x$; (c) $\hat{y} = 456$

11. A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows:

Temperature, x : 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0

Converted sugar, y : 8.1 7.8 8.5 9.8 9.5 8.9 8.6 10.2 9.3 9.2 10.5

- (a) Estimate the linear regression line of y on x .
- (b) Estimate the amount of converted sugar produced when the coded temperature is 1.75.

Ans. (a) $\hat{y} = 6.414 + 1.809x$ (b) $\hat{y} = 9.58$

12. In a study between the amount of rainfall and the quantity of air pollution removed, the following data were collected:

Daily Rainfall (centimeter)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1	7.5
Pollution Removed (Micrograms per cubic meter)	126	121	116	118	114	118	132	141	108

- (a) Find the equation of the regression line to predict the pollution removed from the amount of daily rainfall.
- (b) Estimate the amount of pollution removed when the daily rainfall is $x=4.8$ centimeter.

13. The following data relate to advertising expenditure (in lakh taka) and their corresponding sales in crores taka of a firm in last five years:

Advertising Expenditure, x : 10 12 15 23 20
 Sales, y : 14 17 23 25 21

- (a) Find the regression line of sales on expenditure.
- (b) Estimate the sales of the firm when the advertising expenditure is Tk.30 lakhs.
- (c) Find the equation of the regression line of expenditure on sales
- (d) Estimate the expenditure of the firm when the sales is 35 crores.
- (e) Calculate the correlation coefficient and the coefficient of determination of advertising expenditure and sales,

Ans. (a) $\hat{y} = 8.608 + 0.712x$; (b) Tk.29.968 crores;
 (c) $\hat{x} = -5 + 1.05y$; (d) Tk.31.75Lakhs; $r = 0.864$ and
 (e) $r^2 = 0.747$.

14. A researcher wants to find out if there is any relationship between the ages of the husbands and the ages of the wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 6 couples whose respective ages are given below:

Age of Husband : x	39	25	35	32	27	37
Age of Wife : y	37	18	25	25	20	30

- (a) Find the regression line of wife on husband and estimate the probable age of wife if the age of husband is 30.
 - (b) Find the regression line of husband on wife and predict the probable age of husband if the age of wife is 40.
 - (c) Compute the Karl Pearson's coefficient of correlation with the help of the regression coefficients and from the original formula.
15. The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age:

Age of cars in years : 2 4 6 8
 Maintenance cost : 10 20 25 30
 (in hundred taka)

Estimate the maintenance cost for a 7 year old car.

16. The following data give price and supply of a commodity for last 9 years:

Supply	80	82	86	91	83	85	89	96	93
Price	145	140	130	124	133	127	120	110	116

- (a) Find a regression line of price on supply.
- (b) Compute the correlation coefficient between price and supply.

Ans. (a) $y=301.75+2x$; (b) $r = -0.96$.

17. (i) Find the regression line of production on rainfall from the following data:

	Rainfall in inches : x	Production in kg : y
Average	30	500
Standard deviation	5	100
Correlation coefficient = 0.8		

- (ii) Also find the most likely production corresponding to a rainfall 40 inches

$$\text{Ans. (i)} \hat{y} = 20 + 16x; \text{(ii)} 660\text{kg.}$$

18. The following data about the sales and advertisement expenditure of a firm is given below:

	Sales (in crores Tk.)	Advertisement Expenditure (in crores Tk.)
Mean	40	6
Standard deviation	10	1.5
Correlation Coefficient $r = 0.9$		

- (i) Find the regression line of sales x on expenditure y.
(ii) Estimate the sales of the firm when the expenditure on advertisement is Tk. 10 crores.
(iii) Find the regression line of advertisement expenditure on sales.
(iv) What should be the advertisement expenditure if the firm proposes a sale target of Tk. 60 crores.

$$\text{Ans. (i)} \hat{y} = 4 + 6x, \text{(ii)} \text{Tk } 64 \text{ crores, (iii)} \hat{x} = 0.6 + 0.135y, \text{(iv)} \text{Tk } 8.7 \text{ crores}$$

19. The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age:

Age of cars in years	2	4	6	8
Maintenance cost (in hundred taka)	10	20	25	30

Estimate the maintenance cost for a 7 year old car.

20. The following table gives the per unit cost (in hundred Taka) and selling price (in hundred Taka) of 8 products of a company

Product	A	B	C	D	E	F	G	H
Cost price	10	15	14	20	31	34	57	65
Selling price	11.5	18.0	18.5	20.9	33.2	39.0	64.2	74

- (a) Fit a suitable regression line and comment.
(b) Estimate the profit incurred by the company for a product whose cost price is Taka 50 (hundred).

CHAPTER - 13

INDEX NUMBERS

13.1. Introduction

It is known that most of the economic or business phenomena change over a period of time, from one region to other, etc. We may be interested in comparing economic condition over time or space. Let us consider the cases:

- i) Consumers may want to know if the current period cost of living is higher or lower than in the past period (if the prices are rising or falling),
- ii) Workers may be interested to compare current wages with the past (if the wages are adjusted to the inflation),
- iii) Businessmen may look at the price currently paid or received with the past,
- iv) Shareholders may be interested to compare the average change in the current price of share with some other previous period,
- v) Government officials may ponder the difference between the current general price levels with the past,
- vi) Businessmen may want to compare the profit of selling commodities in two different places,
- vii) An economist may want to compare the consumption pattern of two groups of population.

A certain kind of yardstick can greatly facilitate the desired comparison in all these cases and a million more. Such a yardstick is provided by index number.

A.L. Bowley mentioned that index numbers are used to measure the changes in some quantity which we can not observe. A.M. Tuttle defined as a single ratio which measures the combined change of several variables between two different times, places or situations.

An index in its simplest form requires a comparison of two values of the same series at different time periods. The comparison is done by making a ratio of the two values. For example, if the relative change in the prices of the commodities in 2008 is to be compared with that of 2007, then 2007 is the base period and 2008 is the current period. For the convenience of comparison, the index number is expressed in percentage form, i.e. the base year factor is considered as 100. In business, index number can be defined as follows:

Definition: Index number is a pure number which measures the relative change of price or quantity or value of a commodity or a group of commodities of a particular year called current year with respect to some standard year called base year.

Current Period: The year for which index number calculated is called current year.

Base Period: The year compare to which an index number calculated is called base year. The base year index number is taken as 100.

In general index number may be defined as "An index number is a statistical device designed to measure the relative change in the level of a phenomenon with respect to time, geographical location or other characteristics such as income etc".

When we say that the index number of wholesale prices is 125 for the period June, 2011 compared to June, 2010, it means there is a increase in the prices of wholesale commodities to the extend of 25 percent.

Originally, the index numbers were developed for measuring the effect of changes in the price level. An Italian economist named Giovani Rinaldo Carli in 1764-constructed first price index number. Columbus discovered America in 1492. It had a great impact on the economy of Italy. As a result, the prices of the essential commodities were raised significantly. He computed the price index for the important three commodities named food stuffs, fuel and win for the year 1775 taking 1500 as base year. Through this index number he got a good picture for the relative changes of the prices of these commodities for the last 250 years. Today index numbers have wide range of applications. It is used by economists to measure change in price, by psychologists to measure difference in I.Q., by historians to measure change in population and by labor unions to assure that wages increases with general price level.

Today index numbers occupy a place of great prominence in business statistics. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies. That is why index numbers are called the Barometers of economic activities.

12.2. Some Characteristics of Index Numbers

An index number has the following characteristics:

- (i) Index numbers are specialized types of average.
- (ii) It is a pure number.
- (iii) It is always expressed in percentage.
- (iv) Index numbers measure the change in the level of a phenomenon.
- (v) Index numbers measure the effect of changes over a period of time.

13.3. Uses Index Number

Index numbers are indispensable tools of economic and business analysis. They are also used in all sciences—natural, social and physical. The main uses of index numbers can be summarized as follows:

1. Index numbers are used as Economic Barometers;
2. Index numbers help in studying trends and tendencies of prices, sales, etc.;
3. Index numbers help in formulating decisions and policies;
4. Consumer price index or cost of living index numbers are particularly useful for
 - i) Formulating economic policy, adjusting income payments, and measuring real earning.
 - ii) Measuring the purchasing power of money.
 - iii) Wage negotiations and wage contracts.

13.4. Problems in the Construction of Index Numbers

Different problems are faced in the construction of different types of index numbers. Here we shall discuss the problems for constructing price index only. They are:

- i) Purpose of index number.
- ii) Selection of base period.
- iii) Selection of commodities to be included.
- iv) Collection of data for index number.
- v) Type of average to be used.
- vi) Selection of appropriate weight.
- vii) Selection of an appropriate formula

(i) Purpose of index number

The purpose of the index numbers should be rigorously defined. The purpose of index numbers would help in deciding about the nature of data to be collected, the choice of the base year, the formula to be used and other related matters. If we are going to construct the consumer's price index, we must know the class of consumers whose cost of living we intend to measure. Whether it is the cost of living of the lower class, garment's workers or industrial workers. Great care should be exercised to include the commodities used by the selected class. In fact, before constructing index numbers, we must precisely know what we want to measure, and what we intend to use this measurement for.

(ii) Selection of base period

The base period is the period with which comparison of relative changes are made. It may be a year, month or a day. The index for base period is always

taken as 100. Usually, the selection of base period would primarily depend upon the purpose of the index. However, the following points should take into consideration in choosing a base period.

- (a) The base period should be a period of normal and stable economic conditions. It should be free from abnormalities and random or irregular fluctuations like strikes, lockouts, booms, depressions, famines, wars, earthquakes, etc. In such cases, we can take an average of a few years as base. The process of averaging will reduce the effect of extremes.
- (b) The base year should not be too far from the current year. Since index numbers are helpful in decision-making and economic policies are often a matter of short period, we should choose a base year, which is relatively close to the current year. If the base year is too far in the past, we cannot make valid meaningful comparisons since there might have been appreciable change in the tastes, customs, habits and fashion of the people during the intervening period. This would have affected the consumption pattern of the various commodities to a marked extent making comparison difficult. For example, for deciding increase in dearness allowance at present say for the year 2012 there is no advantage in taking 2000 or even 2005 as the base year. The comparison should be the preceding year or the year after which dearness allowance has not been revised.
- (c) Fixed base or chain base. While selecting the base year, a decision has to be made whether the base shall remain fixed or not. If the period of comparison is fixed for current years, it is called fixed base method. If, on the other hand, the prices of the current year are linked with the prices of the preceding year and not with the fixed year or period, it is called chain base method. Chain base method is useful in cases where there are quick and frequent changes in fashion, tastes, and habits of the people. In such cases comparison with the preceding year is more worthwhile.

(iii) Selection of commodities to be included

In constructing index numbers, all items cannot be included and hence one has to select a sample. The commodities should be selected in such a manner that they are representative of the tastes, habits and customs of the people for whom the index is meant. The purpose of index shall help in deciding the number of commodities. If the purpose of an index is to measure cost of living of taxi driver of Chittagong city we should select those commodities or items which are consumed by persons belonging to this group and due care should be taken not to include the goods which are not ordinarily consumed by the individuals of the selected groups.

(iv) Collection of data for index number

This is perhaps the most tedious problem in the construction of index number. The data is the set of prices and quantities of the selected commodities consumed for different periods that constitute the raw material for the construction of consumer price index. The data should be collected from reliable sources such as standard trade journals, official publications, periodical special reports from the procedures, and exporters etc. or through field agency. Since data play the important role in the construction of index number, their accuracy, comparability, sample representativeness and adequacy should be bear in mind. In any case the data should be strictly pertaining to what is being measured. For example, for the construction of retail price index numbers, the prices for an adequate number of commodities should be obtained from fair price shops, departmental stores, and not from wholesale dealers.

(v) Type of average to be used

Since index numbers are specialized average, a judicious choice of average to be used in their construction is of great importance. Usually (i) arithmetic mean, geometric mean and mode may be used in constructing index number.

Median is the easiest to calculate of all the three, completely ignores the extreme observations while arithmetic mean, though easy to calculate, is unduly affected by extreme observations. Moreover, neither arithmetic mean nor median are reversible and hence do not reflect typical movements of prices and quantities. Theoretically speaking, geometric mean is the best average in the construction of index numbers because of the following reasons:

- (a) In the construction of index numbers we are concerned with ratios or relative changes and the geometric mean gives equal weights to equal ratio of change;
- (b) Geometric mean is less susceptible to major variations as a result of violent fluctuations in the values of the individual items and
- (c) Index numbers calculated by using this average are reversible and, therefore, base shifting is easily possible. The geometric mean index always satisfies the time reversal test. Despite theoretical justification for favouring geometric mean, arithmetic mean is more popularly used while constructing index numbers, as it is simpler to compute than geometric mean. However, in the interest of greater accuracy geometric mean should be preferred.

(vi) Selection of appropriate weight

Various items or commodities included in the index are not of equal importance. For example, for the construction of cost of living index commodities, like rice, wheat, kerosene, and clothing etc., included in the

index are not of equal importance, proper weights should be attached to them to take into account their relative importance. This is done by assigning weights. The term weight refers to the relative importance of the different items in the construction of the index. For example, for constructing price index, usually quantity consumed or produced in the base year or current year or their average may be used as weights. Similarly, for the construction of quantity index, price per unit of the items for the base year or current year or their average can be used as weights. The selection of the weight depends upon the purpose of the index number, the type of formula employed and the data that are available.

(vii) Selection of an appropriate formula

A large number of formulae have been devised for constructing the index. A decision has therefore to be made as to which formula is the most suitable for the purpose. Prof. Irving Fisher has suggested that an appropriate index is that which satisfies time reversal and factor reversal test. Theoretically, Fisher's method is considered as ideal for constructing index numbers since it satisfies both the

time reversal and factor reversal test. However, the choice of the formula depends upon the availability of the data regarding the prices and quantities of the selected commodities in the base and current year.

13.5. Classification of Index Numbers

In business and economics there are three types of index numbers. They are

- i) Price index,
- ii) Quantity index and
- iii) Value index.

Sometimes index numbers are constructed for some special purposes which are called special purpose index numbers.

This book pays more attention to the price and quantity index, because other two has very limited application and even if one understands the concepts of price and quantity index, it would not be difficult to understand the remaining two.

(i) Price index

Perhaps the most widely used, popular and important index is the Price Index (PI). The consumer price index (CPI) traces the movement of retail prices. It has a number of important uses related to the measurement of price inflation and changes in the cost of living. For example, it is used

- a) As an economic indicator: it is watched closely as a measure of success or failure of government economic policy and, in this sense, it is utilized

by both business men and union leaders as well as by individuals as a guide to making economic decision. The consumer price index is the best available indication of changes in monthly living costs.

- b) As a price deflator: another important use of the consumer price index is as a price deflator. This results from the fact that economists, forecasters and business decision makers are concerned with economic models that are representative of the complex workings of our economy. The Consumer price index is often used to adjust nominal wages to real wages by adjusting for changes in the cost of living.
- c) As an escalator: the index is used as a measure of the "cost of living".

(ii) Quantity Index

A quantity index measures the changes in the amounts of goods produced over a period or from one industry to another industry. For example, Government may be interested in comparing the quantity produced in different industries by computing quantity indices for production in a number of industries such as steel, chemicals, garments, etc. which would provide average measures of changing output in these industries over two periods, industries located in different places.

(iii) Value Index

A value index reflects the combined movement of prices and quantities of goods and services. It is measured as the relative change in total expenditure in the current year as compared to the base year. Value-index numbers have less practical use than price and quantity indices. This is because we are usually interested in the separate movement of these two factors, and this can not be determined from the composite changes.

Special Purpose Index

Some special forms of index numbers may be needed for special use. For example house rent index which is used to compare the relative change in the house rent in current year as compared to some convenient previous year. This index also provides an indication if there is any money inflation over the two periods under consideration. Since the construction of this type of index is same as that of price index or quantity index and the use of such index is very special, this index is beyond discussion of this handout.

13.6. Features of Index Numbers

Depending on its definition, nature and interpretation, index numbers can be characterized by the following features:

- a) Index numbers are averages of special type: As we know, average measures is a single number which represents a particular type of

characteristic of data set and this average can be used for comparison of two types of data sets only if the units of measurements of observations for two sets are the same. Hence, index number is also considered as particular type of average that is unit free and can be used for comparison of two sets of data even if they are not measured in the same units.

- b) Index numbers measure the changes in the level of phenomenon in percentage form: In order to make a valid comparison of changes in some phenomenon over two periods or two places, it is convenient if one of them is assigned as 100. Hence, like other relative measures (such as relative measures of dispersion, relative efficiency, etc.), as a relative measure index number is also expressed in percentage form meaning that the value of the phenomenon at the period with which the change is compared is 100, that means the base year index is considered as 100 which makes the comparison straightforward.
- c) Index numbers measure the changes in a variety of phenomena that cannot be measured directly: According to Bowley "Index numbers are used to measure the changes in some quantity which can not be observed directly." For example, it is not possible to directly measure cost of living in quantitative terms, by index number we can only study the changes in it which may occur due to variation in the factors like rate of inflation, price, family members, quality of the items consumed, etc.
- d) Index numbers measure the effect of changes in relation to time or place: From the definition of index number it is clear that index number measures the relative change in some phenomenon over periods or between locations or in categories. For example, cost of living may be different at two places, over two periods or in two categories of people.

In the following few sections and subsections different types of price index numbers will be discussed.

13.7. Methods of Constructing Index Number

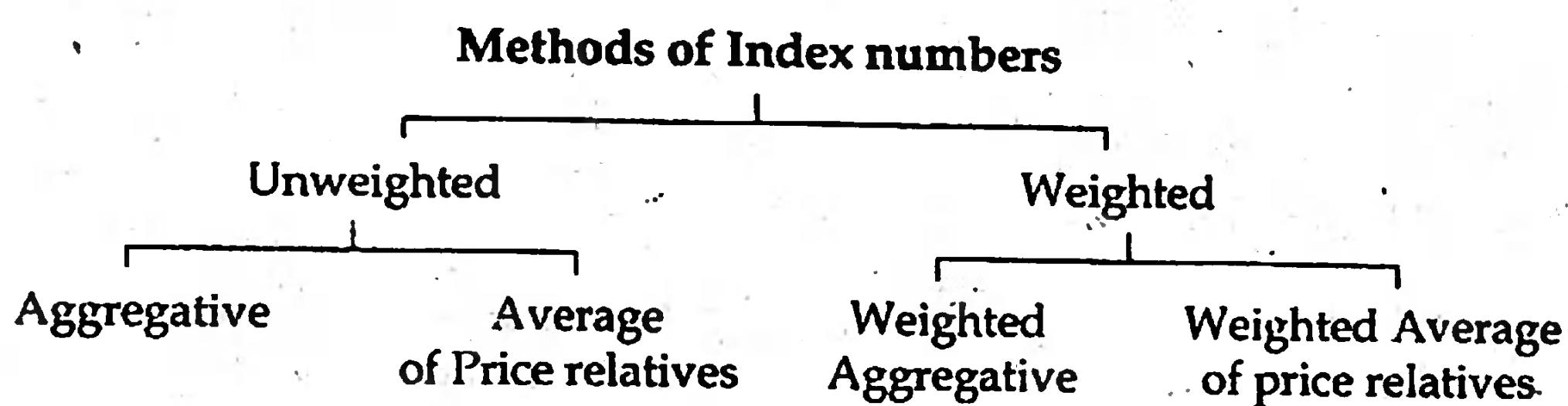
So far a number of formulae for the construction of index number have been developed. They can be classified under two major categories. They are

- a. Unweighted index numbers and
- b. Weighted index numbers

In case of unweighted index numbers, weights are not expressly assigned, whereas weights are accounted for various items in case of weighted index numbers. Each of these two types are further be divided again under two categories, such as-

- a. Simple aggregative index numbers and
- b. Simple average of price relatives

The following chart illustrates the various methods of index numbers



13.7.1. Unweighted Index numbers. Unweighted index numbers are those where index numbers are computed using only prices of the commodities, no importance is given to the respective weights. There are two approaches of construction of unweighted index numbers, viz. (i) Simple aggregative price index method and (ii) Simple Average of Price Relative Method.

Now we shall discuss different methods of constructing index numbers.. Before discussing different methods of constructing index numbers, we define the following:

- p_0 : Base year price per unit of a commodity;
- p_1 : Current year price per unit of a commodity;
- q_0 : Base year quantity of the commodity;
- q_1 : Current year quantity of the commodity;
- P_{01} : Price index of the current year 1 taking base year as 0;
- P_{10} : Price index of the base year 0 taking current year as 1;
- Q_{01} : Quantity index of the current year 1 taking base year as 0;
- Q_{10} : Quantity index of the base year 0 taking current year as 1;
- V_{01} : Value index of the current year 1 taking base year as 0.

Simple Aggregative Price Index Method

Under this method, the total prices for all commodities in the current year are divided by the total prices for those commodities in the base year and the quotient is multiplied by 100. The formula for constructing index number by simple aggregative method is

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100,$$

where $\sum p_1$ = Total price of various commodities for current year, and $\sum p_0$ = Total price of various commodities for base year.

Steps in the construction of simple aggregative price index

The method of constructing index number by simple aggregative method is very simple and the steps involved in this method are:

- i) Add the current year prices of various commodities and obtain $\sum p_1$
- ii) Add the base year prices of various commodities and obtain $\sum p_0$
- iii) Divide $\sum p_1$ by $\sum p_0$ and multiply the quotient by 100.

Example 13.7.1. The following data refer to the prices of some essential commodities for the year 2006 and 2007. Construct index number by simple aggregate method and comment.

Table 13.1. Prices of some essential commodities

Commodity and unit	Price (in Taka) of 2006	Price (in Taka) of 2007	Commodity and unit	Price (in Taka) of 2006	Price (in Taka) of 2007
Rice (per kg)	32	40	Milk (per litre)	35	44
Oil (per litre)	75	90	Meat (per kg)	200	240
Dal (per kg)	60	78	Vegetable (per kg)	12	15

Solution. The formula for price index by simple aggregative method is

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

Table 13.1. Calculation of simple aggregative price index number:

Commodity and unit	Price (in Taka) of 2006 (p_0)	Price (in Taka) of 2007 (p_1)
Rice (kg)	32	40
Oil (Litre)	75	90
Dal (kg)	60	78
Milk (Litre)	35	44
Meat (kg)	200	240
Vegetable (kg)	12	15
Total	$\sum p_0 = 414$	$\sum p_1 = 507$

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100 = \frac{507}{414} \times 100 = 122.46.$$

This means that as compared to 2006, there is net increase of $(122.46 - 100) = 22.46\%$ in 2007, in the prices of commodities included in the index. Or on an average prices of the items in 2007 is 122.46% of that in 2006.

Limitations of the method

It is obvious that this method is very easy to calculate, however, this method has two drawbacks. They are

- (i) Equal weight is given to all the items irrespective of their relative importance.

- (ii) The units in which prices of commodities are given affect the price index. For example, if in the above example the price of oil per bottle is given instead of per litre it will make a difference. Suppose the price per bottle of 5 litre in 2006 is taka 375 and in 2007 is taka 450.

The index would be $\frac{450}{375} \times 100 = 120$.

- Due to these limitations to construct this simple price index, sometimes an alternative method is used which is known as simple average of price relative method.

Average of Price Relatives Method

In this method price relatives are computed for various items included in the index and then average of the relatives is obtained commonly using arithmetic mean or geometric mean. For example, when arithmetic mean is used for averaging the relatives, the formula for computing the index is

$$P_{01} = \frac{\sum \left[\frac{P_1}{P_0} \times 100 \right]}{N} = \frac{\Sigma P}{N}; \text{ where } P = \frac{P_1}{P_0} \times 100.$$

Here N is the number of items whose price relatives are averaged.

Again, when geometric mean is used for averaging the price relatives, the formula for computing the index is

$$P_{01} = \text{Anti-log} \left(\frac{\sum \log \left[\frac{P_1}{P_0} \times 100 \right]}{N} \right) = \text{Anti-log} \frac{\sum \log P}{N}; \text{ where } P = \frac{P_1}{P_0} \times 100.$$

Steps in the construction of index numbers using simple average of price relatives method

The following are the steps involved in the construction of index number by simple average of price relatives method using arithmetic mean

- (i) Obtain the price relative by dividing the price of each commodity in the current year (p_1) by its price in the base year (p_0) and express this in percent, i.e., obtain $\frac{P_1}{P_0} \times 100$ for all items. Let P be value of this

percentage ratio, which means calculate $P = \left[\frac{P_1}{P_0} \times 100 \right]$.

- (ii) Add up all the values of P and divide by the number of items N.

The following are the steps involved in the construction of index number by simple average of price relatives method using geometric mean

- iii) Take the logarithm of all the P values obtained in step (i), that means, calculate $\log P$,
- iv) Add up all these logarithmic values, and obtain $\Sigma \log P$,
- v) Divide this sum by the number of items to obtain average of logarithmic values,
- vi) Finally, calculate the Antilog of average to obtain the required price index.

Example 13.7.2. The following data refer to the prices of some essential commodities for the year 2006 and 2007. Compute price index based on (i) Simple average of price relatives and (ii) Geometric mean of price relatives.

Commodity and unit	Price (in Taka) of 2006	Price (in Taka) of 2007	Commodity and unit	Price (in Taka) of 2006	Price (in Taka) of 2007
Rice(kg)	32	40	Milk (Litre)	35	44
Oil (Litre)	75	90	Meat (kg)	200	240
Dal (kg)	60	78	Vegetable(kg)	12	15

Solution. (i) Let us construct the following table for the computation of index based on simple average and geometric mean of price relatives.

Table 13.2. Computation of price index based on Simple average and Geometric mean of price relatives

Commodity and unit	Price (in Taka) of 2006 (p_0)	Price (in Taka) of 2007 (p_1)	$P = \frac{P_1}{P_0} \times 100$	Log P
Rice(kg)	32	40	125.00	2.10
Oil (Litre)	75	90	120.00	2.08
Dal (kg)	60	78	130.00	2.11
Milk (Litre)	35	44	125.71	2.10
Meat (kg)	200	240	120.00	2.08
Vegetable (kg)	12	15	125.00	2.10
Total			745.71	$\Sigma \log P = 12.57$

Price index based on simple average of relatives is

$$P_{01} = \frac{\sum \left[\frac{P_1}{P_0} \times 100 \right]}{N} = \frac{\Sigma P}{N} = \frac{745.71}{6} = 124.49.$$

(ii) Price index based on geometric mean of relatives

$$P_{01} = \text{Antilog} \frac{\sum \log P}{N} = \text{Antilog} \frac{12.57}{6} = \text{Anti-log } 2.09 = 124.24.$$

It is observed from the above calculations that the index numbers obtained by the two methods are almost same.

Merits and limitations of this method

Merits. This method has the following two advantages over the previous method:

- (i) Extreme items do not influence the index. Equal importance is given to all the items.
- (ii) The index is not influenced by the units in which prices are quoted or by the absolute level of individual prices. Relatives are pure numbers and are, therefore, independent of the original units.

Limitations. The method has the following two drawbacks:

- (i) Difficulty is faced with regard to the selection of an appropriate average.
- (ii) It is kind of unweighted index number, so main drawback of this method is that equal importance or weight is given to all items included in the index numbers that is not proper. As such, unweighted index numbers are of little use in practice.

13.7.2 Weighted Index Numbers. The main drawback of unweighted aggregative index is that it considers equal importance to all items included in the construction of index numbers. The weighted aggregative index number is such an index which is free from this limitation. For constructing price index, weights of the quantity of the commodities of the current year or the base year or their averages are usually taken as weights. Similarly, for the construction of quantity index, prices of the base year or current year or their average prices are taken as the weights.

Like unweighted index numbers, weighted index numbers are also of two types:

- (i) Weighted aggregative price index
- (ii) Weighted average of price relatives.

Weighted aggregative price Index

If w_j is the weight associated with the j th commodity then the general weighted aggregative price index is given by

$$P_{01} = \frac{\sum w_j p_{1j}}{\sum w_j p_{0j}} \times 100.$$

Based on the use of different types of weights, a number of formulae have been developed for the construction of price index. The following are some popularly used weighted aggregative methods that are named according to the person who have suggested them.

1. Laspeyre's method
2. Paasche's method
3. Dorbish and Bowley's method
4. Fisher's Ideal method
5. Marshall-Edgeworth's method
6. Kelly's method

In all the formulae index numbers are expressed as percentages. Each of these indices will convey information on the percentage change in price of all commodities in current year with respect to the price in the base period.

1. Laspeyre's price index. In this method, base year quantities are taken as weights. A German Economist Laspeyre suggest this formula in 1817. The formula for constructing the index number is

$$L P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

where, $L P_{01}$ = Price index by Laspeyre's Method

p_1 = Price in the current year

p_0 = Price in the base year

q_0 = Quantity in the base year

Steps in the construction of index number by Laspeyre's method

- a) Multiply current year price of each commodity by the respective base year quantity and add up these products for all commodities to get $\sum p_1 q_0$.
- b) Multiply base year price of each commodity by the respective base year quantity or weight, and add up these products for all commodities to get $\sum p_0 q_0$.
- c) Divide $\sum p_1 q_0$ by $\sum p_0 q_0$ and multiply the quotient by 100. This gives the required index by Laspeyre's method.

Laspeyre's price index is also sometimes known as the Consumer Price Index.

The Laspeyre's price index is easy to calculate. This is because the weightings do not change unless we change the base period. Besides this, the index makes it easier to compare one index with another over years. It is also possible to make meaningful comparisons of the change in prices and

buying power over time, since it considers only the change in prices per given number of units without changing the number of units.

This index is widely used in practice. The main drawback of this method is that it does not take into consideration the changes in the consumption pattern that take place with the passage of time. As a result, this index is expected to overestimate or to leave an upward bias in the index. Because, in course of time, when the prices of commodities increase, it is very likely to reduce the quantity consumed, in that case, if the base year quantities are used as the weights, over importance of the items is accounted. Instead of using current year weight with the current year prices in the numerator use of base year quantities would overestimate the actual change, hence the index is overestimated. For the same reason, in other words, it is also said that Laspeyre's price index overstates the impact of inflation.

Example 13.7.3: The following data refer to the prices of some essential commodities for the year 2006 and 2007. Compute price index of the year 2007 taking 2006 as base year by using Laspeyre's formula and comment.

Commodity and unit	2006		2007	
	Price (in Taka)	Quantity	Price (in Taka)	Quantity
Rice(kg)	32.00	32.0 kg	40.00	25.0 kg
Oil (liter)	75.00	4.0 liter	90.00	3.0 liter
Dal (kg)	60.00	2.0 kg	78.00	1.5 kg
Milk (liter)	35.00	20 liter	44.00	15.0 liter
Meat (kg)	200.00	10.0 kg	240.00	7.0 kg
Vegetable(kg)	12.00	22.0 kg	15.00	20.0 kg

Solution. The formula of the price index by Laspeyre's method is

$$L P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Table 13.3. Computation of index by Laspeyre's method

Commodity and unit	2006		2007		$P_0 q_0$	$P_1 q_0$
	Price P_0	Quantity q_0	Price p_1	Quantity q_1		
Rice(kg)	32.00	32.0	40.00	25.0	1024	1280
Oil (liter)	75.00	4.0	90.00	3.0	300	360
Dal (kg)	60.00	2.0	78.00	1.5	120	156
Milk (liter)	35.00	20	44.00	15.0	700	880
Meat (kg)	200.00	10.0	240.00	7.0	2000	2400
Vegetable(kg)	12.00	22.0	15.00	20.0	264	330
					$\Sigma P_0 q_0 = 4408$	$\Sigma P_1 q_0 = 5406$

Using the formula, we get, $L P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{5406}{4408} \times 100 = 122.64$.

This means that as compared to 2006, there is $(122.64 - 100) = 22.64\%$ increase in the average price of items in 2007. Or in average prices of the items in 2007 is 122.64% of that in 2006.

2. Paasche's Price Index. In this method, the current year quantities q_1 are taken as weight. A German Economist Paasche first used the formula in 1874. Price index by Paasche' method is given by:

$$P P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Where p_1 = prices in the current year

p_0 = prices in the base year

q_1 = quantities in the current year and

$P P_{01}$ = Price index by Paasche's method.

Steps in the construction of Paasche's Price Index

- Multiply current year price of each item under consideration by the respective current year quantity and add up these products for all items to get $\sum p_1 q_1$
- Multiply base year price of each commodity under consideration by the respective base year quantity and add up these products for all commodities to get $\sum p_0 q_1$
- Divide $\sum p_1 q_1$ by $\sum p_0 q_1$ and multiply the quotient by 100. It gives the required price index by Paasche's method.

Because its weightings are based on the current year, this index reflects combined changes in prices and consumption patterns well. However, this index is expected to underestimate or to show a downward bias in the index. Because, people tend to spend less on goods when their prices are rising, hence the use of the current year weighting produces an index which tends to underestimate the rise of prices, that means, it has a downward bias. For the same reason, unlike Laspeyre's price index, Paasche's price index understates the impact of inflation. Paasche's method is particularly helpful because it combines the effects of changes in price and consumption patterns. Thus, it is a better indicator of general changes in the economy than the Laspeyre's method.

Example 13.7.4. Compute price index of the year 2007 taking 2006 as base year by Paasche's method from the data used in example13.4.3.

Table 13.4. Computation for Paasche's index.

Commodity And unit	2006		2007		p_0q_1	p_1q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1		
Rice(kg)	32.00	32.0	40.00	25.0	800	1000
Oil (liter)	75.00	4.0	90.00	3.0	225	270
Dal (kg)	60.00	2.0	78.00	1.5	90	117
Milk (liter)	35.00	20	44.00	15.0	525	660
Meat (kg)	200.00	10.0	240.00	7.0	1400	1680
Vegetable(kg)	12.00	22.0	15.00	20.0	240	300
					$\sum p_0q_1$ = 3280	$\sum p_1q_1$ = 4027

Formula for price index by Paasche's method is

$$P_P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$\text{We get, } P_P_{01} = \frac{4027}{3280} \times 100 = 122.77.$$

This means that as compared to 2006, there is $(122.77 - 100) = 22.77\%$ increase in the average price of items in 2007 or in an average prices of the items in 2007 is 122.77% of that in 2006.

Differences between the Laspeyre's and Paasche's Indices.

The following are some differences between Laspeyre's and Paasche's indices:

- (i) In Laspeyre's method, base year quantities are taken as weights, whereas current year quantities are taken as weights in Paasche's method.
- (ii) In Laspeyre's method, importance is given to only the changes of prices of the commodities, whereas in paasche's method, importance is given to both the changes of prices and quantities of the commodities.
- (iii) Laspeyre's index tends to overestimate or to show an upward bias, whereas Paasche's index tends to underestimate or to show a downward bias.
- (iv) From a practical point of view, Lasperre's index is often preferred to Paasche's index for the simple reason that in Laspeyre's index weights are the base year quantities and do not change from one year to the next. On the other hand, the use of Paasche's index requires the continuous use of new quantity weights for each period considered and in most cases those weights are different and expensive to obtain.

Remarks. Laspeyre's index is preferred to Paasche's index from the practical point of view.

3. Price Index by Dorbish and Bowley Method. This method is the simple arithmetic mean of Laspeyre's and Paasche's indices. Dorbish and Bowley first used the method in 1871. The formula for this method is

$$\begin{aligned} {}_{DB}P_{01} &= \frac{1}{2} \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 + \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100 \right] \\ &= \frac{1}{2} \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1} \right] \times 100 = \frac{1}{2} ({}_{L}P_{01} + {}_{P}P_{01}) \end{aligned}$$

Where, p_1 = prices of the commodities in the current year

p_0 = prices in the base period

q_1 = quantities in the current period

q_0 = quantities in the base period

${}_{DB}P_{01}$ = Price index by Dorbish and Bowley method.

Since the price index by Dorbish and Bowley's method is the arithmetic mean of Laspeyres' and Paasche's index numbers, it always lies between the Laspeyres' and Paasche's indices.

Example 13.7.5. Compute price index by Dorbish and Bowley's method by using the data of example 13.4.4.

Solution. Price indices computed from example 13.4.3 and 13.4.4 are

${}_{L}P_{01} = 122.64$ and ${}_{P}P_{01} = 122.77$. Hence the price index by Dorbish and Bowley's is

$${}_{DB}P_{01} = \frac{1}{2} \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 + \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100 \right] = \frac{1}{2} [122.64 + 122.77] = 122.71.$$

Remarks. It is seen that this index lies between the indices of Laspeyre's and Paasche's.

4. Fisher Ideal Price Index. The Fisher's ideal price index is the geometric mean of Laspeyre's and Paasche's price indices. This formula for computing index number was first used by Fisher in 1920. The formula for constructing the index is

$${}_{F}P_{01} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}} \times 100 = [{}_{L}P_{01} \times {}_{P}P_{01}]^{\frac{1}{2}}$$

where, ${}_{F}P_{01}$ = Fisher's ideal price index

p_1 = prices of all goods and services in the current period

p_0 = prices in the base period

q_1 = quantities in the current period

q_0 = quantities in the base period.

This index also lies between Laspeyre's and Paasche's indices. It is smaller than the Dorbish and Bowley's index, since it is the geometric mean of Laspeyre's and Paasche's indices whereas Dorbish and Bowley's index is the arithmetic mean of the Laspeyre's and Paasche's indices.

Fisher's index is called Ideal index because of the following reasons:

- The formula is based on the geometric mean that is theoretically considered as the best measure of average for constructing index numbers.
- The formula takes into account prices and quantities of both current year as well as base year.
- The method is free from bias. The weight biases embodied in Laspeyre's and Paasche's methods are crossed geometrically and thus eliminated completely.
- The method satisfies both time reversal and factor reversal tests which justifies its superiority over other indices.

However, there are some shortcomings of this method as well. For example, calculation of index number is more tedious than other methods. Even if someone wants, sometimes it is not possible to use this method, because, both the current and base year quantities might not be available.

Steps in the construction of Fisher's Ideal Price Index

- Multiply current year price of each commodity under consideration by the respective base year quantity or weight, and add up these products for all commodities to get $\Sigma p_1 q_0$
- Multiply base year price of each commodity under consideration by the respective base year quantity or weight, and add up these products for all commodities to get $\Sigma p_0 q_0$
- Multiply current year price of each item under consideration by the respective current year quantity or weight, and add up these products for all items to get $\Sigma p_1 q_1$
- Multiply base year price of each commodity under consideration by the respective base year quantity or weight, and add up these products for all commodities to get $\Sigma p_0 q_1$
- Divide the value obtained in step (i) by that in step (ii),
- Divide the value obtained in step (iii) by that in step (iv)
- Multiply the quotient obtained in step (v) by that in step (vi) and take the square root of the result

- h) Multiply the result obtained in step (vii) by 100 to get the required index number by Fisher's method.

Example 13.7.6. Compute price index of 2007 taking 2006 as base year by Fisher's ideal method by using the data of Example 13.4.3. We make the following table for computation:

Table 13.6. Computation table for Fisher's ideal index.

Commodity and unit	2006		2007		P_0q_0	P_0q_1	P_1q_0	P_1q_1
	Price P_0	Quantity q_0	Price P_1	Quantity q_1				
Rice(kg)	32.00	32.0	40.00	25.0	1024	800	1280	1000
Oil(liter)	75.00	4.0	90.00	3.0	300	225	360	270
Dal(kg)	60.00	2.0	78.00	1.5	120	90	156	117
Milk(liter)	35.00	20.0	44.00	15.0	700	525	880	660
Meat(kg)	200.00	10.0	240.00	7.0	2000	1400	2400	1680
Vegetable(kg)	12.00	22.0	15.00	20.0	264	240	330	300
					$\Sigma P_0q_0 = 4408$	$\Sigma P_0q_1 = 3280$	$\Sigma P_1q_0 = 5406$	$\Sigma P_1q_1 = 4027$

Putting the values in the formula of Fisher's index we have

$$F_{P_{01}} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}} \times 100 = \left[\frac{5406}{4408} \times \frac{4027}{3280} \right]^{\frac{1}{2}} \times 100 \\ = (1.2264 \times 1.2277)^{1/2} \times 100 = 122.71$$

Alternative method for finding Fisher's ideal index

We can also find Fisher's index by using the values of Laspeyr's and Paasche's indices. Fisher's ideal index is the geometric mean of the Laspeyr's and Paasche's indices. We have already found the values of Laspeyr's and Paasche's indices from the examples 13.4.3 and 13.4.4 as

$$L_{P_{01}} = 122.64 \quad P_{P_{01}} = 122.77$$

Hence the Fisher's ideal index is

$$\left[L_{P_{01}} \times P_{P_{01}} \right]^{\frac{1}{2}} = (122.64 \times 122.77)^{1/2} = 122.71$$

This means that as compared to 2006, there is $(122.71 - 100) = 22.71\%$ increase in the average price of items in 2007 or in an average prices of the items in 2007 is 122.71% of that in 2006.

It is observed that Fisher's ideal index lies between Laspeyrs' and Paasche's indices and it is also less than Dorbish and Bowley's index since it is the

geometric mean of the Laspeyrs' and Paasche's indices whereas Dorbish and Bowley's index is the arithmetic mean of the Laspeyrs' and Paasche's indices.

5. Marshall-Edgeworth price index. In this method the arithmetic mean of base year and current year quantities are taken as weight to calculate the index number. The formula for Marshall-Edgeworth index is

$$ME P_{01} = \frac{\Sigma p_1(q_0 + q_1)/2}{\Sigma p_0(q_0 + q_1)/2} \times 100 = \frac{\Sigma p_1(q_0 + q_1)}{\Sigma p_0(q_0 + q_1)} \times 100 = \frac{\Sigma p_1q_0 + p_1q_1}{\Sigma p_0q_0 + p_0q_1} \times 100;$$

where, $ME P_{01}$ = Price index by the method of Marshall-Edgeworth

p_1 = prices in the current period,

p_0 = prices in the base period,

q_1 = quantities in the current period,

q_0 = quantities in the base period.

The advantage of this formula is that it considers the base year and current year quantities as weights and the disadvantage is that it needs current year quantities every time whenever it is constructed.

Steps in the construction of Marshall-Edgeworth Price Index

- i) Multiply current year price of each commodity under consideration by the respective base year quantity and add up these products for all commodities to get Σp_1q_0 .
- ii) Multiply current year price of each item under consideration by the respective current year quantity and add up these products for all items to get Σp_1q_1 .
- iii) Multiply base year price of each commodity under consideration by the respective base year quantity and add up these products for all commodities to get Σp_0q_0 .
- iv) Multiply base year price of each commodity under consideration by the respective base year quantity and add up these products for all commodities to get Σp_0q_1 .
- v) Add Σp_1q_0 and Σp_1q_1 .
- vi) Add Σp_0q_0 and Σp_0q_1 .
- vii) Divide the result obtained in step (v) by the result in (vi) and multiply the quotient by 100. It gives the price index by the Marshall-Edgeworth method.

For illustration of this method, let us again consider the data used for previous methods. Here, the table to be constructed is the same as that for Fisher's index. So, using the values of different sum of products from the table 13.6. used in Fisher's index, we get,

$$M_E P_{01} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100 = \frac{5406 + 4027}{4408 + 3280} \times 100 = 122.70$$

Remark. It is a simple, readily constructed measure, giving a very close approximation to the results obtained by the Fisher's ideal index. The index always lies between Laspeyrs' and Paasche's indices.

6. Kelly's Method. In this method; neither base year nor current year quantities are taken as weights. Instead, the quantities of some reference year or the average quantity of two or more years may be taken as weights.. The formula for this index is

$$K P_{01} = \frac{\sum p_1 q}{\sum p_0 q} \times 100 ; \text{ where } q = \text{fixed weight.}$$

This method is also known as the fixed weight aggregative index and is currently in great favour in the construction of index number series.

The important advantage of this method is that it does not need yearly changes in the weights. Weights of any representative period other than base period or current period can be selected as the weights. Selection of such weight can improve the accuracy of the index. This weight may be kept fixed until new data are available in order to revise the index.

The drawback of this method is that it does not give any importance to the weight either base period or current period.

13.7.3. Weighted Average of Price Relatives Method. In a previous section the unweighted average of price relative method is discussed where relative changes in the prices of commodities in current period with respect to the base period are averaged using different types of averages. The relative weights of the items were not considered. We can also compute weighted average of price relative by using the quantity consumed as weight. The formula for constructing the weighted average of price relatives index using base is given by

$$P_{01} = \frac{\sum \left[\frac{p_1}{p_0} \times 100 \right] (p_0 q_0)}{\sum p_0 q_0} = \frac{\sum PV}{\sum V} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

where, V = value weights ($= p_0 q_0$) = base period prices and quantities that determines total values in base period $P = \left[\frac{p_1}{p_0} \times 100 \right]$ = price relative.

This formula is equivalent to Laspeyr's method for any given data.

Steps in the construction of weighted average of price relatives index

The following steps are followed while computing the weighted average of price relatives index:

- i) Obtain percentage price relatives for each item under consideration, that means, obtain $\frac{P_1}{P_0} \times 100$.
- ii) Multiply price of each item in base year by the corresponding weight, and obtain the value weight $p_0q_0 = V$ for every item.
- iii) Multiply the percentages obtained in step (i) for each item by the corresponding value weight (p_0q_0) obtained in step (ii) assigned to that item.
- iv) Add up the results obtained from several multiplication carried out in step (iii).
- v) Add up the products obtained in step (ii) to obtain sum of the value weights, that means $\sum p_0q_0 = \sum V$.
- vi) Divide the sum obtained in step (iv) by the sum of the weights obtained in step (v). The result is the required index number.

On the other hand, if we wish to compute a weighted average of price relative using $V = p_0q_1$, then the above formula becomes,

$$P_{01} = \frac{\sum \left[\frac{P_1}{P_0} \times 100 \right] (p_0q_1)}{\sum p_0q_1} = \frac{\sum PV}{\sum V} = \frac{\sum P_1 q_1}{\sum p_0 q_1} \times 100.$$

This formula is equivalent to Paasche's method for any given data.

However, instead of using arithmetic mean, the geometric mean may also be used for averaging relatives. The weighted geometric mean of relatives is computed in the same manner as the unweighted geometric mean of relatives index number except that weights are introduced by applying them to the logarithms of the relatives. When this method is used the formula for computing the logarithm of index becomes,

$$\log(P_{01}) = \frac{\sum P \times \log V}{\sum V}$$

where, V = value weights i.e., p_0q_0 for each item and $P = \frac{P_1}{P_0} \times 100$.

Advantages of Weighted Average of Price Relative Method

The following are few advantages of weighted average of price relative indexes as compared to weighted aggregative price index:

- i) Different index numbers constructed using average price relative with the same base can be combined to form a new index.
- ii) Weighted average of price relative method is suitable to construct an index by selecting one item from each of the many subgroups of items. In such case, the values of each subgroup may be used as weights.
- iii) In course of time, if it is necessary to incorporate a new commodity in the formula instead of an old formerly used one, the relative for the new item may be spliced to the relative for the old one, using the former value weights.
- iv) The price or quantity relatives for each single item in the aggregate are, in effect, themselves a simple index that often gives valuable information for analysis.

13.7.4. Quantity or volume index numbers. Price indices measure changes in the price level of certain commodities. On the other hand quantity or volume index numbers measure the changes in the physical volume of goods produced, distributed or consumed. These indices are important indicators of the level of output in the economy or parts of it.

In constructing quantity index numbers, the problems facing the statistician are similar to the ones faced by him in constructing price indices. In this case we measure changes in quantities, and when we weigh, we use price as weights.

The quantity indices can be obtained easily by replacing p by q in the various formulae discussed earlier. Thus the quantity indices by different methods are:

Quantity index by the simple aggregative method is

$$Q_{01} = \frac{\sum q_1}{\sum q_0} \times 100,$$

where, $\sum q_0$ = sum of quantities in the base year 0 and

$\sum q_1$ = sum of the quantities of commodities in current year 1.

The method of construction of this index is very simple and the steps involved in this method are:

- i) Add the current year quantities of various items and obtain $\sum q_1$
- ii) Add the base year quantities of various items and obtain $\sum q_0$
- iii) Divide $\sum q_1$ by $\sum q_0$ and multiply the quotient by 100.

Important weighted aggregative quantity indices are:

Laspeyre's quantity index :

$$L Q_{01} = \frac{\sum P_0 q_1}{\sum P_0 q_0} \times 100$$

Paasche's quantity index :

$$PQ_{01} = \frac{\sum p_1 q_1}{\sum p_1 q_0} \times 100$$

Dorbish and Bowley's index:

$$DBQ_{01} = \frac{LQ_{01} + PQ_{01}}{2}$$

Fisher's quantity index:

$$FQ_{01} = \sqrt{LQ_{01} \times PQ_{01}}$$

Marshall-Edgeworth quantity index:

$$MEQ_{01} = \frac{\sum q_1(p_0 + p_1)}{\sum q_0(p_0 + p_1)} \times 100$$

Kelly's quantity index:

$$KQ_{01} = \frac{\sum q_1 p}{\sum q_0 p} \times 100$$

Example 13.7.7. The following data refer to the quantities of some essential commodities for the year 2006 and 2007. Construct simple aggregative quantity index and comment.

Commodity and unit	Quantities of 2006	Quantities of 2007	Commodity and unit	Quantities of 2006	Quantities of 2007
Rice(kg)	32	25	Milk (litre)	20	15
Oil (litre)	4	3	Meat (kg)	10	7
Dal (kg)	2	1.5	Vegetable(kg)	22	20

Solution. We know the formula for quantity index by simple aggregative method is

$$Q_{01} = \frac{\sum q_1}{\sum q_0} \times 100$$

The following table is constructed for necessary calculations:

Table 14.7. Construction of simple aggregate quantity index.

Commodity and unit	Quantity] of 2006 q ₀	Quantity of 2007 q ₁
Rice(kg)	32.0	25.0
Oil (litre)	4.0	3.0
Dal (kg)	2.0	1.5
Milk (litre)	20	15.0
Meat (kg)	10.0	7.0
Vegetable(kg)	22.0	20.0
Total	90.00	71.50

The simple aggregative quantity index is

$$Q_{01} = \frac{71.5}{90} \times 100 = 79.45$$

Interpretation. The quantity of the basket of commodities under consideration has been decreased by $(100 - 79.45)\% = 20.55\%$ in 2007 in compare to 2006. This happens due to the increase of prices of commodities in 2007 compare to 2006.

Example 13.7.8. Determine the quantity index numbers for 2003 taking 2002 as base year from the following data by using (i) Laspeyre's, (ii) Paasche's, (iii) Fisher's and (iv) Marshall-Edgeworth methods and interpret the results.

Year	Article I		Article II		Article III		Article IV	
	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity
2002	5.00	5	7.75	6	9.60	4	12.50	9
2003	6.50	4	8.80	5.5	9.75	4	12.75	9

Solution. The steps involved in the calculation of quantity index numbers are analogous to those in price index numbers. Let us construct the following table for necessary calculations:

Table 13.8. Computation of quantity index numbers.

Article	2002		2003		$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
	P_0	q_0	P_1	q_1				
I	5.00	5	6.50	4	5	5	6.5	4
II	7.75	6	8.80	5.5	7.75	6	8.8	5.5
III	9.60	4	9.75	4	9.6	4	9.75	4
IV	12.50	9	12.75	8.5	12.5	9	12.75	8.5
Total					239.05	222.4	221.775	207.275

(i) Laspeyre's quantity index is :

$$LQ_{01} = \frac{\sum P_0 q_1}{\sum P_0 q_0} \times 100 = \frac{207.275}{222.40} \times 100 = 93.20.$$

(ii) Paasche's quantity index is :

$$PQ_{01} = \frac{\sum P_1 q_1}{\sum P_1 q_0} \times 100 = \frac{221.775}{239.05} \times 100 = 92.77$$

(iii) Fisher's quantity index is :

$$FQ_{01} = \sqrt{LQ_{01} \times PQ_{01}} = \sqrt{93.20 \times 92.77} = 92.99$$

Interpretation. In all cases the average quantity demanded of the basket of items has been reduced in 2003 than in 2002. For example, Laspeyre's index shows that there is a $(100 - 93.20)\% = 6.80\%$ reduction in the consumption of the basket of commodities over year.

13.7.5. Value Index Numbers. Value means price time's quantity. Thus value index V is the sum of the values of a given year divided by the sum of the values for the base year. The formula, therefore is

$$V = \frac{\sum P_1 q_1}{\sum P_0 q_0} \times 100$$

Where, $\sum P_1 q_1$ = Total value of all commodities in the current year

And $\sum P_0 q_0$ = Total value of all commodities in the base year

However, if the values are given directly, then the value index number is

$$V = \frac{\sum V_1}{\sum V_0} \times 100$$

Since this type of index numbers take both the price and quantity in account, they need not be weighted. However, these index numbers are not popularly used because the situation meant by price and quantities are not fully meant by the values. A value index does not distinguish between the effects of its components, viz. price and quantity. The value index is not widely use.

13.8. Test of Accuracy of Index Number

Several formulae have been suggested for constructing index numbers. The problem is to select the most appropriate formula in a given situation. Professor Irving Fisher has suggested two tests for selecting an appropriate formula. They are

1. Time Reversal Test, and
2. Factor Reversal Test

1. Time Reversal Test. The test says that the index number of current year to the base year should be reciprocal of the index number of the base year to the current year. Suppose P_{01} is the price index of the current year 1 to the base year 0 and P_{10} is the index number of the base year 0 to the current

year 1 then time reversal text says that $P_{01} = \frac{1}{P_{10}}$. That is $P_{01} \times P_{10} = 1$.

The formula for calculating an index number should be such that it will give the same ratio between one period of comparison and the other, no matter which of the two is taken as base. Or putting in another way, the index number for forward should be the reciprocal of the backward, except for a constant of proportionality.

Thus, if the time script of any index formula is interchanged then the resulting index should be the reciprocal of the original index. Symbolically,

$$P_{01} = \frac{1}{P_{10}} \text{ or, } P_{01} \times P_{10} = 1.$$

If the product of these two indices is not unity, it is assumed that there is a time bias in the method. As it has been discussed before that there is an upward and downward biases in cases of Laspeyre's and Paasche's indices respectively due to selection of quantity of a single period as the weight; so these two formula do not satisfy the time reversal test (illustrated below). For example, index number by Laspeyre's method is

$$L P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \text{ (Omitting the percentage)}$$

By interchanging the base year with current year Laspeyre's index are

$$L P_{10} = \frac{\sum P_0 q_1}{\sum P_1 q_1}$$

$$\text{Hence, } L P_{01} \times L P_{10} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_0 q_1}{\sum P_1 q_1} \neq 1.$$

That means Laspeyre's index does not satisfy time reversal test.

For Paasche's index, we have,

$$P P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \text{ and } P P_{10} = \frac{\sum P_0 q_0}{\sum P_1 q_0}$$

$$\text{Hence, } P P_{01} \times P P_{10} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0} \neq 1,$$

That means Paasche's index does not satisfy time reversal test.

However, it can be proved that the time reversal test is satisfied by the index numbers constructed by the:

- i) Fisher's Ideal formula
- ii) Simple geometric mean of price relatives
- iii) Marshall-Edgeworth formula
- iv) Kelly's Method

For Fisher's ideal index, we have

$$F P_{01} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}} \text{ and } F P_{10} = \left[\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0} \right]^{\frac{1}{2}}$$

$$\text{Hence, } {}_F P_{01} \times {}_F P_{10} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}} \times \left[\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0} \right]^{\frac{1}{2}} = 1$$

That means time reversal test is satisfied by Fisher's index.

Similarly, it can be verified for other formulae.

2. Factor Reversal Test. The factor reversal test says that the product of price index with the quantity index should be equal to the corresponding value index. Symbolically, if P_{01} and Q_{01} are the price and quantity indices respectively, then

$$P_{01} \times Q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0} = V_{01} = \text{value index.}$$

If the product is not equal to the value ratio, then there is an error in one or both of the index numbers.

Like time reversal test, Laspeyre's and paasche's indices do not satisfy the factor reversal test. For example,

For Laspeyre's price index and quantity indices are given by

$${}_L P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \text{ and } {}_L Q_{01} = \frac{\sum P_0 q_1}{\sum P_0 q_0} \text{ respectively,}$$

$$\text{and value index is given by } V_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

$$\text{Thus, } {}_L P_{01} \times {}_L Q_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_0 q_1}{\sum P_0 q_0} \neq \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

That means, factor reversal test is not satisfied by the Laspeyre's index.

Again, Paasche's price index and quantity indices are given by,

$${}_P P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \text{ and } {}_P Q_{01} = \frac{\sum P_1 q_1}{\sum P_1 q_0} \text{ respectively,}$$

$$\text{thus, } {}_P P_{01} \times {}_P Q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_1 q_1}{\sum P_1 q_0} \neq \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

That means, factor reversal test is not satisfied by the Paasche's index.

For Fisher's Ideal price and quantity indices, we have

$${}_F P_{01} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}} \text{ and } {}_F Q_{01} = \left[\frac{\sum P_0 q_1}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_1 q_0} \right]^{\frac{1}{2}}$$

$${}_F P_{01} \times {}_F Q_{01} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}} \times \left[\frac{\sum P_0 q_1}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_1 q_0} \right]^{\frac{1}{2}} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

= Value index

That means, factor reversal test is also satisfied by the Fisher's ideal index.

Fisher's index is called Ideal index since it satisfies both the time and factor reversal tests.

Besides these two tests, some authors have suggested two other tests. These are

1. Unit test, and
2. Circular test.

1. Unit Test. According to unit test, the formula for constructing index numbers should be independent of the units in which prices and quantities are quoted. This test is satisfied only by simple aggregative index method.

2. Circular Test. This is another test for the index number. This test is based on shifting the base period and can be considered as an extension of the time reversal test. For instance, if there are three periods, this test implies that the multiplication of price indices constructed by shifting the base period in a circular manner is unity. If we consider three periods denoted by 0, 1 and 2. Then the circular test is given by $P_{01} \times P_{12} \times P_{20} = 1$.

This test is not met by most of the common methods used in the construction of index numbers or by any of the weighted index numbers. It is met by simple geometric mean of price relatives and the weighted aggregative fixed weights.

Example 13.8.1. Construct Fisher's index number by using following data and show that it satisfies time reversal and factor reversal tests

Commodity	2004		2005	
	Quantity	Price	Quantity	Price
I	20	12	30	14
II	13	14	15	20
III	12	10	20	15
IV	8	6	10	4
V	5	8	5	6

Solution. The following table is constructed for necessary calculation for the construction of Fisher's index number.

Table 13.9. Table for construction of Fisher's index numbers.

Commodity	2004		2005		P_1q_0	P_0q_0	P_1q_1	P_0q_1
	q_0	p_0	q_1	p_1				
I	20	12	30	14				
II	13	14	15	20				
III	12	10	20	15				
IV	8	6	10	4				
V	5	8	5	6				
Total					870	630	1090	782

We know, for Fisher's ideal index,

$$F P_{01} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}}, \quad F P_{10} = \left[\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0} \right]^{\frac{1}{2}}$$

and $F Q_{01} = \left[\frac{\sum P_0 q_1}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_1 q_0} \right]^{\frac{1}{2}}$

and time reversal test is satisfied if $F P_{01} \times F P_{10} = 1$

and factor reversal test is satisfied if

$$F P_{01} \times F Q_{01} = V_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

Now, substituting the values of sum of products from the table 13.9., we have

$$F P_{01} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \right]^{\frac{1}{2}} = \left[\frac{870}{630} \times \frac{1090}{782} \right]^{\frac{1}{2}} = 1.387392$$

$$F P_{10} = \left[\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0} \right]^{\frac{1}{2}} = \left[\frac{782}{1090} \times \frac{630}{870} \right]^{\frac{1}{2}} = 0.720777$$

$$F Q_{01} = \left[\frac{\sum P_0 q_1}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_1 q_0} \right]^{\frac{1}{2}} = \left[\frac{782}{630} \times \frac{1090}{870} \right]^{\frac{1}{2}} = 1.247058$$

and $V_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0} = \frac{1090}{630} = 1.730159$

Thus, $F P_{01} \times F P_{10} = 1.387392 \times 0.720777 = 1$

and $F P_{01} \times F Q_{01} = 1.387392 \times 1.247058 = 1.730159$

Hence, both time reversal and factor reversal tests are satisfied by Fisher's ideal index number (showed).

13.9. The Chain Index Numbers – Change in Base Period

The index numbers, which are discussed earlier, assumed that the base period remains the same throughout the series of the index. Usually the base period is the immediately preceding period of current year or some very recent year. It is impractical, even sometimes difficult to ascertain the accuracy of comparing the changes in some phenomenon in one period with a remote period which is far away from the period of comparison. In course of time, new items have to be included and old ones may have to be taken away for a more representative and accurate index. In that case, it is more realistic to construct an index using no fixed base, but using a series of indices constructed for each period using the immediately preceding period as base. For example, for the index of 2007, the base period would be 2006, for the index of 2006, the base period would be 2005, and for 2005 the base period would be 2004 and so on. Such index numbers are useful in comparing current period's relative change with respect to immediately preceding period. However, if it is desired to associate these relatives to a common base, the results may be chained. Such an index is called a chain index. In other words, the chain index is one in which the figures for each period are first expressed as percentage of the preceding period, then these percentages are linked or chained together by successive multiplication.

For example,

$$P_{02} = P_{01} \times P_{12},$$

$$P_{03} = (P_{01} \times P_{12}) \times P_{23} = (P_{02} \times P_{23})$$

$$P_{04} = \{(P_{01} \times P_{12}) \times P_{23}\} \times P_{34} = (P_{02} \times P_{23} \times P_{34}) = P_{03} \times P_{34}$$

Similarly, $P_{0n} = P_{0(n-1)} \times P_{(n-1)n}$

Here P_{01} is called the first link relative, P_{12} the second link relative and so on.

Steps in the construction of Chain Index. The following step are followed for the computation of chain index:

- i) Express the figure for each period as percentage of the preceding period. The percentages so obtained are called the link relatives.
- ii) Chain these link relatives together by successive multiplication to form a chain index. Chain index of any year is the average link relative of that period multiplied by the chain index of previous period divided by 100.

Hence, the formula for a chain index is given by

Chain index for current period =

$$\frac{\text{Link relative of current period} \times \text{Chain relative of preceding period}}{100}$$

The chain index number is useful for long-term comparison whereas link relatives are useful for comparison with the immediately preceding period.

Note that chain relatives differ from fixed relatives in computation. Chain relatives are computed from link relatives whereas fixed base relatives are directly computed from the original data. For instance,

$$\text{Link relative} = \frac{\text{Price relative for current period}}{\text{Price relative for previous period}} \times 100$$

and Price relative =

$$\frac{(\text{Current period's link relative}) \times (\text{previous period's price relative})}{100}$$

Conversion of Chain base index (CBI) to fixed base index (FBI).

$$\text{Current period FBI} = \frac{(\text{Current period CBI}) \times (\text{Previous period FBI})}{100}$$

Example 13.9.1. For the following data of wholesale prices of certain commodity construct the index number by the chain base method taking 1997 as base period using (i) link relatives and (ii) without using link relatives

Year	Price	Year	Price
1997	237	2001	260
1998	239	2002	270
1999	243	2003	277
2000	252	2004	295

Solution. (i) Computation of chain base index numbers using link relatives are shown in following table.

Table 13.10. Chain base indexes using link relatives.

Year	Price	Link Relatives	Chain base indexes (Base period 1997=100)
1997	237	100.00	100.00
1998	239	$\frac{239}{237} \times 100 = 100.84$	$\frac{100.84}{100} \times 100 = 100.84$
1999	243	$\frac{243}{239} \times 100 = 101.67$	$\frac{101.67}{100} \times 100.84 = 102.53$
2000	252	$\frac{252}{243} \times 100 = 103.70$	$\frac{103.70}{100} \times 102.53 = 106.33$

2001	260	$\frac{260}{252} \times 100 = 103.17$	$\frac{103.17}{100} \times 106.33 = 109.70$
2002	270	$\frac{270}{260} \times 100 = 103.85$	$\frac{103.85}{100} \times 109.70 = 113.92$
2003	277	$\frac{277}{270} \times 100 = 102.59$	$\frac{102.59}{100} \times 113.92 = 116.88$
2004	295	$\frac{295}{277} \times 100 = 106.59$	$\frac{106.59}{100} \times 116.88 = 124.47$

This means that from 1997 to 1998 there was a 0.84 percent increase in price, from 1997 to 1999 there was a 2.53 percent increase. In this way, if we are interested in finding out increase of average price from 1997 to 1998, from 1997 to 1999, from 1997 to 2000, etc. we have to compute the chain indexes.

The chain index for different periods as compared to 1997 can also be computed directly as the relative change of price of that period as compared to 1997. For example, chain index for 2004 is given by

$$\frac{P_{2004}}{P_{1997}} \times 100 = \frac{106.59}{100} \times 116.88 = 124.47$$

Index for other periods are computed in the same way.

The calculation of chain index without using link relatives are shown in the following table.

Year	Price	Chain index
1997	237	100.00
1998	239	$\frac{239}{237} \times 100 = 100.84$
1999	243	$\frac{243}{237} \times 100 = 102.53$
2000	252	$\frac{252}{237} \times 100 = 106.33$
2001	260	$\frac{260}{237} \times 100 = 109.70$
2002	270	$\frac{270}{237} \times 100 = 113.92$
2003	277	$\frac{277}{237} \times 100 = 116.88$
2004	295	$\frac{295}{237} \times 100 = 124.47$

Table 13.11. Chain base indexes without using link relatives

Year	Price	Chain base indexes (Base period 1997=100)	Year	Price	Chain base indexes (Base period 1997=100)

However, since concept of chain base index is based on the link relatives, it is suggested to compute this type of index using link relatives.

Example 13.9.2. Prepare fixed base index (FBI) numbers from the chain base indices (CBI) given below:

Year : 1998	1999	2000	2001	2002	2003	2004
CBI : 100.84	102.53	99.55	98.64	105.00	109.30	107.50

Solution. We know, FBI = $\frac{(\text{Current period CBI}) \times (\text{Previous period FBI})}{100}$

Computation of fixed base indices are shown in following table.

Table 13.12. Computation of the FBI number from CBI.

Year	Chain base index	Conversion	Fixed base index
1998	100.84		100.84
1999	102.53	$\frac{102.53 \times 100.84}{100}$	103.39
2000	99.55	$\frac{99.55 \times 103.39}{100}$	102.93
2001	98.64	$\frac{98.64 \times 102.93}{100}$	101.53
2002	105.00	$\frac{105.00 \times 101.53}{100}$	106.60
2003	109.30	$\frac{109.30 \times 106.60}{100}$	116.52
2004	107.50	$\frac{107.50 \times 116.52}{100}$	125.26

Example 13.9.3. Calculate chain base index number and fixed base index number from the following price figures.

Commodity	2002	2003	2004	2005
X	23	25	30	33
Y	55	62	68	75
Z	40	45	51	60

Solution. Computation of chain base index is shown in following table:

Table 13.13. Chain base index numbers.

Commodity	Link Relatives Based on Preceding Year			
	2002	2003	2004	2005
X	100	$\frac{25}{23} \times 100 = 108.70$	$\frac{30}{25} \times 100 = 120.00$	$\frac{33}{30} \times 100 = 110.00$
Y	100	$\frac{62}{55} \times 100 = 112.73$	$\frac{68}{62} \times 100 = 109.68$	$\frac{75}{68} \times 100 = 110.29$
Z	100	$\frac{45}{40} \times 100 = 112.50$	$\frac{51}{45} \times 100 = 113.33$	$\frac{60}{51} \times 100 = 117.65$
Total Link Relatives	300	333.92	343.01	337.94
Average	100	111.31	114.34	112.65
Chain Index	100	$\frac{111.31 \times 100}{100} = 111.31$	$\frac{114.34 \times 111.31}{100} = 127.27$	$\frac{112.65 \times 127.27}{100} = 143.36$

Computation of fixed base index is shown in following table:

Table 13.14. Fixed base index numbers

Commodity	Fixed Relatives (Base Period = 2002=100)			
	2002	2003	2004	2005
X	100	$\frac{25}{23} \times 100 = 108.70$	$\frac{30}{23} \times 100 = 130.43$	$\frac{33}{23} \times 100 = 143.48$
Y	100	$\frac{62}{55} \times 100 = 112.73$	$\frac{68}{55} \times 100 = 123.64$	$\frac{75}{55} \times 100 = 136.36$
Z	100	$\frac{45}{40} \times 100 = 112.50$	$\frac{51}{40} \times 100 = 127.50$	$\frac{60}{40} \times 100 = 150.00$
Total Link Relatives	300	333.92	381.57	429.28
Average (FBI))	100	111.31	127.19	143.28

Advantages and disadvantages of Chain base Method

Advantage. The following are a few advantages of chain base index:

- i) The chain base indices enable us to make comparison with the immediately previous period that is very useful in business decision-making.
- ii) The Chain base method permits to introduce new items and delete the existing one, which might be obsolete, without recalculation of entire series.

- iii) Weights can be adjusted as frequently as possible. This flexibility is of great significance in many types of index numbers.
- iv) The index numbers calculated by the chain base method are relatively free from cyclical and seasonal variations.

Disadvantage. The following are the disadvantages of chain base index:

- i) The chain base index is not useful for long-term comparisons in a time series.
- ii) The process of chaining link relatives is computationally difficult.

13.10. Base Shifting, Splicing and Deflating the Index Numbers

Base Shifting :

In course of time the index number constructed for a given base period may become, or it might be necessary to compare the constructed index number with some other base period. In that situation, index number of the given base period is to be changed and a new index is to be constructed based on a new base period, such an action is referred to the base shifting. The reasons of such a shift can state as follows:

- i) The base period is either too old or distant from the current period. Hence the constructed index has become almost useless for a meaningful comparison with the current period. Because, base period is to be chosen so that it is not far distant from the period of comparison. For example, if cost of living index of 2007 is compared with that of 1977, then such a comparison is useless, because base period is far away from the period comparison.
- ii) Comparison is to be made with another series of index numbers having different base. For example, suppose the cost of living index for a certain class of people is available with 2000 as base (i.e, 2000 = 100), but an interested researcher wants to compare the cost of living changes in that class of people with those of another class of people for which the corresponding index is given with the base year 2005. In such case, it is necessary to shift the base of the first series from 2000 to 2005.

However, when base is to be shift, one possibility is to reconstruct all index numbers using new base period which is sometimes almost impossible due to lack of data. In that case, the following formula for shifting base is suggested

$$\text{New index of any year} = \frac{\text{Index with old base period}}{\text{Index with new base period}} \times 100.$$

Example 13.10.1. The following table relates to the price index numbers taking 2000 as base period:

Year	Index	Year	Index
2000	100	2004	165
2001	110	2005	188
2002	130	2006	195
2003	147	2007	210

Shift the base from 2000 to 2005 and recomputed the index numbers.

Solution. Computation of new index with base period 2005 is shown in the following table:

Table 13.14. Computation of index with base period 2005

Year	Index (2000 = 100)	Index Numbers (2005 = 100)	Year	Index (2000 = 100)	Index Numbers (2005 = 100)
2000	100	$\frac{100}{188} \times 100 = 53.19$	2004	165	$\frac{165}{188} \times 100 = 87.77$
2001	110	$\frac{110}{188} \times 100 = 58.51$	2005	188	$\frac{188}{188} \times 100 = 100.00$
2002	130	$\frac{130}{188} \times 100 = 69.15$	2006	195	$\frac{195}{188} \times 100 = 103.72$
2003	147	$\frac{147}{188} \times 100 = 78.19$	2007	210	$\frac{210}{188} \times 100 = 111.70$

The new series with 2005 as base is obtained by dividing each entry of the old index by the index of 2005 (by 188) and multiplied by 100.

It should be noted that the above method of shifting the base will not necessarily coincide with the method in which the index number is constructed anew with the new base. However, since it is difficult to do otherwise in practice, this simple method is often employed regardless of whether a complete recalculation of index would produce the identical results.

Splicing

The task of combining two or more overlapping series of index numbers into one continuous series is called splicing. The need for splicing arises for securing continuity in comparison. It might happen often that an index is discontinuous due to its too old base period. A new index may be needed with the same items, but some recent year as base. In that case, we have to connect the new index number with that of one discontinuous the second number would be spliced to the first one with the result that the index

would enable comparison with the old base. The reverse is also possible. Suppose we have index numbers with a base of 1995 and another series of index numbers with a base of 2000, suppose both index numbers are continuing, then we can splice the index numbers with base 2000 into index numbers with base 1995 for all index numbers. It is also possible to splice the first index series to the second one and have a common index with base period 2000 for all indexes.

Two series of index numbers with different bases are spliced into a continuous series of indices to make it with a common base period using the following formula:

Forward Splicing Approach. This approach is used for splicing an old series of indices to make it continuous with a new series of indexes. The formula used for this purpose is:

Required index =

$$\frac{\text{Old index number with existing base}}{100} \times \text{Index number to be spliced}$$

Backward Splicing Approach. This approach is used for splicing a new series of indexes to make it continuous with an old series of indexes. The formula used for this purpose is:

Required index =

$$\frac{100}{\text{Old index number with existing base}} \times \text{Index number to be spliced}$$

The process of splicing is very simple and similar to that used in shifting. The process is illustrated below:

Example 13.10.2. The index numbers for price index A is calculated starting from 1995 and continued till 2000. Another price index B is also calculated starting from 2000 and continued till 2005. The index numbers are given below:

Year	Index A	Index B	Year	Index A	Index B
1995	100		2001		120
1996	110		2002		128
1997	130		2003		140
1998	147		2004		155
1999	165		2005		164
2000	188	100			

- Splice the index B to index A, so that a continuous series of index number from 1995 up to date are available with same base period 1995.

- ii) Also splice the index A to index B, so that a continuous series of index number from 1995 up to date are available with same base period 2000.

Solution. The splice of index B to index A with base 1995 and A to B with base period 2000 is shown in following table

Table 13.15. Index B spliced to index A and A spliced to B.

Year	Index A	Index B	Index B spliced to index A (1995=100)	Index A spliced to index B (2000=100)
1995	100		100	$\frac{188}{100} \times 100 = 188$
1996	110		110	$\frac{188}{100} \times 120 = 225.60$
1997	130		130	$\frac{188}{100} \times 128 = 240.64$
1998	147		147	$\frac{188}{100} \times 140 = 263.20$
1999	165		165	$\frac{188}{100} \times 155 = 291.40$
2000	188	100	$\frac{188}{100} \times 100 = 188$	$\frac{188}{100} \times 164 = 308.32$
2001		120	$\frac{188}{100} \times 120 = 225.60$	120
2002		128	$\frac{188}{100} \times 128 = 240.64$	128
2003		140	$\frac{188}{100} \times 140 = 263.20$	140
2004		155	$\frac{188}{100} \times 155 = 291.40$	155
2005		164	$\frac{188}{100} \times 164 = 308.32$	164

The spliced index in column 4 now refers to 1995 as base period and in column 5 refers to 2000 as base period.

Deflating

When prices of commodities rise, the money value or purchasing power of money declines. For example, if the money incomes of the consumers remain the same and prices are doubled in 2008 than that in 2005, then it is

assumed that the purchasing power of money reduced to half in 2008 in comparison of 2005, it means, if in 2005, a consumer's income was Taka 15000.00 per month and he could afford 20 kg of rice at the rate of Taka 22 per kg, that means he used to spend $22 \times 20 =$ Taka 440 for rice. Now if the price of rice rises to Taka 44 per kg in 2008 and his income and other expenditures remain the same, he could purchase only 10 kg of rice with the same amount. Thus the term Deflating refers to correcting or adjusting the money value at a time period after considering the changes in price levels which can be done by dividing money value by the appropriate price index of the same time period.

So, when the price rises, the money wages are deflated by the price index to the figure of real wages. The real wages enable us to see whether earner is better off or worse-off as a result of price change. Thus, real wage or income is determined by using the formula:

$$\text{Real Wage or Income} = \frac{\text{Money Wage}}{\text{Price Index}} \times 100$$

Here, the price index should be the consumer price index as it would reflect the change in purchasing power of the wage earner. Thus,

$$\begin{aligned}\text{Real Wage Index} &= \frac{\text{Real Wage of Current Period}}{\text{Real Wage of Base Period}} \times 100 \\ &= \frac{\text{Index of Money Wage}}{\text{Consumer Price Index}} \times 100\end{aligned}$$

Example 13.10.3. In support of the demand of salary adjustment, Government Officials working in a Government Office have supplied the following data

Year	2003	2004	2005	2006	2007
Pay	10,000	10500	11000	16000	16800
Price index	112.5	118.7	140.6	178.4	210.9

- i) Compute the real wages based on the given pay and price indexes
- ii) Compute the amount of pay they need in 2007 to provide buying power equal to that they enjoyed in 2003.

Solution. (i) The calculation of real wages based on the pay and price indexes are shown in following table.

Table 13.16. Computation of real wages.

Year	Pay	Price index	Real Wages = Col (2)/Col(3) × 100
2003	10000	112.5	$\frac{10000}{112.5} \times 100 = 8888.89$
2004	10500	118.7	$\frac{10500}{118.7} \times 100 = 8845.83$
2005	11000	140.6	$\frac{11000}{140.6} \times 100 = 7823.61$
2006	16000	178.4	$\frac{16000}{178.4} \times 100 = 8968.61$
2007	16800	210.9	$\frac{16800}{210.9} \times 100 = 7965.86$

That means the real pay the Officials are receiving is Taka 7965.86

(ii) In order that the Officials have the same buying power in 2007 as they had in 2003, their payment should be

$$\frac{10000}{112.5} \times 210.9 = 18746.67$$

Example 13.10.4. The following table gives the weekly wages (in Tk) of a worker and general price index number during 1996- 2005. Prepare the index number to show the changes in real wages of the worker.

Year	Weekly wages	Price index No.	Year	Weekly wages	Price index No.
1996	300	100	2001	500	150
1997	360	110	2002	540	175
1998	420	128	2003	560	210
1999	460	140	2004	600	230
2000	480	146	2005	630	250

Solution. Real weekly wage for 1997 is computed using the formula Real

$$\text{Wage or Income} = \frac{\text{Money Wage}}{\text{Price Index}} \times 100$$

and real wage index is computed using the formula

$$\text{Real Wage Index} = \frac{\text{Real Wage of Current Period}}{\text{Real Wage of Base Period}} \times 100$$

Index number showing the changes in the real weekly wages of the workers are computed in following table

Table 13.17. Real weekly wage index of worker.

Year	Weekly wages	Price index no.	Real weekly wages	Real weekly wage index numbers
1996	300	100	$\frac{300}{100} \times 100 = 300.00$	$\frac{300}{300} \times 100 = 100.00$
1997	360	110	$\frac{360}{110} \times 100 = 327.27$	$\frac{327.27}{300} \times 100 = 109.09$
1998	420	128	$\frac{420}{128} \times 100 = 328.13$	$\frac{328.13}{300} \times 100 = 109.38$
1999	460	140	$\frac{460}{140} \times 100 = 328.57$	109.52
2000	480	146	$\frac{480}{146} \times 100 = 328.77$	109.59
2001	500	150	$\frac{500}{150} \times 100 = 333.33$	111.11
2002	540	175	$\frac{300}{100} \times 100 = 308.57$	102.86
2003	560	210	$\frac{560}{210} \times 100 = 266.67$	88.89
2004	600	230	$\frac{600}{230} \times 100 = 260.87$	86.96
2005	630	250	$\frac{630}{250} \times 100 = 252.00$	84.00

13.11. Consumer Price Index Number (Cost of Living Index Number)

The consumer price index, also known as the cost of living index or retail price index, is constructed to study the effect of changes in the prices of a basket of goods and services on the purchasing power of a particular group of people during current period as compared with some base period. Change in the cost of living of an individual between two periods means the change in his money that will be needed for him to maintain the same standard of living in both periods. Thus the cost of living index numbers are intended to measure the average increase in the cost of maintaining the same standard in a given year as in the base year. Since the consumption habits of people differ widely from class to class (such as poor, low income, middle income, rich, etc.) and even within the same class from region to region, the changes in the level of prices affect different classes differently.

and consequently the general price index numbers usually fail to reflect the effects of changes in the general price level on the cost of living of different classes of people. Cost of living index numbers are, therefore, compiled to get a measure of the general price movement of the commodities consumed by different classes of people.

Since the factors like the size of family, its age composition, its income or occupation, the place or the region etc. are not taken into account while computing the cost of living index numbers, it should not be interpreted as a measure of standard of living.

The consumer price index numbers are constructed by the following two methods:

Aggregate Expenditure Method or Weighted Aggregate Method

This method is akin to the Laspeyre's method of constructing weighted index. In this method quantities of various commodities consumed by a particular class of people are assigned weights on the basis of quantities consumed in the base year. That means,

$$\text{Consumer Price Index} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100.$$

13.11.1. Uses of consumer price index number. Though index numbers are originally designed to study the general level of prices or purchasing power of money, today these are used extensively for a variety of purposes in economics, business management, etc. and for quantitative data relating to production, consumption, profits, etc. phenomena for comparing changes in two periods, places, etc. However, the main uses of index numbers are:

1. Index numbers are used as Economic Barometers;
2. Index numbers help in studying trends and tendencies of prices, sales, etc;
3. Index numbers help in formulating decisions and policies;
4. Consumer price index or cost of living index is particularly useful for purposes
 - i) Formulate economic policy, adjust income payments, and measure real earning
 - ii) Measure the purchasing power of money
 - iii) Wage negotiations and wage contracts.

13.11.2. Construction of CPI. The consumer price index is a weighted aggregate price index with fixed weights. The need for weighting arises because the relative importance of various commodities for different classes of people is not the same. The percentage of expenditure on different

commodities by an average family constitutes the individual weights assigned to the corresponding price relatives, and the percentage expenditure of some essential commodities like (i) food, (ii) clothing, (iii) fuel and lighting, (iv) house rent (v) miscellaneous.

The consumer price index is constructed by the following two methods:

Aggregate Expenditure Method or Weighted Aggregate Method:

This method is very similar to the Laspeyre's method of constructing weighted index. In this method, the quantities of various items consumed by a particular class of people are assigned weights on the basis of quantities consumed in the base period. Mathematically, it is stated as:

$$\text{Consumer Price Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 .$$

Where, p_1 and p_0 are the prices in current period and base period respectively, and q_0 's are the quantities consumed in base period.

Family Budget Method or Method of Weighted Average of Price Relatives

In this method the family budget of a large number of people, for whom the index is to be constructed, are cautiously studied, then aggregate expenditure of an average family on various items is estimated. These values constitute the weights. Mathematically, consumer price index in this method is

$$\text{Consumer Price Index} = \frac{\sum PV}{\sum V}$$

Where, P's are price relatives, that means $P = \frac{p_1}{p_0} \times 100$ and V is the value weight that means $V = p_0 q_0$

$$\text{or, if } V = \text{weight } W, \text{ then } CPI = \frac{\sum PW}{\sum W}$$

Example 13.11.5. Calculate the cost of living index form the following information

Item	Price		Weight
	Base year	Current year	
Food	39	47	4
Transport	8	12	1
Clothing	14	18	3
House rent	12	15	2
Misc.	25	30	1

Solution. Here the information about prices of commodities of base and current year along with common weights are given, so for the calculation of cost of living index number, we have to use the formula given by

$$\text{CLI} = \frac{\sum \text{PW}}{\sum W} \text{ where, } P = \frac{P_1}{P_0} \times 100.$$

The calculation of price relatives and PW are shown in following table:

Table 13.18. Calculations of price relatives and PW.

Item	P_0	P_1	W	$P = \frac{P_1}{P_0} \times 100$	PW
Food	39	47	4	120.51	482.04
Transport	8	12	1	150.00	150.00
Clothing	14	18	3	128.57	385.71
House rent	12	15	2	125.00	250.00
Misc.	25	30	1	120.00	120.00
Total			11		1387.75

Using the formula we have, $\text{CLI} = \frac{\sum \text{PW}}{\sum W} = \frac{1387.75}{11} = 126.16$.

Example 13.11.6. An enquiry into the budgets of middle class families in Chittagong gave the following information:

Item	Food	Rent	Clothing	Fuel and electricity	Transport and others
Expenses on	35%	15%	20%	10%	20%
Price (2008)	150	50	100	20	60
Price (2009)	174	60	125	25	90

What changes in the cost of living figure of 2009 have taken place as compared to 2008?

Solution. Here the information about prices of items of base and current year along with common weights (percentage of expense on items) are given, so for the calculation of cost of living index number, we have to use the formula given by $\text{CLI} = \frac{\sum \text{PW}}{\sum W}$ where, $P = \frac{P_1}{P_0} \times 100$.

The calculation of price relatives and PW are shown in following table.

Table 13.19. Calculations of price relatives and PW.

Item	p_0 (2008)	p_1 (2009)	W	$P = \frac{P_1}{P_0} \times 100$	PW
Food	150	174	35	116	4060
Rent	50	60	15	120	1800
Clothing	100	125	20	125	2500
Fuel and electricity	20	25	10	125	1250
Transport and others	60	90	20	150	3000
Total			100		12610

Using the formula we have, $CLI = \frac{\sum PW}{\sum W} = \frac{12610}{100} = 126.10$.

Hence, compared to 2008, the cost of living has gone up by $(126.10 - 100)\% = 26.10\%$ in 2009.

Example 13.11.7. An increase of 50% of cost of a certain consumption goods raises the cost of living of a certain family by 5%. What percentage of its cost of living was due to buying that article before the change in the price?

Solution. Let the cost of article before rise be X . So, after increase, it became $(X + 50X/100) = 1.5X$, hence the rise was $1.5X - X = 0.5X$ which is equal to an increase of 5% in the cost of living. Let the cost of living before increase was Y , so after increase of 5%, it became $(Y + 5Y/100) = 1.05Y$, or the increase was $1.05Y - Y = 0.05Y$. Thus, $0.5X = 0.05Y$ or $X = 0.1Y$, that means 10% of the cost of living was due to expenditure on that article was.

Example 13.11.8. During a certain period the cost of living index goes up from 110 to 200, and the salary of a worker is also raised from Taka 3250 to Taka 5000. Does the worker really gain, if so, by how much in real terms?

Solution. We the real wage = $\frac{\text{Actual wage}}{\text{Cost of living index}} \times 100$,

Given the real wage and cost of living index in base period are respectively Taka 3250 and 110

So, the real wage of Taka 3250 is $\frac{3250}{110} \times 100 = \text{Taka } 2954.54$

The real wage and cost of living index in current period are respectively Taka 5000 and 200.

So, the real wage of Taka 5000 is $\frac{5000}{2000} \times 100 = \text{Taka } 2500$ which is less than Taka 2954.54.

Since, the real wage of Taka 5000 in current period is less than that of Taka 3250 in base period, the worker does not really gain, rather the real wage of worker is decreased by Taka $(2954.54 - 2500) = \text{Taka } 454.54$.

Questions

1. What does an index number measure? Distinguish between price index, quantity index and value index. Also state two important uses of index numbers.
2. What do you mean by index number? Briefly discuss its applications in business.
3. How does simple aggregative index differ from weighted aggregative index? Mention some important formula of weighted aggregative index numbers along with their advantages and disadvantages.
4. What is an index number? Discuss the steps involved in the construction of an index number.
5. Distinguish between simple aggregative index and simple average of price relatives. Mention the steps of simple aggregative index.
6. Briefly discuss the importance and uses of index number in business.
7. How does Laspeyres's method differ from Paasche's method? Why Fisher's formula is called an ideal index number?
8. What do you mean by a good index number? What are the tests of index numbers? Discuss time reversal test and factor reversal test in context to the Laspeyres's and Paasches index numbers.
9. What is Fisher's ideal index? Why is it so called? Show that it satisfies the time reversal and factor reversal tests.
10. How does the Marshall-Edgeworth formula differ from Fisher's formula? What is Kelly's formula? 'Fisher's formula is better than Marshall Edgeworth formula' – Justify.
11. What are the criteria of a good index number? Describe the tests, which should be satisfied by a good index number.
12. What are the factor reversal test and time reversal test? Show that Laspeyres's and Paasche's formula do not satisfy these tests.
13. What do you mean by circular test? How does it differ from time reversal test? Verify whether Laspeyres's formula, Paasche's formula or Fisher's formula satisfy it.
14. Describe the problems faced in the construction of an index number. Differentiate between unweighted and weighted index numbers. Also distinguish between price index numbers and quantity index numbers with examples.

15. What do you mean by weighted aggregative index numbers? Define all weighted index numbers known to you.
16. How does price index number differ from quantity index number? Write down the formula for Laspeyre's and Paasches' quantity index numbers.
17. Distinguish between fixed base and chain base index numbers. Also point out their relative merits and demerits.
18. Define chain index number and state its advantage and disadvantages. Also discuss the steps in construction of chain index number.
19. Explain the concepts of base shifting, splicing and deflating. How does deflating is used in business?
20. What do you mean by consumers price index? Discuss the steps involved in the construction of it. Also mention the uses of consumer's price index.
21. Describe the main problems for the construction of index number.
22. What do you mean by cost of living index? The cost of living index of School teachers of Chittagong city in June 2011 is found as 190 as compared to December 2010 – comment on the statement.
23. Discuss various steps involved in the construction of cost of living index for garments workers in Chittagong City.
24. Mention the points to be remembered in selection of a base period during construction of cost of living index. Hence, discuss how you can construct the cost of living index for small businessmen in Chittagong city.

Exercises

25. The following are the prices of six different commodities for 2004 and 2005. Compute price index for the year 2005 considering 2004 as base year by using (i) simple aggregative method and (ii) average of price relatives method.

Commodity	Price in 2004 (in Taka)	Price in 2005 (in Taka)
A	40	50
B	60	60
C	20	30
D	50	70
E	80	90
F	100	110

Ans. i) 117.14, ii) using AM 122.92, using GM 121.17

26. From the data given below, compute price index of 2007 taking 2006 as base year using (i) simple aggregative method and (ii) average price relatives method.

Commodity	Price in 2006 (in Taka)	Price in 2007 (in Taka)
A	120	150
B	165	172
C	120	135
D	150	165
E	180	194

Ans. (i) 11.02, (ii) using AM 111.90, using GM 110.95

27. The following table gives the weekly wages (in Tk) of a worker and general price index number during 2001- 2010. Prepare the index number to show the changes in real wages of the worker.

Year	Weekly wages	Price index No.	Year	Weekly wages	Price index No.
2001	500	100	2006	650	220
2002	540	120	2007	670	245
2003	560	150	2008	700	260
2004	600	175	2009	720	300
2005	630	210	2010	780	310

Ans. For 2002 = 90, for 2001 = 74.67, and so on

28. For the data given below, calculate price and quantity index of the year 2006 taking 2006 as base year year using the formula given by Laspeyer's, Paasche's, Fisher's, Marshall-Edgeworth and Dorbish, and comment. Also verify whether time reversal test and factor reversal tests are satisfied by the Laspeyer's, Paasche's, and Fisher's index numbers.

Item	For 2006		For 2007	
	Price	Quantity	Price	Quantity
Bread	50	32	60	30
Meat	25	15	30	14
Tea	15	2	17	2
Jelly	40	3	46	2
Misc.	15	2	21	2

29. Compute price index numbers from the following data using (i) Laspeyer's formula, (ii) Paasche's formula and Fisher's formula and comment

Item	2006		2007	
	Quantity	Price	Quantity	Price
Rice	12	10	15	12
Wheat	15	7	20	5
Fish	24	5	20	9
Meat	5	16	5	14

Ans: Laspeyre's 118.82, Paasche's 112.77, Fisher's 115.76

30. Calculate Laspeyre's and Paasche's price and quantity indices for the following data

Commodity	2009		2010	
	Price	Quantity	Price	Quantity
A	4	10	5	12
B	6	8	7	10
C	10	5	12	4
D	3	12	4	15
E	5	7	5	8

Ans. Price indices are 119.14, 119.31, quantity indices are 111.48, 111.65

31. Compute suitable index number using the following data and comment.

Commodity	Price		Base Year quantity
	Base Year	Current year	
A	16	20	50
B	11	12	100
C	14	19	60
D	20	25	30

32. Compute suitable index number using the following data and comment.

Commodity	Price		Current Year quantity
	Base Year	Current year	
A	16	20	30
B	11	12	70
C	14	19	45
D	20	25	25

33. Compute Laspeyer's, Paasche's, Fisher's, Marshall-Edgeworth and Dorbish index number using the following data and comment.

Commodity	Price		Quantity	
	Base Year	Current year	Base Year	Current year
A	6	10	50	56
B	2	2	100	120
C	4	6	60	60
D	10	12	30	64

34. Calculate the cost of living index from the following data.

Item	Price		Weight
	Base year	Current year	
Food	150	160	7
Clothing	25	30	3
House rent	50	54	1
Transport	40	46	2
Misc.	15	21	2

35. Using following data, show that Fisher's Ideal index number satisfies the time reversal and factor reversal tests.

Commodity	Price per unit		Number of units	
	Base year	Current year	Base year	Current year
A	6	10	50	56
B	2	2	100	120
C	4	6	60	60
D	10	12	30	24
E	8	12	40	36

36. The following data gives the average wholesale price of five groups of commodities for the years 2003 to 2007; compute the chain base index numbers and comment.

Commodity	2003	2004	2005	2006	2007
A	2	3	4	2	7
B	3	6	9	4	3
C	4	12	20	8	22
D	5	7	22	16	18
E	3	8	11	14	12

37. The following table gives the per capita income and cost of living index of a particular community. Calculate the real income taking into account the rise in the cost of living.

Year	Cost of living index base 1983	2004
1983	100	360
1984	104	400
1985	115	480
1986	160	520
1987	210	550
1988	260	590
1989	300	610
1990	320	650

38. Prepare the fixed base index numbers from the following chain base index numbers given below:

Year:	1999	2000	2001	2002	2003	2004
Chain Index:	94	104	104	93	103	102

39. From the following chain base index numbers construct fixed base index numbers

Year:	1996	1997	1998	1999	2000	2001	2002
Chain Index:	100	105	105	105	96	94	95

40. Calculate the price index number for the year 2010 with 2005 as base year from the following data by using i) Laspeyre's, ii) Paasche's, iii) Fisher's, iv) Marshall-Edgeworth and v) Dorbish formula and comment.

Commodity	2005		2010	
	Unit price	Money value	Quantity	Money value
A	10	50	4.5	60
B	35	350	9	400
C	100	350	3.5	375
D	84	672	6.5	750

Also calculate the corresponding quantity index numbers and comment. Show that Fisher's Ideal formula satisfies the time reversal and factor reversal tests.

(Here money value means the total value of the commodity)

41. Calculate the quantity index number for the year 2010 with 2005 as base year from the following data by using i) Laspeyre's, ii) Paasche's, iii) Fisher's formula, iv) Marshall-Edgeworth and v) Dorbish formula and comment. Also calculate the price index using Fisher's formula and show that Fisher's Ideal formula satisfies the time reversal and factor reversal tests.

	Commodities			
Quantity in kg	A	B	C	D
In 2005	8	10	15	20
In 2010	6	7	10	15
Price per kg				
In 2005	20	50	40	20
In 2010	40	60	50	20

42. Calculate the price index number for the year 2007 with 2006 as base year from the following data by using i) Laspeyre's, ii) Paasche's and iii) Fisher's formula. Also calculate the price index using Fisher's formula and show that Fisher's Ideal formula satisfies the time reversal and factor reversal tests.

	Commodities			
Price per kg	A	B	C	D
In 2006	118	110	75	220
In 2007	126	125	90	250
Quantity in kg				
In 2006	18	10	15	20
In 2007	16	8	12	17

43. From the chain base index given below prepare fixed base index numbers

Year: 2000 2001 2002 2003 2004

Chain base Index numbers: 80 140 130 110 90

Ans. 80, 112, 145.6, 160.16, 144.10

44. From the following data prepare a whole sale price index for the year 2005

Commodity	Price in 2004	Price in 2005	Weights
A	20	25	22
B	18	30	14
C	11	12	7
D	15	20	8

Applications

45. A certain publishing company began its business of publishing college textbooks in 2004. It is interested in determining how its sales have changed compared to its first year. A summary of the company's record shows how many new books it published in each year in different areas:

Area	Price in 2004	Price in 2005	Price in 2006
Science	25	28	35
Arts	30	30	40
Commerce	112	130	150
Engineering	20	25	30

Using 2004 as the base year, calculate the un-weighted aggregative quantity index for 2005 and 2006. Also interpret the results.

46. The information on wages of workers of a large company are given below:

Category of workers	Wage per day			
	2004	2005	2006	2007
Class A	50	56	60	80
Class B	65	70	80	100
Class C	90	100	120	150
Class D	100	115	130	200

Using 2004 as base period, calculate the un-weighted aggregate wage index for rest of the periods and comment.

47. The administrator of a private hospital has compiled the information regarding food needed for their patients. Average prices of the commodities for different years are given below:

Commodity	Prices per kg			
	2005	2006	2007	2008
Rice	30	32	38	43
Vegetable	11	14	18	25
Fish	90	100	120	150
Meat	250	270	300	325

Express the prices in 2006-2008 in terms of un-weighted aggregate index and comment.

48. In a study of group health insurance policies, commissioned by a reputed company, the following sample of average individual rates was collected. Using 2004 as the base period, calculate average of relative price index for each year and comment.

Insurance group	2004	2005	2006	2007
Teachers	30	32	38	43
Govt. employee	11	14	18	25
Students	90	100	120	150
Engineers	250	270	300	325

49. In order to know about the change in the prices of rooms in a new hotel, the following information were recorded:

Category of rooms	Prices per room per night			Frequency of rooms rented in 2006
	2006	2007	2008	
Ordinary	130	140	160	143
Luxury	300	400	500	125
Suit	490	550	700	150
common	100	110	130	325

Help the company determine the relative change in prices of different years using 2006 as the base period.

50. After hearing many complaints that the prices of the fruits of a particular wholesale fruit shop during summer, the shop decided to see whether this is true. Based on the following data, help the shop keeper to calculate the appropriate aggregate price indices for each month.

Fruit	Price per kg			Amount sold in June
	June	July	August	
Apples	100	110	130	70
Oranges	90	105	120	75
Watermelon	50	60	65	100
Papaya	30	35	50	60
Grapes	200	220	300	25

(Hints: since only base year prices are given, you have to use Laspeyre's formula)

51. A manufacturing company has collected the following production information about the company's principal products. Calculate weighted aggregative quantity index for the years using price of 2005 as weight.

Products	Quantities produced			Cost of production/unit in 2005
	2005	2006	2007	
A	592	600	550	145
B	456	444	478	220
C	72	90	102	340

52. A veterinarian has noticed that he has treated a number of pets this past winter. He wonders whether this number was spread across the 3 winter months evenly or whether he treated more pets in any certain month. Using December as the base period, calculate the weighted average of relative quantity index for January and February.

Pets	Number treated			Price per visit Average for three months
	December	January	February	
Cats	100	200	95	Tk. 60
Dogs	125	75	200	Tk. 70
Parrots	15	20	15	Tk. 90
Snakes	10	5	5	Tk. 120

53. A survey by the milk products association produced the following information. Construct Laspeyres index with 2004 as the base period and comment.

Products	Average price per kg		Total quantity (in lacs) 2004
	2004	2007	
Sweet	120	220	2.7
Liquid milk	40	48	13.1
Butter	395	530	7.5

54. A garment manufacturing industry has examined the pricing trends of clothing items for a client. The following table contains the results of the survey (shown in unit price).

Product	Prices per unit			
	2005	2006	2007	2008
Jeans	130	150	230	43
Jackets	210	260	320	25
Shirts	90	100	120	150

Calculate unweighted average of relative index for each year using 2005 as the base period.

55. Construct quantity index numbers for 2008 considering 2007 as base year using Laspeyre's formula, ii) Paasche's formula, iii) Fisher's formula, iv) Marshall-Edgeworth and v) Dorbish formula and comment.

Year	Article							
	I		II		III		IV	
	P	Q	P	Q	P	Q	P	Q
2007	192	19	35	20	100	10	225	17
2008	222	17.5	22	16	100	8.5	200	15

56. Prepare price index numbers for 2006 with 2003 as base year from the following data by using, i) Laspeyre's formula, ii) Paasche's formula, iii) Fisher's formula, iv) Marshall-Edgeworth and v) Dorbish formula and comment. Show that Fisher's Ideal index satisfies the time reversal and factor reversal tests

Year	Article							
	I		II		III		IV	
	P	Q	P	Q	P	Q	P	Q
2003	112	9	15	8	78	10	125	6
2006	188	7.5	22	6	100	8.5	200	5

57. A long distance phone company has collected price and sales volume data for phone calls from Bangladesh to USA. The data for three rate schedules are as follows:

Period	Price per call 2000	Price per call 2006	Total number of calls in 2000 (in thousands)
Day	35	30	135
Evening	30	28	140
Night	27	20	50

Construct a Laspeyre's price index using 2000 as base period and comment on the change.

58. In order to take decision about the adjustment of wages of labours with the hike of prices of different necessary commodities in the market in June 2011 as compared to December 2010, the authority of a factory collected the following information:

Year	Commodity (average price and consumption)							
	Rice		Dal		Vegetable		Oil	
	Price /kg	Quantity	Price /kg	Quantity	Price /kg	Quantity	Price /liter	Quantity
Dec 2010	35	30	85	120	20	16	85	3
June 2011	50	25	4	3	32	12	105	2.5

Compute the appropriate price index and suggest wages to be increased in July 2011.

59. After hearing many complaints that the prices of the fruits of a particular wholesale fruit shop during summer, the shop decided to see whether this is true. Based on the following data, help the shop keeper to calculate the appropriate aggregate price indices for each month.

Fruit	Price per kg			Amount sold in June
	June	July	August	
Apples	100	110	130	70
Oranges	90	105	120	75
Watermelon	50	60	65	100
Papaya	30	35	50	60
Grapes	200	220	300	25

(Hints: since only base year prices are given, you have to use Laspeyre's formula)

60. An increase of 60% of cost of a fuel raises the cost of living of a certain family by 5%. What percentage of its cost of living was due to buying that fuel before the change in the price?
61. An increase of 25% of house rent raises the cost of living of a certain family by 12%, what percentage of its cost of living was due to paying the house rent before the change.
62. During a certain period the cost of living index goes up from 200 to 310, and the salary of government employees is also raised from Taka 16250 to Taka 20200. Do the employees really gain, if so, by how much in real terms?
63. The cost of living index has been raised as 225 in 2010 as compared to 2005, while at the same period the wages of workers has been raised from Taka 8100 to Taka 10300, are the workers really gained by the raise of wages, if so, by how much in real term?

CHAPTER - 14

TIME SERIES ANALYSIS AND FORECASTING

14.1. Introduction

Forecasting or prediction is an essential tool in any decision-making. Its uses vary from determining inventory requirements for a local shoe store to estimating annual sales of general stores. Time series analysis is one quantitative method that can be used to determine patterns in data collected over time and to project these patterns to arrive at an estimate for the future. Hence, time series analysis helps cope with uncertainty about the future.

In this chapter, we will deal with issues involved in analyzing a special type of data set, called time series data. In particular, we will be interested in analyzing the measurements through time on a particular variable. However, the statistical data can be arranged in a number of ways—according to magnitude or size, according to place of occurrence or geographical location and according to the time of occurrence or in chronological order. When the data are arranged according to the time of their occurrence, they form a time series. Thus time series is the arrangement of statistical data in chronological order of time such as hourly, daily, weekly, fortnightly, monthly or yearly. For example, monthly product, monthly sales, quarterly corporate earnings, daily closing price of share, etc. Time series are of particular importance in the field of business and economics because variables like price, wages, production, sales, profits, etc. vary from one time period to another. A time series is a set of observations taken at specified times, usually at equal intervals. According to W.Z. Harisch "A time series may be defined as a sequence of values of some variables corresponding to successive points of times". Time series analysis is used to detect patterns of change in statistical information over a regular interval of time.

Definition. A time series is a set of numerical measurements on a time-dependent variable of interest arranged over a regular interval of time.

Suppose, y_1, y_2, \dots, y_k be the values of a variable for k different time periods t_1, t_2, \dots, t_k respectively, thus y is a function of t . Symbolically, $y = f(t)$.

Time series data typically possess special characteristics that necessitate statistical methods for their analysis. This type of data may be analyzed to achieve some or all of the following objectives.

14.2. Objectives of Time Series Analysis

The main objective of analyzing the time series data is to get a concrete idea about the past behaviour of data so that appropriate course of action for future can be taken. However, the objective can be pointed out as follows:

- i) To identify the pattern and trends, and isolate the influencing factors or effects
- ii) To apply the idea obtained from analyzing the pattern of time series data for future planning and control.

14.3. Importance of Time Series Analysis in Business Decision-Making

Time series analysis is of great importance to business executives. It is extremely useful for him/her in decision-making due to the following reasons:

- i) This is the most popular and so far the effective method for business forecasting.
- ii) It helps in understanding the past behaviour of economic process and in predicting the future behaviour.
- iii) It helps in planning future operations.
- iv) It helps in evaluating current achievement.
- v) It facilitates comparison of same phenomenon over two or more periods.

14.4. Components of a Time Series

It can be easily observed that the patterns or behaviour of different time series are different. This variation is due to presence of the affects of a number of inherent components. Hence, one way of thinking about the variation in behaviour of an actual observed series is to regard it as being made up of various components. Traditionally, there are four possible components, which have been considered with the notion that any or some or all might be present in any particular series of data. These components seem to be independent of each other and seem to be influencing the overall time series data. The components are as follows:

- i) Secular Trend (T_t)
- ii) Cyclical components (C_t)
- iii) Seasonal components (S_t)
- iv) Irregular components (I_t)

14.5. Mathematical Models for Time - Series Analysis

There are two mathematical models, which are commonly used for the decomposition of a time series into different components. These are:

- (i) Multiplicative model
- (ii) Additive model

14.5.1. Multiplicative model. In traditional or classical time series analysis, it is assumed that there is a multiplicative relationship among these four components. Let Y_t denotes the value of a series at time t . Symbolically,

$$Y_t = T_t \times S_t \times C_t \times I_t;$$

where, T_t = Trend component, S_t = Seasonal component, C_t = Cyclical component and I_t = Irregular component.

For example, if $T_t = 450$, $S_t = 1.4$, $C_t = 1.6$ and $I_t = 0.9$, then, $Y_t = 907.2$

This particular model is appropriate for those situations in which percentage changes best represent the movement in the series.

14.5.2. Additive model. According to this model, a time series is the sum of its four components. Symbolically,

$$Y_t = T_t + S_t + C_t + I_t$$

For example, if trend = 250, seasonal fluctuation = 45, cyclical fluctuation = 12 and irregular component = -6, then

$$T_t = 250 + 45 + 12 - 6 = 301$$

This model assumes that all the components of a time series are independent of one another. For example, it assumes that trend has no effect on the seasonal and cyclical components, nor seasonal swings have any influence on cyclical fluctuation and vice-versa. However, in most of the business and economic time series, this assumption is not true. For example, the seasonal or cyclical fluctuations may virtually be wiped off by very sharp rising or falling trend. Similarly strong and powerful seasonal swings may strengthen or even precipitate a change in cyclical fluctuations. This model also assumes that the different components are absolute quantities expressed in original units and can take positive and negative values.

Both models may be equally acceptable, although it is easier to understand the techniques associated with time series if the multiplicative model is referred. However, for the proper analysis of time series data, sometimes it might be convenient to treat some factors as additive and others as multiplicative.

Some experts thought that in some situations, the appropriate model may be the combination of the above two models. For example, the time series models may be of the form :

$$Y_t = T_t \times S_t + C_t \times I_t \quad \text{or,} \quad Y_t = T_t + S_t \times C_t + I_t;$$

Remarks. As most of the time-series relating to economic and business phenomenon conform to the multiplicative model, the additive model is used rarely.

14.6. Description of Time Series Components

The important components of a time series are

- i) Secular trend
- ii) Seasonal variation
- iii) Cyclical fluctuation and
- iv) Irregular variation.

A brief description of different components of time series data and some of their important aspects are provided below:

14.6.1. Secular Trend. Many time series met in practice exhibit a tendency of either growing or reducing fairly steadily over time. This tendency of a time series data over a long period of time is called secular trend. Some series increase slowly, some increase fast, others decrease at varying rate, and some remain relatively constant for long periods of time. For example, despite short-run deviations from the trend, gross domestic product of a country, prices or productions of a firm or a country, money supply show an upward trend over the years due to increasing population. On the other hand population growth or death rate shows a downward trend. The time series data can be represented by a Histogram, which exhibits its long-term trend.

The various types of trends are divided into two classes viz

- (i) Linear or straight line trends, and
- (ii) Non-linear trends.

It is to be noted that by secular trend, we mean smooth, regular, long-term movement of the data – sudden or irregular movements either in upward or downward direction have nothing to do with the trend. For illustration, let us consider the following Histogram of yearly car ownership rate in UK.

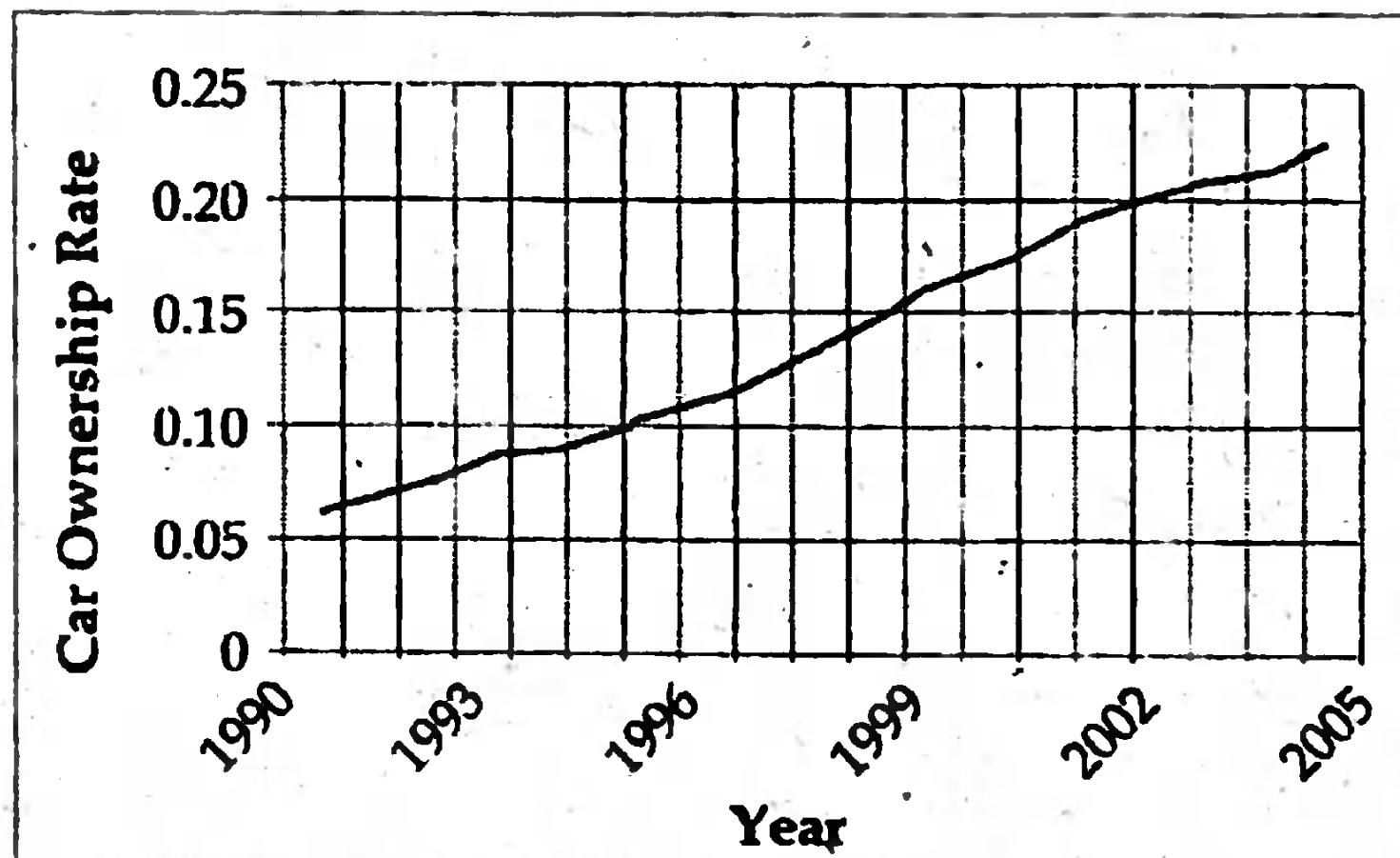


Fig. 14.1. Yearly Car ownership rate in UK.

It is clear from the Figure 14.1 that the behaviour of this time series is characterized by a fairly steady upward trend. The rate of growth of car ownership is less rapid in the later part of the period than in middle, raising the possibility of a slowly growing trend pattern.

Factors affecting trend

There are several factors that affect trend in time series data. Some of the important factors are:

- i) Population: The ever-increasing population of a country is responsible for increasing trends in series like prices, production, sales, etc.
- ii) Technology, Institution and culture: Downward or upward trends in some factors are caused by technological, institutional or cultural changes. For example, progress in automobile industry reduces the road accident and increases the number of cars, buses, trucks etc. On the other hand, better medical facilities, improved sanitation, diet, etc. reduce the death rate, consequently, contribute to a rise in birth rate.

Reasons for studying Trends

The reasons for studying the trend in a time series data are pointed below:

- i. It allows us to describe the historical pattern of time series data.
- ii. It permits us to project past pattern or trends, into the future.
- iii. It facilitates us to eliminate the trend component from the series in order to obtain the de-trended series that is useful for studying other components of time series data.

14.6.2. Seasonal Variation. Seasonal variations are like cycles, but they occur over short and repetitive calendar periods. By seasonal variation we mean a periodic movement that repeats itself with remarkable similarity at a regular interval of time, the period being no longer than one year. Hence, seasonal component of a time series data is the repetitive and predictable movement of observations around the trend line during particular time intervals of the year. In order to measure or to detect the seasonal component, the data must be given in small unit of time, such as hours, days, weeks, months or quarters. Many business and economic time series met in practice consists of quarterly or monthly observations.

Seasonal variation arises as a result of natural changes in the seasons during the year or may result due to habits, customs, or festivals that occur at the same time of every year. For example, retail sales of spices is the highest in the month of Eid, retail sales of egg is the lowest in Ramadan, retail sales of many products tend to be relatively high in December, because of Christmas shopping, etc. Although the amplitude of seasonal variations may vary, their period is fixed being one year; as a result seasonal variations appear in

daily, monthly or quarterly series and do not appear in any annual series. Let us consider the following graph of quarterly prices of share of a company.

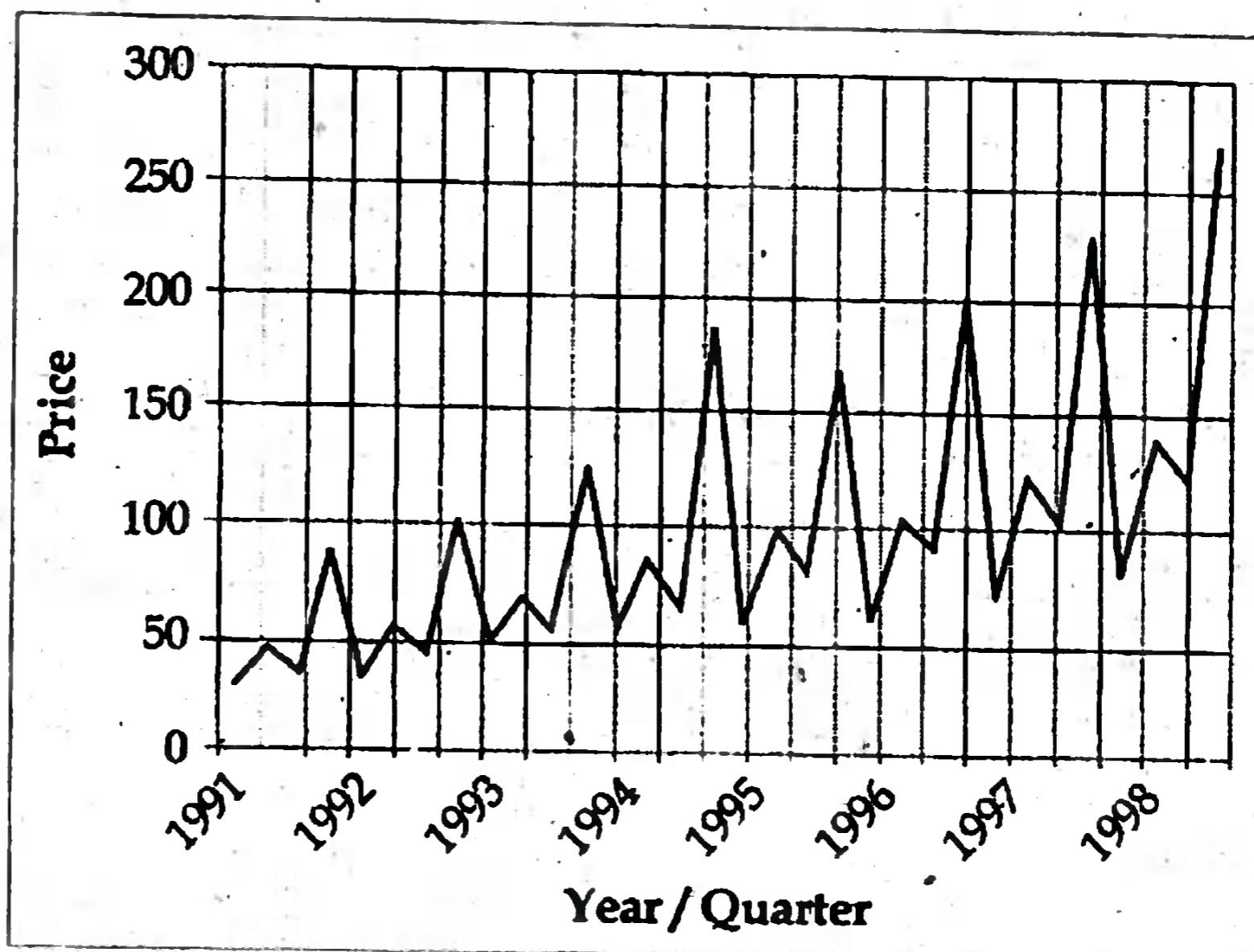


Fig. 14.2. Quarterly Prices of the Share of a Company Over Eight Years.

It is seen from the Figure 14.2 that the fourth quarter prices tend to be relatively high, while those in the first quarter are quite low. Hence the seasonal behaviour of the data is quite clear from the figure where there is an obvious pattern repeating each year. The prices in the second quarter are somewhat higher than those of the immediately preceding or succeeding quarter, while those of the fourth quarter are much higher yet. The figure also makes clear that there is another component of the time series. Apart from the obvious seasonality, there is a noticeable upward trend in prices over the period covered by the data. So the data consists of minimum two components viz. trend and seasonal components.

Factors affecting seasonal variations

Seasonal variations in a series are also caused by some factors. Some of the important factors are:

- i) **Climate and weather factors:** Change in the climate and weather conditions such as humidity, heat, rainfall, etc. act on different product and industries differently and cause change in demand of them. For example, during winter there is greater demand for woolen clothes and hot drinks etc. whereas in summer cotton clothes, cold drinks have a greater sale.
- ii) **Social customs or Religious factors:** Due to some religious festival the sales or demand of particular product varies over the year. For example, during Ramadan the demand or sales of chickpeas is higher than any other period of the year, similarly, demand of eggs is lower in this

period, again, due to Christmas the demand of cakes or sweetmeats is higher than any other month of the year, during valentines day or New year, the demand of flower is higher than any other time of the year, during Durga Puja or Eid-ul-azha the demand of ready made garments is higher than any other period and so on.

14.6.3. Cyclical fluctuation. A cycle is a wave like pattern about a long-term trend that is usually apparent over a number of years. The term cycle refers to the recurrent variations in time series that usually last longer than a year. One complete period is called a cycle. Many business and economic time series met in practice appear to exhibit oscillatory, or cyclical, pattern unconnected with seasonal behaviour. They are not necessarily regular, but follow rather smooth pattern of upswings or downswings. Examples of cycles include the business cycle that record periods of economic recession and inflation, long term product-demand cycles and cycles in the monetary and financial sectors. There are four well-defined periods or phases in the business cycle, namely, i) prosperity, ii) decline, iii) depression and iv) improvement. Different phases of business cycle are shown in following graph.

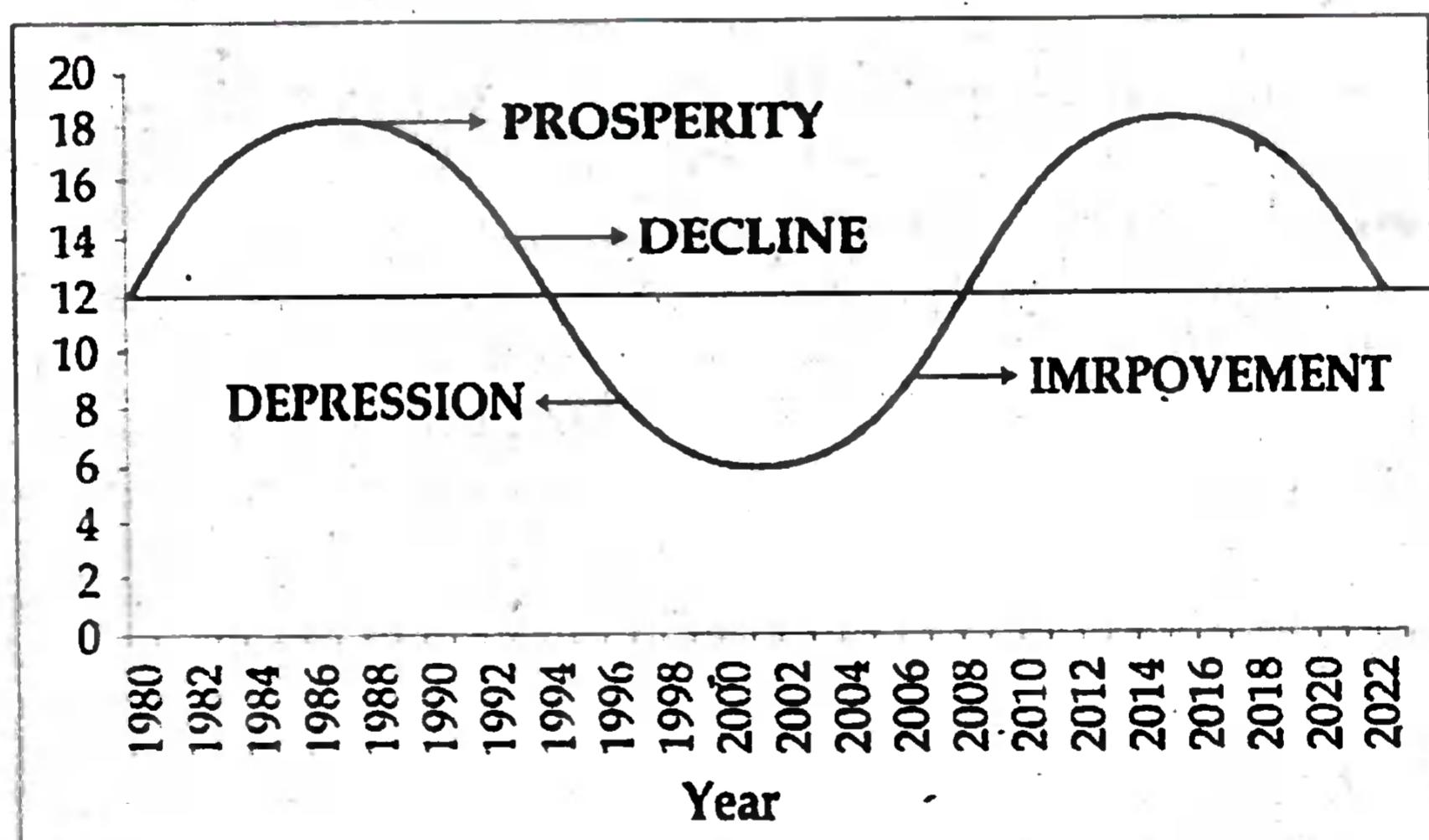


Fig. 14.3. Phases of Business Cycle.

In Figure 14.3 the period 1980-87 is the period of prosperity, after that period, it comes to its normal condition in 1993-94 (let us consider this as the normal period), the curve starts declining and it continues up to 1988-94, so this period is called the period of decline. However, the curve reaches its lowest peak in the period 1995-2001 which can be considered as the period of depression, and after 2001 the factor again starts rising and reached its normal position in 2008-09, and an improvement over the normal period is observed after 2008-09, so the period 2008-2015 is termed as the period of improvement.

Difference between business cycles and seasonal variation

We know the period of seasonal variation is less than a year and this type of variation is found only in the data collected for the period of less than a year, such as hourly, daily, monthly, quarterly, etc. On the other hand, a business cycle has the following features:

- i) The business cycles or cyclical variations are of longer duration than a year. A business cycle may be of any duration but normally the period of business cycle is 2-10 years. Moreover, they do not ordinarily exhibit regular periodicity as successive cycles vary widely in timing, amplitude and pattern. For example, the 23 cycles of general business in the USA between 1954 and 1949 averaged 40 months, in duration individual cycles differed greatly. The shortest period lasted only 20 months and the longest lasted for 29 months.
- ii) The fluctuations in a business cycle results from a different set of causes. The period of prosperity, decline, depression and improvement viewed as four phases of business cycle which are generated by factors other than the factors which are responsible for seasonal variations such as weather, social customs, etc.

14.6.4. Irregular variation. So far we have discussed three sources of variability in a time series. If the components of a time series are found to be trend, seasonal and cyclical, we would expect the series very smooth and rather easily projected forward to produce forecast. But this may not be the case in practice. Because of an inevitable presence of another component known as irregular element, induced by the multitude of factors influencing the behaviour of any actual series and whose pattern becomes rather unpredictable on the basis of past experience.

Random variation comprises the irregular changes in a time series are not caused by any other component. It tends to hide the existence of the other more predictable components. Hence, irregular variation refers to such variation in business activity, which does not repeat in a definite pattern. This may happen due to catastrophic events like earthquake, flood, fire, war, strike etc. all these parts of variation of a time series data cannot be explained by trend, seasonal or cyclical variation.

There are two reasons for identifying the irregular components in time series data. These are:

- i) To suggest that on occasions it may be possible to explain certain moments in the data as due to specific causes and to simplify further analysis.
- ii) To emphasize the fact that prediction of economic conditions is always subject to certain degree of error owing to the unpredictable erratic influences, which may enter.

14.7. Measurement of Trend Component

Of the four components of time series, secular trend represents the long-term direction of the time series. One way to describe the trend component is to fit a line visually to a set of points on a graph. Any given graph, however, is subject to more or less different interpretations by different individuals. Trends may be linear or curvilinear. Here, we will be mainly dealing with linear trends, and concepts about the non-linear trends will be provided in short. However, the following are the methods used to eliminate linear trend component from a given time series data:

- i) Graphic or free hand curve method
- ii) Semi-average method
- iii) Least squares method
- iv) Moving average method

14.7.1. Graphic method. The given data are plotted on a graph paper and a free hand trend line fitted to the data is obtained just by inspection. A freehand curve drawn through the data values is often an easy and perhaps, adequate representation of data. Forecasts can be obtained simply by extending the line. A trend line fitted by this method should conform to the following conditions:

- i. The trend line should be smooth – a straight line or mix of long gradual curves.
- ii. The numerical sum of vertical deviations of the observations below the trend line should be equal to the numerical sum of deviations above the line.
- iii. The sum of squares of deviations of the observations from the trend line should be as small as possible.
- iv. The trend line should bisect the cycles so that area above the trend line should be equal to the area below the line, not only for the entire series but as much as possible for each full cycle.

Example 14.7.1. Fit a trend line to the following data by using freehand curve method :

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales Turnover (Taka in lac)	80	85	97	110	160	94	86	174	180	200	135	120	105

Solution. At first we plot the turnover against year in a graph paper. Then we draw a free hand trend line on the graph, the estimated trends are the values lying on the line corresponding to respective years. Forecast of the turnover can be obtained simply by extending the trend line.

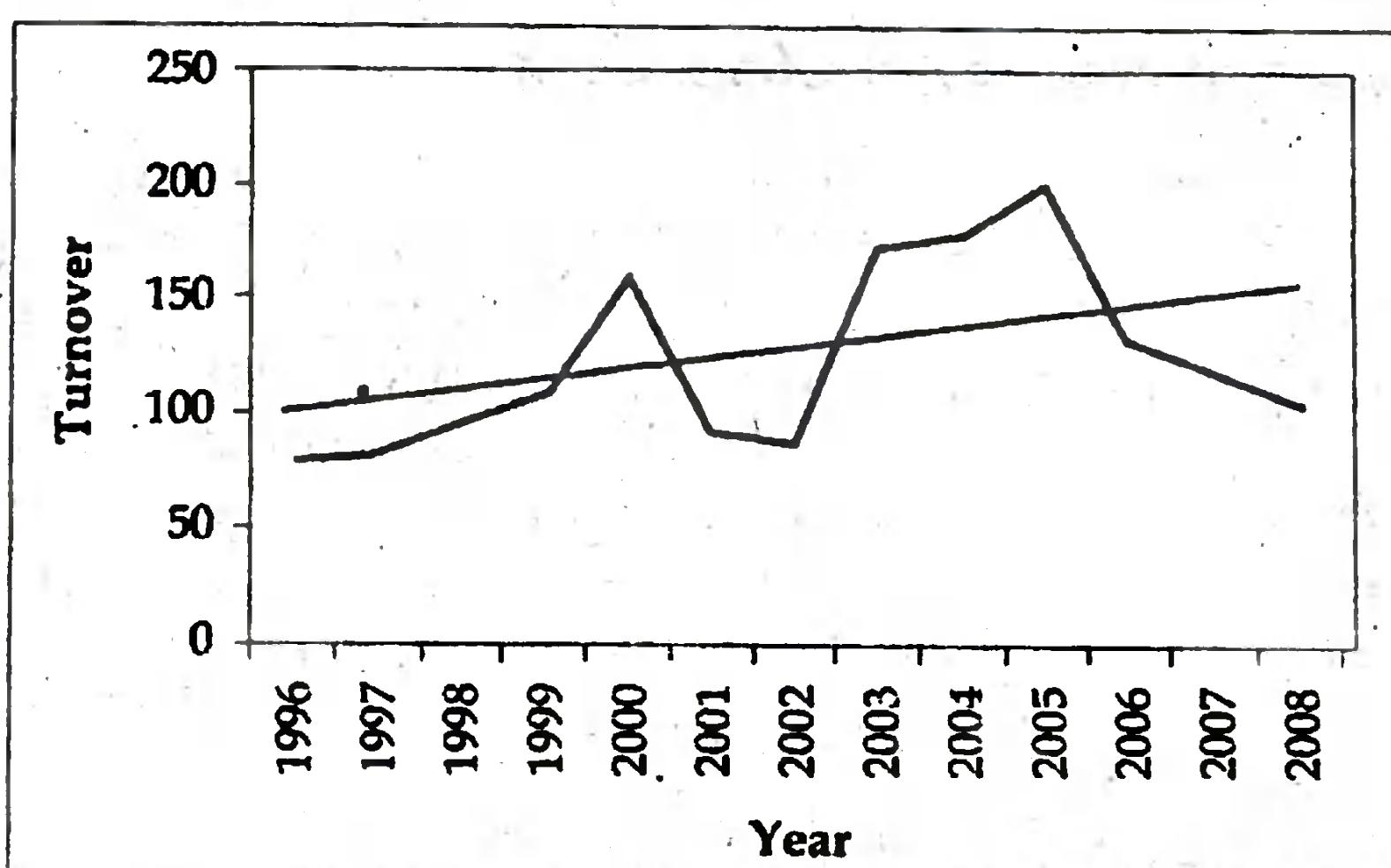


Fig. 14.4. Freehand Trend values of Turnover.

The trend values obtained by freehand method are as follows:

Table 14.1. Trend values of turnover obtained by freehand curve method.

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales Turnover (Taka in lac)	101	105	109	113	117	121	125	129	133	137	141	145	149

Merits and Demerits of the Freehand curve method

Apparently this method is very simple to determine the trend values of a time series data. However, this method has following merits and demerits.

Merits:

- This is the simplest, quickest and easiest method of estimating trend values.
- This method is very flexible in the sense that it can be used irrespective of the nature of trend component, whether it is linear or non-linear.
- If the statistician knows the past behaviour of the data series, it is possible to obtain secular trend by this method, even sometimes better than by any other mathematical method.

Demerits:

- From its nature of fitting, it is clear that this method is very subjective, because the trend line depends on the personal judgment and therefore what happens to be a good fit to one individual may not be so for other.
- The trend line drawn cannot have much value if it is used as a basis for prediction.
- It is very time consuming to construct a free hand curve if a careful and conscientious job is to be done.

14.7.2. Semi-average method. In this method the given data are divided into two equal parts preferably with equal number of periods. If there is odd number of years or period like 7, 9 etc. the middle year is ignored and the two equal parts are formed. An average of each part is computed and the two points thus obtained are centered corresponding to the middle period and shown on the graph. A straight line is drawn through these two points. The values lying on this line describe the trends. By projecting the line it is possible to forecast the future values.

Example 14.7.2. Apply the method of semi-average of measuring the trend of sales of a commodity from the following data:

Year:	1985	1986	1987	1988	1989	1990
Sales (in '000):	20	24	22	30	28	32

Solution. Since there are six years, we will take an average of the first three years namely 1985, 1986, 1987 and last three years namely 1988, 1989, 1990. These averages will be plotted corresponding to the middle period i.e. 1986 and 1989 respectively. When these two points are joined we will get the trend line by the method of semi-averages.

Now, average of first three observations is 22 which is plotted corresponding to 1986 and average of last three observations is 30 which is plotted corresponding to 1989. The trend values are the values corresponding to the straight line joining two averages as shown in figure 7.5a. The trend values in this method are determined using the following method. It has been found that the average corresponding to 1986 is 22 and to 1989 is 30. So there is an increment of $(30 - 22) = 8$ over three periods 1986 – 1989, that means each year's increment is $8/3$. This value can be considered as the correction factor for the calculation of trends for other periods. Thus, for trend value corresponding to the following year 1987, add up $8/3$ with 22 which is given by $(22 + 8/3) = 24.67$. Similarly for each subsequent year trend value is calculated adding $8/3$ to the previous period's trend value. On the other hand, for every preceding year subtract $8/3$ from the estimated trend value of following year. For example, trend value for the year 1985 is calculated as $(22 - 8/3) = 19.33$. In this way, trend values for any expected year can be determined by extending the line, that means, adding $8/3$ to each of the preceding year. The estimated trend values of sales obtained by semi average method along with original time series of sales are shown in figure 14.5.

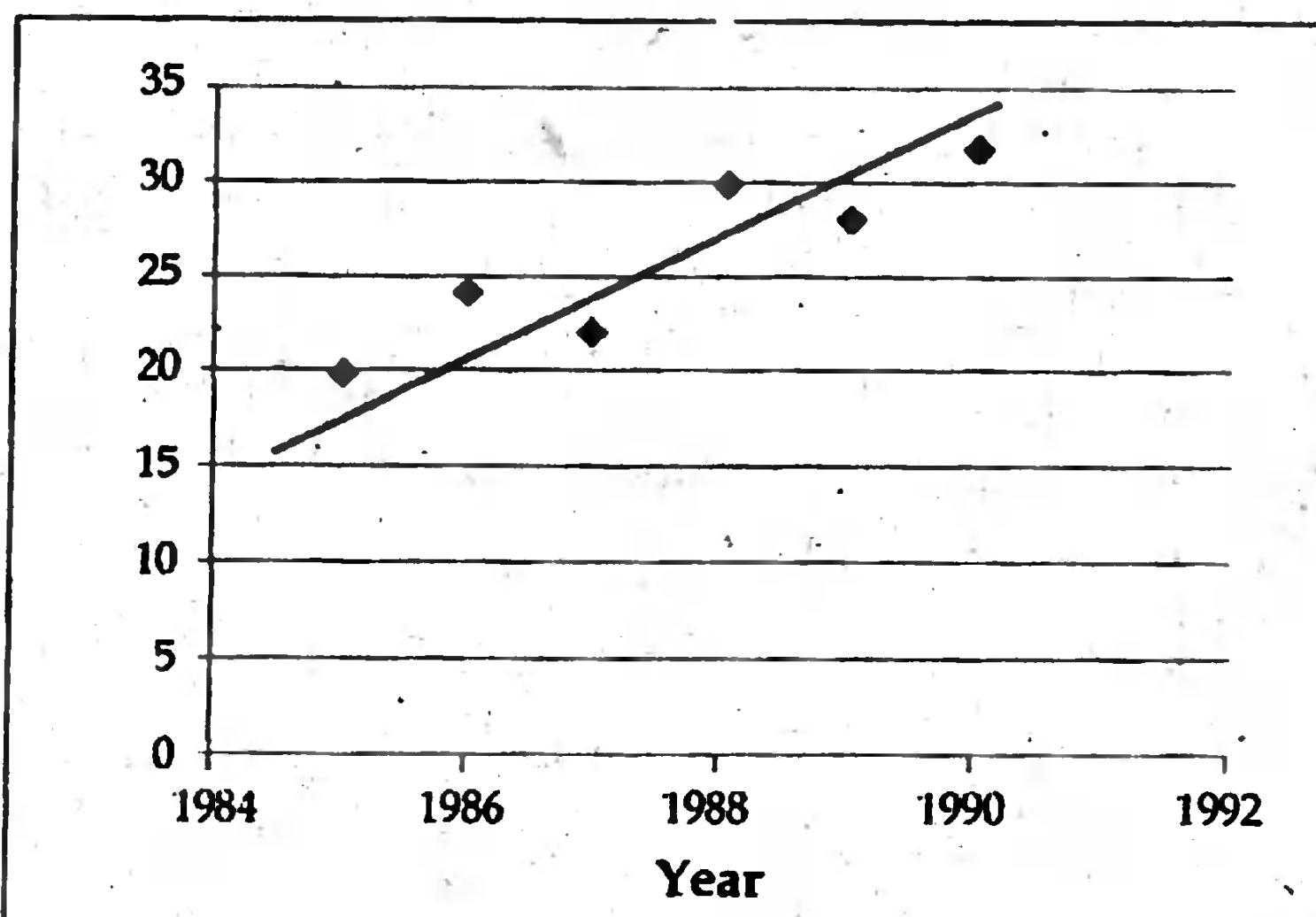


Fig. 14.5. Trend line obtained by semi-average method.

The trend values of sales obtained by semi-average method are shown in following table.

Table 14.2. Trend values of sales obtained by semi-average method:

Year:	1985	1986	1987	1988	1989	1990
Sales (in '000):	20	24	22	30	28	32
Trend component:	19.33	22.00	24.67	27.33	30.00	32.67

Example 14.7.3. Fit a trend line to the following data by semi-average method:

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales Turnover (Taka in lac)	80	85	97	110	160	94	86	174	180	200	135	120	105

Solution. Since there are thirteen years, in order to make these data into two equal parts having six years in each part, we are to ignore the middle year 2002. Now, let the parts be 1996 – 2001 and 2003 – 2008 respectively. The average turnovers for these two periods are 104.33 and 152.33 respectively. By plotting these averages corresponding to the mid-year of two periods, namely 1998.5 and 2005.5 respectively, we get the trend line as shown in figure 14.6. However, it is possible to forecast the trend values by extending the line.

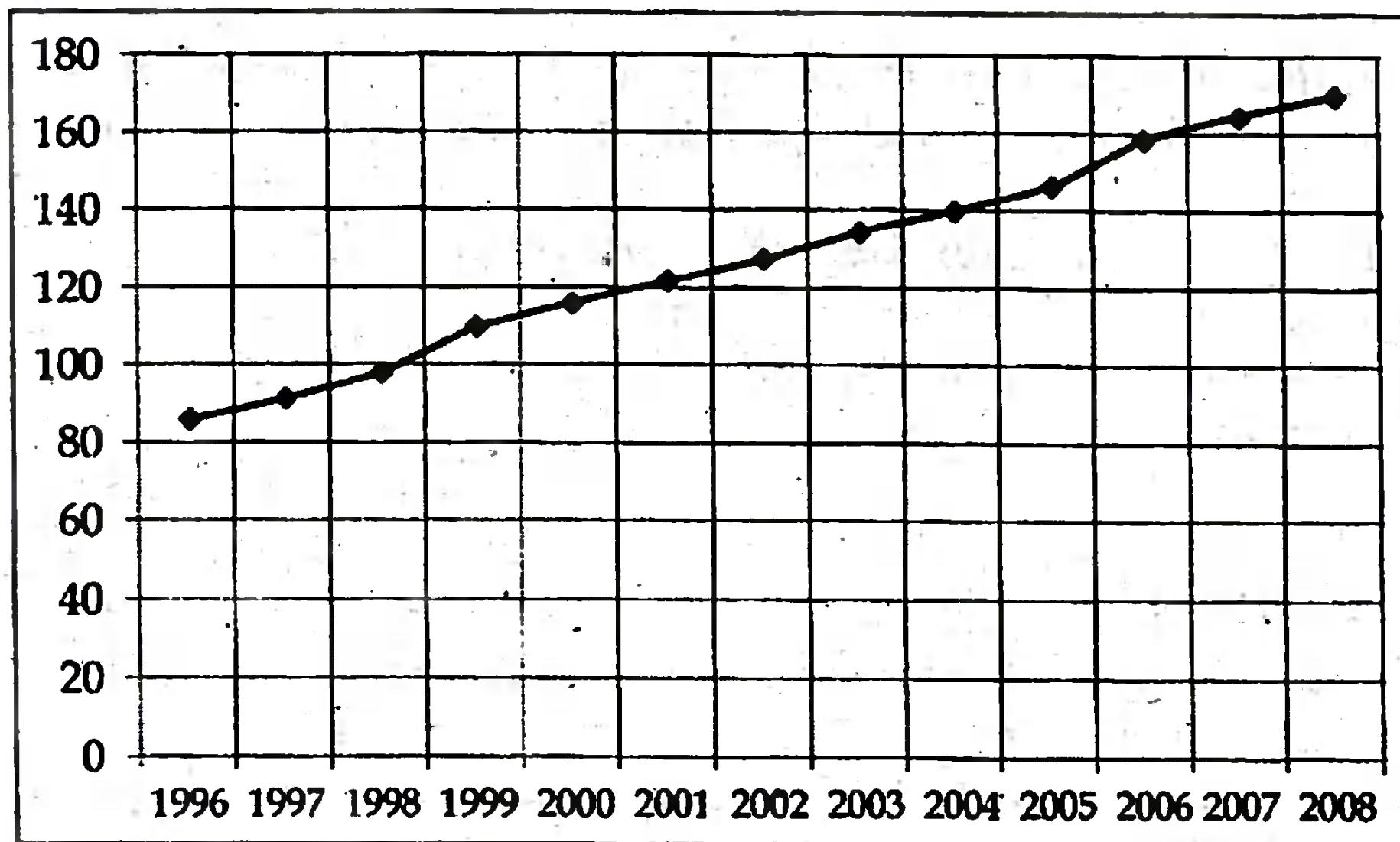


Fig. 14.6. Trend line obtained by semi-average method.

The trend values obtained by the semi average method are shown below:

Table 14.3. Trend values of turnover obtained by semi-average method.

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales Turnover (Taka in lac)*	80	85	97	110	160	94	86	174	180	200	135	120	105
Turnover Trends	86	92	98	110	116	122	128	134	140	146	158	164	170

Merits and Demerits of the Semi-average method. Like the free hand curve method, the semi-average method of determining the trend has the following merits and demerits.

Merits:

- i. This is a more logical and easier method of determining trend than freehand curve method.
- ii. This method does not take that much time to find the trend component.
- iii. This is an objective method of determining trend as everyone who ever applies this method is bound to get the same result.

Demerits:

- i. This method assumes straight line relationship between the time period and the observations, so if the relationship deviates much from the linearity, then forecast by this method will be biased and less reliable.
- ii. Since this method is based on arithmetic mean, it also bears the limitations of this mean. For example, if there are some extreme values in either half or both half of the series, then the trend line will

not indicate the true picture of the growth factor. This danger is the greatest if the time period represented by the average is small.

14.7.3. Fitting Linear Trend by the Method of least square. We have already discussed two methods of obtaining trend values from a time series data. One is free hand curve method, by which one can obtain linear or non-linear trend which are very subjective and that is why we cannot rely on such trend values. The semi average method provides only linear trends obtained by joining two averages, which is although free from subjective error, but are not based on strong mathematical logic. Because, the only trend line obtained by this method may not fit the given data well. Here, we will discuss the method by which it is possible to obtain linear or non-linear trend by establishing suitable functional relationship between the time variable and the objective variable. This method is known as the Least squares method. The name is due to the fact that sum of the squares of differences between the observed value and the predicted trend values is the least. The trend line obtained by this method is considered as the best line. Thus, with the help of the method of least squares we can fit either a straight-line trend or a second-degree parabola. The equation of straight line or linear trend is given by, $y = a + bx$, where y represents the variable, x represents the transformed time period, a and b are two constants, a is the intercept and b is the slope of line. In this case the method of estimating parameter is same as that of linear regression line. While the equation of second degree parabola is given by : $y = a + bx + cx^2$.

Steps involved in Least squares method

The following steps are involved in determining trend by least squares method:

- i) Plot the observed data against time period in a graph paper and observe the nature of relationship between time variable and observations.
- ii) Consider a linear trend line or second-degree parabola according to as it is observed from the graph. Consider converted time period as independent variable and the observations as dependent variable.
- iii) Fit the postulated line or parabola by using least squares method.
- iv) Obtain the estimated trend values putting the coded values of time variable in the fitted equation of trend line.

Translating, converting or coding time variable

Usually the independent variable time is measured in terms of weeks, quarters, months or years. In order to fit a trend equation of observed variable on the time period, it is convenient to convert these traditional measures of time into the form that can simplify the computation process.

This process of conversion is called coding. The convenient way of coding is to find the mean time and then subtract that value from each of other sample time periods. For example, suppose a time series consists of only three periods 2006, 2007 and 2008, if we consider numbers in trend line equation, we will find the resultant calculations tedious. Instead, we can transform the values 2006, 2007 and 2008 into corresponding values -1, 0 and 1 respectively, where, 0 represents the mean (2007), -1 represents the first year ($2006 - 2007 = -1$) and 1 represents the last year ($2008 - 2007 = 1$). This is the case of odd number of periods, however, in practice we may have even number of periods. In that case, if we subtract the mean period from other periods, the fraction 0.5 will be part of numbers, which will again make the calculations tedious. To simplify this problem and to remove the matter of fraction 0.5, multiply each time period by 2. Two examples are given below to illustrate these two cases.

However, we get two advantages from this type of translation of time. First, it eliminates the need to square the large numbers like 2006, 2007, etc. and secondly, the method sets the mean year of independent variable (x) equal to zero and simplifies the computations of constants in the trend line, particularly, the value of the constant a becomes equal to the mean of dependent variable.

Fitting Linear Trend line

We know this method is based on two conditions such as

- i) The sum of deviations of the actual values of y and the fitted trend values is zero, and
- ii) The sum of squares of deviations of the actual and fitted values is the least from the fitted line.

The line that satisfies these two conditions is known as the line of best fit. Let $y = a + bx$ be our trend line, where y represents the variable, x represents the time period, a and b are two parameters. We know the least square estimates of parameters a and b are given by

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

Thus the fitted trend line is given by $\hat{y} = a + bx$.

However, as we know that translation of time variable provides us with some advantage, and for translated time variable x , the mean time \bar{x} becomes zero, so the estimates of a and b become

$b = \frac{\sum xy}{\sum x^2}$ and $a = \bar{y}$ which are simpler than the previous values.

Example 14.7.4. Fit a trend line to the following data by Least squares method and estimate the trend component of time series data.

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales Turnover (Taka in lac)	80	85	97	110	160	94	86	174	180	200	135	120	105

Solution. We have to fit the linear trend line $y = a + bx$, where y represents the sales turnover and x represents the coded year. The values of a and b are to be computed using the formula

$$b = \frac{\sum xy}{\sum x^2} \text{ and } a = \bar{y}$$

Let us construct the following table for necessary calculations of sum of squares and sum of product.

Table 14.4. Calculations of Trend values of turnover by Least Squares method.

Year	Turnover (y)	Coded period (x) Taking origin at 2002	x^2	xy	Estimated Trend values
1996	80	-6	36	-480	96.52
1997	85	-5	25	-425	101.28
1998	97	-4	16	-388	106.04
1999	110	-3	9	-330	110.80
2000	160	-2	4	-320	115.56
2001	94	-1	1	-94	120.32
2002	86	0	0	0	125.08
2003	174	1	1	174	129.84
2004	180	2	4	360	134.60
2005	200	3	9	600	139.36
2006	135	4	16	540	144.12
2007	120	5	25	600	148.88
2008	105	6	36	630	153.64
Total	$\Sigma y = 1626$	$\Sigma x = 0$	$\Sigma x^2 = 182$	$\Sigma xy = 867$	

From table 14.4, we have, $\Sigma x = 0$, $\Sigma y = 1626.00$, $\Sigma x^2 = 182$ and $\Sigma xy = 867$.

Using the formula mentioned above, we get the estimated values of b and a as $b = 4.76$ and $a = 125.08$

The fitted equation of trend line is $\hat{y} = 125.08 + 4.76 x$

From the fitted equation, the trend values are estimated as in the following way:

$$\text{For 1996, } \hat{y}_{1996} = 125.08 + 4.76(-6) = 96.52$$

$$\text{For 1997, } \hat{y}_{1997} = 125.08 + 4.76(-5) = 101.28 \text{ and so on.}$$

We can also project the trend value for the year 2010 just by putting the value of x equal to 8 which is $\hat{y}_{2010} = 125.08 + 4.76(-8) = 163.16$. Thus we can say that the expected sales turnover for the year 2010 is taka 163.16 lac. In this way trend value for any expected time period can be estimated.

The trend values of turnover obtained by least squares method along with original are presented in figure 14.7.

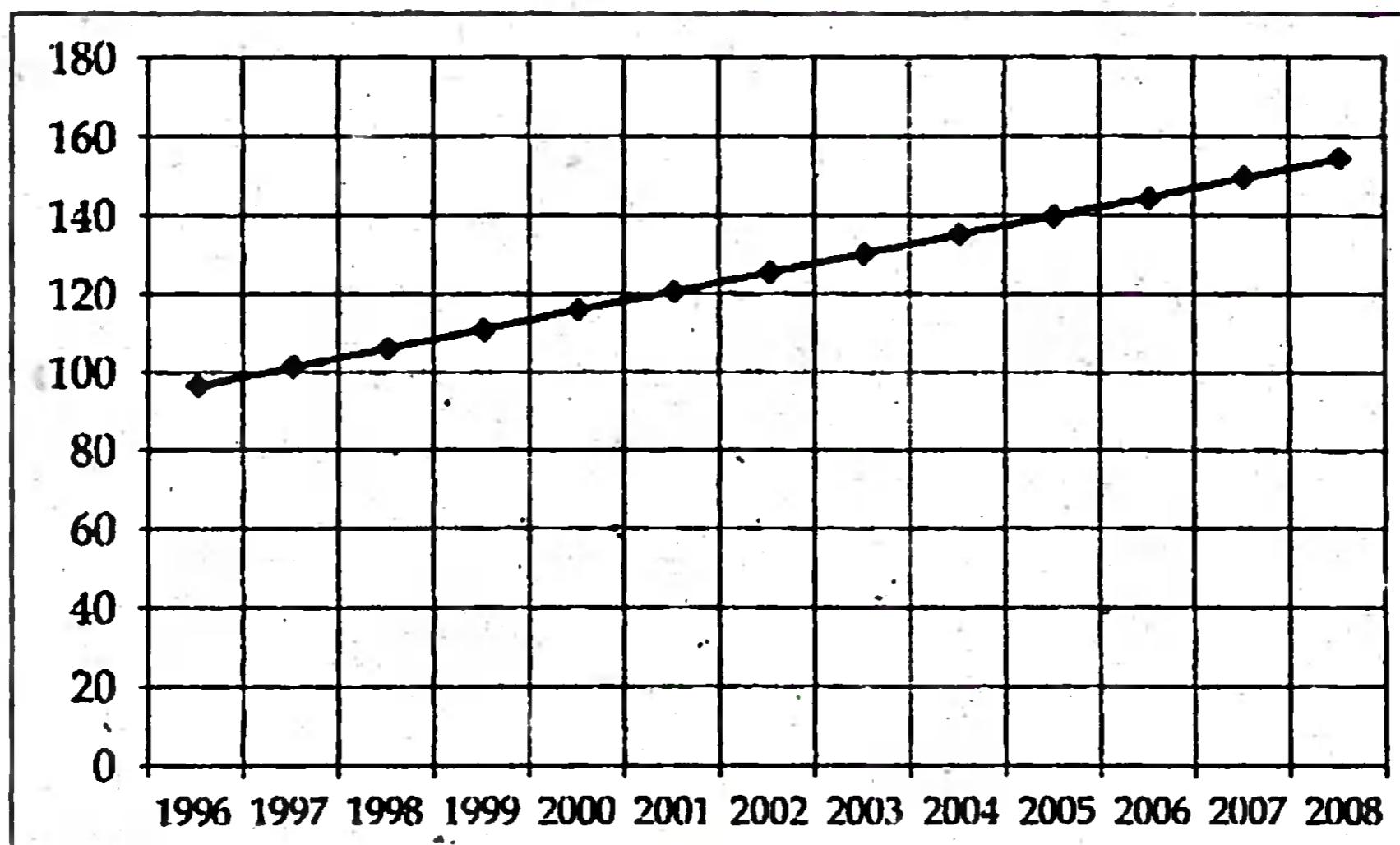


Fig. 14.7. Least squares trend values of turnover.

Example 14.7.5. The following are the annual profits in thousand of taka in a certain business. Use the method of least squares to determine the trend values of profit. Also forecast the annual profits for 2009.

Year	Profit	Year	Profit
1999	60	2004	85
2000	72	2005	95
2001	75	2006	101
2002	65	2007	107
2003	80	2008	115

Solution. We have to fit the linear trend line $y = a + bx$, where y represents the sales turnover and x represents the coded year. Here, the number of years is 10, which is even. So, in order get the convenient values of coded year (x), we have to multiply difference of each year from mean year by 2.

For example, here the mean year is 2003.5, the first value of coded year is calculated as $(1999 - 2003.5) \times 2 = (-4.5) \times 2 = -9$.

The values of a and b are to be computed using the formula

$$b = \frac{\sum xy}{\sum x^2} \text{ and } a = \bar{y}$$

Let us construct the following table for necessary calculations of sum of squares and sum of product.

Table 14.5. Calculations of Trend values of profit by Least Squares method.

Year	Profit (y)	Coded period (x) [Taking origin at 2003.5 and multiplying by 2]	x^2	xy	Estimated Trend values
1999	60	-9	81	-540	59.22
2000	72	-7	49	-504	65.06
2001	75	-5	25	-375	70.9
2002	65	-3	9	-195	76.74
2003	80	-1	1	-80	82.58
2004	85	1	1	85	88.42
2005	95	3	9	285	94.26
2006	101	5	25	505	100.1
2007	107	7	49	749	105.94
2008	115	9	81	1035	111.78
Total	$\Sigma y = 855$		$\Sigma x^2 = 330$	$\Sigma xy = 965$	

Here $\Sigma x = 0$, $\Sigma y = 855$, $\Sigma x^2 = 330$ and $\Sigma xy = 965$

Using the formula mentioned above, we get the estimated values of b and a as $b = 2.924$ and $a = 85.5$

The fitted equation of trend line is $\bar{y} = 85.5 + 2.924x$

From the fitted equation, the trend values are estimated as in the following way;

For 1999, $\bar{y}_{1999} = 85.5 + 2.924(-9) = 89.22$

For 2000, $\bar{y}_{2000} = 85.5 + 2.924(-7) = 65.06$ and so on.

The trend value for the year 2009 can also be predicted just by putting the value of x equal to $(2009 - 2003.5) \times 2 = 11$ which is $\bar{y}_{2009} = 85.5 + 2.924(11) = 117.62$. Thus we can say that the expected profit for the year 2009 is taka 117.62 thousand. In this way trend value for any expected time period can be estimated.

The estimated trend values of profit and observed profits are shown in figure 14.8.

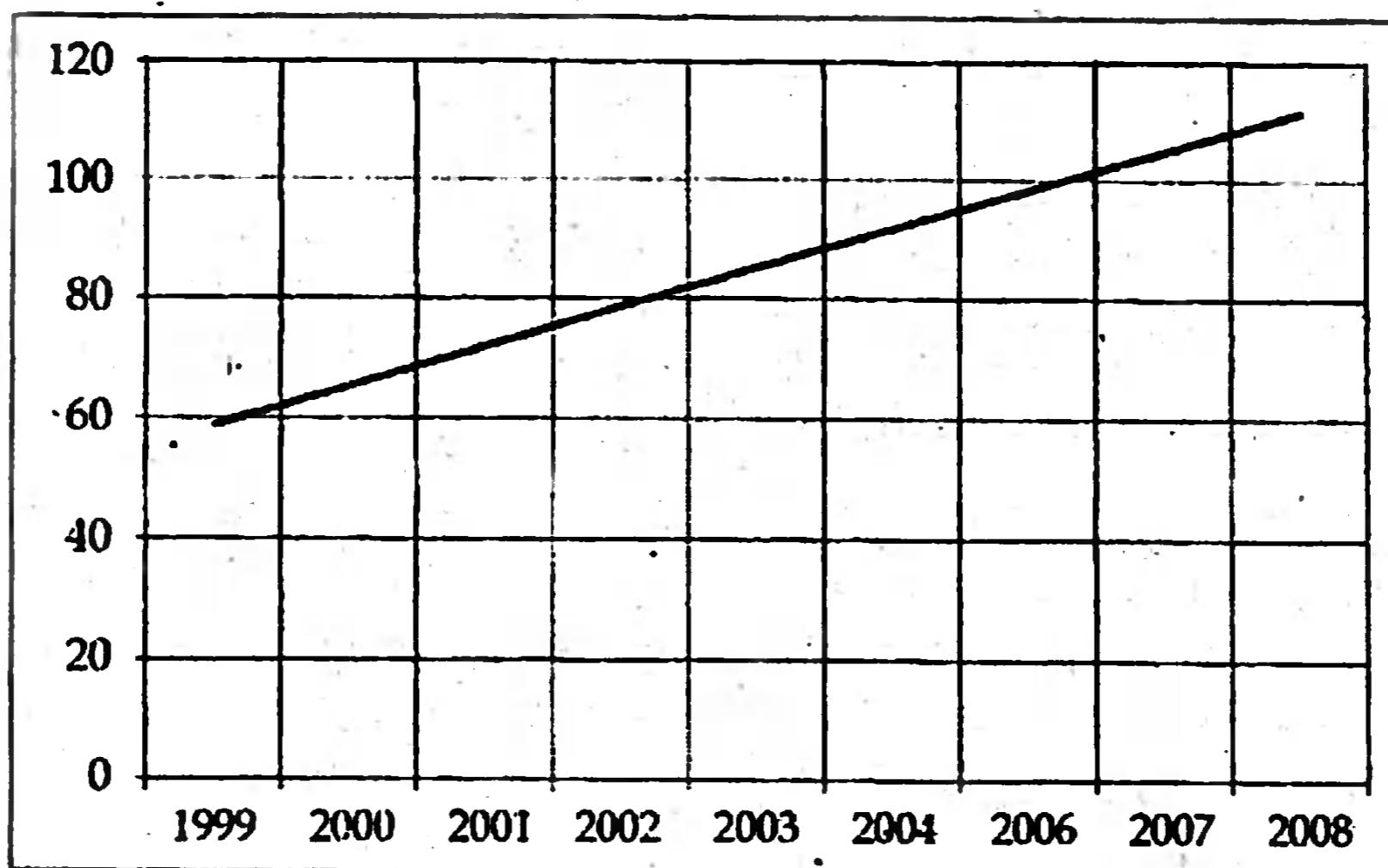


Fig. 14.8. Least squares trend values of profit.

Merits and Demerits of Least Squares Method

It is already mentioned that this method is based on two conditions which are assumed to necessary for getting estimated trend values efficiently. However, this method also has following merits and limitations:

Merits:

- i. This is a mathematical method of measuring trend and as such there is no possibility of subjectiveness.
- ii. The trend line obtained by this method is called the line of best fit, because from this line the sum of squares of differences between the observed time series and estimated trend values is the least.
- iii. This method enables us to compute the trend values for all the given time period in the series.
- iv. This method is the only technique, which enables us to obtain the rate of growth per annum for yearly data in case of linear trend.
- v. The projected trend values obtained by this method are more reliable than any other method.

Demerits:

- i. For using this method, one needs to have higher mathematical knowledge.
- ii. Calculations required in this method are quite tedious and time consuming as compared with other methods.
- iii. If the functional form of the relationship between time period and observations cannot be defined rightly, the trend values obtained by this method may be misleading.

- iv. Fresh calculations become necessary if even single new observations are added.
- v. Future predictions based on this method completely ignore the cyclical, seasonal and erratic fluctuations.
- vi. This method cannot be used to fit growth curves which most of the business and economic time series conform.

Comparison of trend values obtained by free hand method, semi-average method and least squares method

We have already mentioned three methods of measuring trend component of a time series data. Now we shall compare the performance of these three methods. The only criteria sum of squares of deviation or residuals (SSR) of observed time series and estimated trend values is used for here for comparison. The definition of SSR is very simple and is given by $SSR = \sum (y - \bar{y})^2$. The performance of the methods of determining linear trends of a time series data is illustrated with two examples.

Example 14.7.6. Let us consider the data of turnover illustrated earlier for all these three methods. The computed values of SSR obtained by these methods are shown in following table.

Table 14.6. Sum of squares of residuals of trend values obtained by different methods.

Year	Turn-over	Freehand trend values (FHT)	Semi-average trend values (SAT)	Least squares trend values (LST)	SSR For FHT	SSR for SAT	SSR for LST
1996	80	101	86	96.52	441	36	272.91
1997	85	105	92	101.28	400	49	265.04
1998	97	109	98	106.04	144	1	81.72
1999	110	113	104	110.80	9	36	0.64
2000	160	117	116	115.56	1849		1974.91
2001	94	121	122	120.32	729	1936	692.74
2002	86	125	128	125.08	1521	784	1527.25
2003	174	129	134	129.84	2025	1764	1950.11
2004	180	133	140	134.60	2209	1600	2061.16
2005	200	137	146	139.36	3968	1600	3677.21
2006	135	141	158	144.12	36	529	83.17
2007	120	145	164	148.88	625	1936	834.05
2008	105	149	170	153.64	1937	4225	2365.85
Total					15893	17412	15786.77

From the above calculations it is clear that SSR is minimum for least squares method, followed by free hand curve method. The SSR of these two

methods are closer due the fact that the trends in original time series data are not linear, but linear trend line is assumed in both methods.

Example 14.7.7. Determine the trend components of tea production given in following table by (i) semi-average method and (ii) least squares method. Also (iii) calculate the expected production of tea for 2000 by both methods and (iv) comment on the methods.

Year	Production of Tea (in 000 quintal)	Year	Production of Tea (in 000 quintal)	Year	Production of Tea (in 000 quintal)
1975	236	1983	203	1991	256
1976	213	1984	215	1992	304
1977	180	1985	280	1993	291
1978	163	1986	351	1994	277
1979	180	1987	320	1995	274
1980	187	1988	370	1996	272
1981	210	1989	366		
1982	237	1990	325		

Solution. (i) Trends by semi-average method.

There are 22 years, the average of first 11 years (1975 - 1985) is 209.45 which is placed corresponding to the middle of this period 1980, and the average of last 11 years (1986 - 1996) is 309.64 and is placed corresponding to 1991. If we draw a straight line joining these two points, we shall get the required trend line, and the points on the line will provide estimated trend values. Here, it is found that the per year increment over the 11 years from 1981 to 1991 is $(309.64 - 209.45)/11 = 9.11$ thousand quintals (approx). Hence, the trend value of 1981 is obtained by adding 9.11 to 209.45, similarly, 9.11 is added to the previous year for every subsequent year in order to obtain trend values for other years. On the other hand, for trend value of every previous year of 1980, the factor 9.11 is subtracted from the estimated trend of following year. For example, if 9.11 are subtracted from 209.45, the trend value for 1979 is obtained, and so on. The trend values for different year obtained by semi-average method are shown in table 14.7 and figure 14.9.

(ii) Trends by Least Squares method.

Here the total number of periods is 22, which is even. The average of the periods is 1985.5. Thus, coded variable (x) is obtained by subtracting mean year 1985.5 from every year and multiplied the difference by 2, that means $x = (\text{Year} - 1985.5) \times 2$.

Now, we have to fit the trend line $y = a + bx$, where the formula for a and b are given by $a = \bar{y}$ and $b = \frac{\sum xy}{\sum x^2}$ respectively.

From the table 14.7, we get $\sum y = 5710.00$, $\sum xy = 11014.00$ and $\sum x^2 = 3542.00$. Using the formula mentioned above, we get $a = 259.55$ and $b = 3.11$. So, the fitted trend line is $\bar{y} = 295.55 + 3.11 x$.

The estimated trend values for different years are calculated using different values of x and shown in table 14.7 and plotted in figure 14.9.

Table 14.7. Calculations of Trend values of production by semi-average and Least Squares methods and sum of squares of errors.

Year	y	Semi-Average Trend	x	x square	xy	Least squares Trends	SSR for Semi Average	SSR for LS Trend
1975	236	163.90	-21.00	441.00	-4956.00	194.24	5197.75	1743.90
1976	213	173.01	-19.00	361.00	-4047.00	200.46	1598.84	157.25
1977	180	182.12	-17.00	289.00	-3060.00	206.68	4.51	711.82
1978	163	191.23	-15.00	225.00	-2445.00	212.90	797.19	2490.01
1979	180	200.34	-13.00	169.00	-2340.00	219.12	413.90	1530.37
1980	187	209.45	-11.00	121.00	-2057.00	225.34	504.21	1469.96
1981	210	218.56	-9.00	81.00	-1890.00	231.56	73.35	464.83
1982	237	227.67	-7.00	49.00	-1659.00	237.78	86.96	0.61
1983	203	236.78	-5.00	25.00	-1015.00	244.00	1141.40	1681.00
1984	215	245.89	-3.00	9.00	-645.00	250.22	954.47	1240.45
1985	280	255.00	-1.00	1.00	-280.00	256.44	624.77	555.07
1986	351	264.11	1.00	1.00	351.00	262.66	7549.08	7803.96
1987	320	273.22	3.00	9.00	960.00	268.88	2187.94	2613.25
1988	370	282.33	5.00	25.00	1850.00	275.10	7685.23	9006.01
1989	366	291.44	7.00	49.00	2562.00	281.32	5558.52	7170.70
1990	325	300.55	9.00	81.00	2925.00	287.54	597.58	1403.25
1991	256	309.64	11.00	121.00	2816.00	293.76	2876.86	1425.82
1992	304	318.75	13.00	169.00	3952.00	299.98	217.46	16.16
1993	291	327.86	15.00	225.00	4365.00	306.20	1358.39	231.04
1994	277	336.97	17.00	289.00	4709.00	312.42	3595.96	1254.58
1995	274	346.08	19.00	361.00	5206.00	318.64	5195.00	1992.73
1996	272	355.19	21.00	441.00	5712.00	324.86	6919.97	2794.18
Total	5710		0.00	3542.00	11014.00		55139.36	47756.95

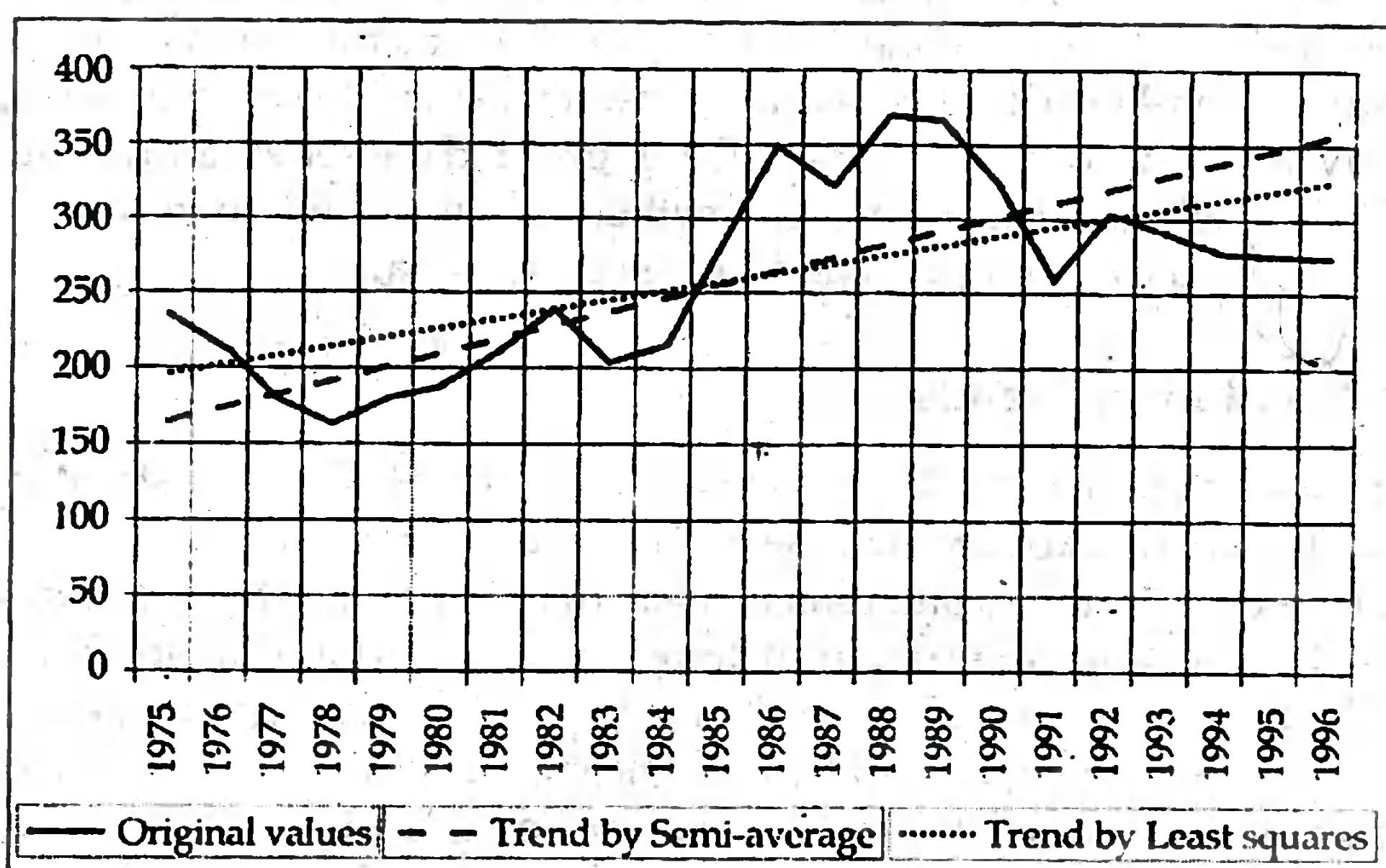


Fig. 14.9. Trend values obtained by semi-average and least squares method.

(iii) Forecast of production for 2000.

By semi-average method. We have found that average increment of production of every year is 9.11 thousand quintals, and the estimated trend by semi-average method for 1996 is 355.19 thousand quintals. Thus, if we add 9.11 quintals with 355.19, the estimated trend value for 1997 will be obtained. In similar way, in order to estimate the production for 2000, we have to add four times of 9.11 with trend value of 1996. Thus, forecasted production for 2000 is given by $355.19 + 4 \times 9.11 = 391.63$ thousand quintals. This value can also be obtained by extending the trend line for semi average method up to 2000.

By Least squares method. We have already obtained the fitted trend line by least squares method as $\bar{y} = 295.55 + 3.11x$.

Now, the coded value of x for 2000 is 29. Putting this value of x in the trend line we get

$$\bar{y}_{2000} = 295.55 + 3.11 \times 29 = 349.74 \text{ thousand quintals.}$$

(iv) For comparison of these two methods of determining trend values, we have to calculate sum of squares of errors or residuals (SSR) for both the models. The SSR is defined as $SSR = \sum (y - \bar{y})^2$.

The SSR values for Semi-average method is 55139. 36 and that for least squares method is 47756.95 as calculated in table 14.7.

So, based on the values of SSR it can be concluded that the least squares trend model is better than semi average method.

It is to be noted here that Free hand curves simply describe the given data values, this is a non-algebraic method and this method is also used for measuring non-linear trends, while semi-average and least squares help to identify a trend equation to describe a given data series and these are algebraic methods. Least squares method is also used to measure the parabolic trend assuming a second-degree polynomial.

14.8. Non-Linear Trends

In previous section and sub-section we dealt with the linear or straight-line trends, which indicate the increase or decrease of a time series at a constant rate. However, in many situations, linear trends may not be found in time series data, instead, the nature of trend may be curvilinear. In that case, straight line cannot express the data adequately. This type of pattern of trends can be better expressed by non-linear curve. However, the following methods of measuring non-linear trends are being used:

- i) Free-hand curve or graphic method
- ii) Moving average method
- iii) Least squares method using parabolic trend equation

14.8.1. Free hand curve method for non-linear trends. We have already discussed the method of free hand curve for linear trends. For non-linear trend, the method is the same, only the difference is that in this case it is necessary to draw free hand non-linear curve instead of a linear trend line, which is a bit difficult. However, it is also stated that free hand curve method involves an element of subjectiveness and it is not recommended for general use, so this method will not be discussed in details here.

14.8.2. Moving Average (MA) Method. A mathematical method of determining non-linear trend of a time series data is the moving average method. A moving average for a time period is the simple arithmetic mean of the values in that time period. The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend. The moving average method is also used for smoothing the irregularity present in a time series data. A note about smoothing the data is provided in section 14.12.

The simplest technique of moving average is called a simple centered $n = (2m + 1)$ -period moving average. The idea here is to replace each actual observation Y_t by the average of itself and its m neighbours on either side that means, replace Y_t by

$$MA_t(2m+1) = Y_t = \frac{1}{2m+1} \sum_{j=-m}^m Y_{t+j}$$

$$= \frac{Y_{t-m} + Y_{t-m+1} + \dots + Y_t + \dots + Y_{t+m-1} + Y_{t+m}}{2m+1}$$

It is to be noted that here $t = m + 1$

The moving average \bar{Y}_t is said to be centered because \bar{Y}_t is the central trend value of $(2m+1)$ observation considered for calculating the moving average, i.e., \bar{Y}_t is MA corresponding to $(m+1)$ th observation of $(2m+1)$ observations, and there are same number of observations in the right and left of \bar{Y}_t and $2m+1$ is always odd for any real number of m . For example, suppose we set $m = 2$, so that $n = 5$ and a 5-period moving average is formed, we then have

$$MA_t(5) = \bar{Y}_t = \frac{Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1} + Y_{t+2}}{5}$$

However, if we consider the first available observation as Y_1 , the first moving average of 5-period would provide value for 3rd period, defined as

$$MA_3(5) = \bar{Y}_3 = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5}{5}$$

This is, of course, just the average of the first five observations and the trend corresponding to the 3rd period. Similarly, \bar{Y}_4 is the average of second through sixth observations, that means,

$$MA_4(5) = \bar{Y}_4 = \frac{Y_2 + Y_3 + Y_4 + Y_5 + Y_6}{5} \text{ and so on.}$$

In this method, first the period of moving average is to be selected depending on the period of data. For example, for the quarterly data, the period of MA is considered as 4, for monthly data it is 12, and for yearly data, it is usually selected on the basis of the length of cycle. This method is commonly applied to the data, which are attributed by cyclical movements. For yearly data, if the selected period of moving average is not equal to the length of cycle, the effect of cyclical variation will not be removed entirely. Often we find the cycles in the data are not of uniform length, in such case, it required to consider the period of moving average equal to or somewhat greater than the average period of cycle in the data. Usually, the necessary period will range between three to ten years for general business series. However, some advanced statistical methods are also available for proper selection of the period of MA.

MA for Odd and Even number of periods

If the chosen period of length of MA is an odd number and given by $n = 2m + 1$, the moving average is centered in the middle $(m + 1)$ th period of the consecutive sequence of n -period values used to compute it, and there will be m observations in either side of it. In this case no MA can be obtained for the first $(n - 1)/2$ periods and for the last $(n - 1)/2$ periods of the series. For

example, if $n = 5$, then, $m = 2$, then it is required to calculate MA using 5 observations which will give the trend value corresponding to $(m + 1)$ th or 3rd period, and in this case MA can not be obtained for the first 2 and last 2 periods.

Again, when the chosen period of length of MA is an even number, the calculated MA is to be placed in between two middle periods. This type of MA falling between two time periods causes various problems, particularly - it does not give trend corresponding to any particular period of the given periods which causes difficulty in graphing. Centering MAs is only the remedy of this problem. This is performed by calculating the 2-period MA of moving averages and placing in mid-period of these two periods, which gives the trend corresponding to a time period of given time span. These types of averages are called the centered moving averages. For example, if $n = 4$, first we have to calculate MA for the first 4 observations and place it between 2nd and 3rd period. The next MA is to be calculated using the observations from 2nd to 5th, and place it between 3rd and 4th period and so on. Thus the centered MA is to be calculated using the first two MAs, and this is placed corresponding to 3rd period, the next 2-period average of MAs is placed corresponding to the 4th period, and so on. In this way, MA trends for the feasible periods can be calculated. If the period of MA is even, say, n , then the MAs cannot be obtained for the first $n/2$ periods and the last $n/2$ periods.

The matter of moving averages for odd and even number of periods are explained below with examples.

Example 14.8.1. Calculate (i) 3-yearly and (ii) 4-yearly moving averages of the number of students studying in MBA in different years as given below.

Year	No. of students	Year	No. of students
1981	332	1986	405
1982	317	1987	410
1983	357	1988	427
1984	392	1989	405
1985	402	1990	438

Solution. (i) 3-yearly averages are calculated using observations of consecutive three years and placed corresponding to the second period of each of three-year periods. For this, at first, three-yearly totals are calculated for each consecutive three years, then averages are calculated dividing these totals by 3. These averages are placed corresponding to the central year of the three years, which are required moving averages. Here, the total number of students for the first three years 1981, 1982, 1983 is 1006 and simple average is 335, this value gives the moving average for the year 1982.

MAs for first period 1981 and last period 1990 cannot be computed. The MAs for the given data are shown in table 14.8.

Table 14.8. 3-yearly and 4-yearly moving averages of number of students.

Year	No. of Students	3-yearly Total	3-yearly M.A.	4-yearly Total	4-yearly M.A.	4-yearly Centered M.A.
1981	332					
1982	317	(332+317+357)=1006	335	(332+317+357+392) = 1398	349.5	
1983	357	(317+357+392)=1066	355	(317+357+392+402) =1468	367	358.25
1984	392	(357+392+402)=1151	384	(357+392+402+405) =1556	389	378
1985	402	(392+402+405)=1199	400	(392+402+405+410) =1609	402.25	395.625
1986	405	(402+405+410)=1217	406	(402+405+410+427) =1644	411	406.625
1987	410	(405+410+427)=1242	414	(405+410+427+405) = 1647	411.75	411.375
1988	427	(410+427+405)=1242	414	(410+427+405+438) =1680	420	415.875
1989	405	(427+405+438)=1270	423			
1990	438					

(ii) First 4-yearly moving total and moving average for the years 1981, 1982, 1983, 1984 are calculated as 1398 and 349.5 respectively and are placed at the center of time span for which these are computed which falls between two given time periods viz. between 1982 and 1983, second 4-yearly moving total and moving average for the next four years 1982, 1983, 1984, 1985 are calculated as 1468 and 367 and are placed between 1983 and 1984. In this way all moving averages are calculated. Again, the centered moving averages are calculated as the 2-period moving average of these 4-yearly MAs. First 4-yearly centered moving average is obtained as the average of 349.5 and 367, and placed between this time spans, corresponding to year 1983. In this way all the 4-yearly and centered moving averages are calculated and presented in table 14.8. Since we have calculated the 4-yearly

moving averages, we will not get the moving averages for the first two years 1981 and 1982, and for the last two years 1989 and 1990.

Example 14.8.2. Calculate 5-yearly and 6-yearly moving average for the following data of a number of commercial industrial failures in a country during 1994 to 2005.

Year	No. of failures	Year	No. of failures
1994	20	2000	11
1995	12	2001	14
1996	12	2002	12
1997	10	2003	9
1998	9	2004	7
1999	13	2005	5

Solution. The required 5-yearly and 6-yearly MAs are calculated using the method as described in the previous Example, and shown in the table 14.9.

Table 14.9. 5-yearly and 6-yearly moving averages of number of failures.

Year	No. of failures	5-yearly moving total	5-yearly moving average	6-yearly moving total	6-yearly moving average	6-yearly centered-moving average
1994	20					
1995	12					
1996	12	63	12.6	76	12.67	
1997	10	56	11.2	67	11.17	11.92
1998	9	55	11	69	11.50	11.33
1999	13	57	11.4	69	11.50	11.50
2000	11	59	11.8	68	11.33	11.42
2001	14	59	11.8	66	11.00	11.17
2002	12	53	10.6	58	9.67	10.33
2003	9	47	9.4			
2004	7					
2005	5					

Trends for first 2 periods and last 2 periods can not be obtained in case of 5-yearly moving average method, and trends for first 3 periods and last 3 periods can not be obtained in case of 6-yearly moving average method.

Example 14.8.3. Determine short-term fluctuations from the data given below.

Year	Seasons			
	Summer	Monsoon	Autumn	Winter
1999	20	70	62	110
2000	22	60	80	162
2001	40	142	88	218
2002	55	160	122	226
2003	65	190	135	300

Solution. The short-term fluctuations are computed using 4-quarterly moving averages and shown in following table.

Table 14.10. Computations of the short-term fluctuations.

Year	Season	Observations	4-quarterly moving total	Paired Sum of 4-quarterly moving total	Centered 4-quarterly M.A. Col.(5)/8	Fluctuations Col.(2)-Col.(7)
(1) 1999	(2) Summer	(3) 20	(4) 262	(5) 526	(6) 65.75	(7) -3.75
	Monsoon	70				
	Autumn	62				
	Winter	110				
2000	Summer	22	(4) 302	(5) 586	(6) 73.25	(7) -51.25
	Monsoon	90				
	Autumn	80				
	Winter	162				
2001	Summer	40	(4) 432	(5) 856	(6) 107	(7) -67.00
	Monsoon	142				
	Autumn	88				
	Winter	218				
2002	Summer	55	(4) 555	(5) 1076	(6) 134.5	(7) -79.50
	Monsoon	160				
	Autumn	122				

	Winter	226		1176	147	79.00
2003	Summer	65	603	1219	152.375	-87.38
	Monsoon	190	616	1306	163.25	26.75
	Autumn	135	690			
	Winter	300				

The fluctuations in column (7) include both regular and irregular fluctuations. The following table shows the seasonal variations, which is nothing but separation of the regular fluctuations from the total fluctuations.

Table 14.11. Calculations of Regular seasonal fluctuations.

Year	Seasons			
	Summer	Monsoon	Autumn	Winter
1999			-3.75	41.50
2000	-51.25	8.00	-10.75	62.50
2001	-67.00	27.00	-35.88	90.00
2002	-79.50	20.25	-20.00	79.00
2003	-87.38	26.75		
Total	-285.13	82.00	-70.38	273.00
Average	-71.28	20.50	-17.60	68.25

Thus, the seasonal variations for different quarters are: Summer -71.28, Monsoon +20.50, autumn -17.60, Winter +68.25.

Merits and limitations of moving average method

The following are the merits and limitations of this method.

Merits:

- i) This method is simple as compared to the method of least squares.
- ii) It is a flexible method of measuring trend. If a few more figures are added to the data, this method does not need to calculate the trend afresh; only few more trend values are to be calculated.
- iii) If the period of moving average happens to coincide with the period of cyclical fluctuations in the data, such fluctuations are automatically eliminated.
- iv) The moving average has the advantage that it follows the general movements of the data and that its shape is determined by the data rather than the statistician's choice of mathematical function.

- v) Most of the components of time series data can be eliminated by this method, particularly; the irregular components can be completely eliminated by this method.

Limitations:

- i) The main limitation of this method is that trend values cannot be computed for all the periods. The longer the period of moving average, the greater the number of periods for which trend values cannot be obtained.
- ii) The second important limitation of this method is that a mathematical model does not represent the method; hence this method cannot be used in forecasting which is one of the major objectives of trend analysis.
- iii) Moving averages are highly affected by the extreme values.
- iv) Because no hard and fast rules are available for the choice of the period, one has to use his own judgment which may incur subjective error.
- v) Although theoretically we can say that if the period of moving average happens to coincide with the period of cycle, the cyclical fluctuations are completely eliminated, but in practice since the cycles are by no means perfectly periodic, no moving average can remove the cycles
- vi) If there is no linear relationship between the time series data, this method is unsafe to determine the trend values.

When to use MA method for non-linear trend

The moving average method of analyzing non-linear trend of time series data is recommended under the following situations:

- i) When the purpose of investigation does not call for current analysis or forecast.
- ii) When the trend line is non-linear.
- iii) The cyclical variations are regular both in period and amplitude.

However, in practice, these conditions rarely hold true.

Other forms of non-linear trends

We have already discussed the least squares method of determining parabolic form of secular trend and moving average method of determining nonlinear trend. Some of the plausible forms of nonlinear trend of a time series are as follows:

- i) Second degree parabola : $y = a + bx + cx^2$
- ii) Third degree polynomial trend : $y = a + bx + cx^2 + dx^3$
- iii) Exponential trend : $y = ab^x$

- iv) Modified exponential trend : $y = a + bc^x$
- v) Gompertz type of trend : $y = a \cdot bc^x$
- vi) Logistic trend : $y = \frac{k}{1 + e^{a+bx}}$

In all of the above mathematical forms, x denotes the transformed or coded series of time. The last three are also known as growth curves. That means, if the trend component of a time series show increasing growth or decline, then the trend can be efficiently determined using these models.

It is to be noted here that, least squares method is usually used for fitting the above mentioned curves. Because, curve fitting by the least squares method is an important and popular method of determining trend component, particularly, when interest lies in making proper projection or forecasting for future period. The appropriate form of mathematical function is generally determined by plotting the time series graphically or from previous experience, and the reliability of the projection depends on choice of appropriate curve.

However, here we shall discuss how the trend components of quadratic or second-degree parabola form can be determined by using least squares method.

14.8.3. Least squares Trends using parabolic trend model. The curvilinear relationship for estimating the trends \bar{y} of a time series data for transformed time variable x is given by $\bar{y} = a + bx + cx^2$.

This type of trend line is called parabola. The least squares method of determining the values of the constants a , b and c produces following three normal equations:

$$\begin{aligned}\Sigma y &= na + b \Sigma x + c \Sigma x^2 \\ \Sigma xy &= a \Sigma x + b \Sigma x^2 + c \Sigma x^3 \\ \Sigma x^2 y &= a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4\end{aligned}$$

Since the time variable (x) is coded in such a way that $\Sigma x = 0$ and $\Sigma x^3 = 0$, we can easily drop out two terms from the above expressions and the expressions reduce to:

$$\begin{aligned}\Sigma y &= na + c \Sigma x^2 \\ \Sigma xy &= b \Sigma x^2 \\ \Sigma x^2 y &= a \Sigma x^2 + c \Sigma x^4\end{aligned}$$

Thus the estimated values of constants in the parabola are given by:

$$a = \frac{\Sigma y - c \Sigma x^2}{n}; b = \frac{\Sigma xy}{\Sigma x^2} \text{ and } c = \frac{n \Sigma x^2 y - \Sigma x^2 \Sigma y}{n \Sigma x^4 - (\Sigma x^2)^2}$$

The estimated values of $\bar{y} = a + bx + cx^2$ are the required trend values that exist in a given time series data.

Example 14.8.4. The prices (in Taka) of a commodity during 1994 -99 are given below. Fit a parabola to these data and estimate the price of the commodity for the year 2000.

Year	1994	1995	1996	1997	1998	1999
Price	100	107	128	140	181	192

Also plot the values on graph paper.

Solution. We have to fit the parabola $\bar{y} = a + bx + cx^2$, where y is the price variable and x is the scaled variable of year. Here, for coded or scaled x variable, we have $\sum x = 0$ and $\sum x^3 = 0$. Thus, the values of constants a , b , and c are to be computed using the following formulae:

$$a = \frac{\sum y - c \sum x^2}{n}; b = \frac{\sum xy}{\sum x^2} \text{ and } c = \frac{n \sum x^2 y - \sum x^2 \sum y}{n \sum x^4 - (\sum x^2)^2}$$

The calculations needed to compute the values of constants are shown in Table 14.11.

Table 14.11. Calculations for parabolic trend line.

Year	Scaled Time x	Price y	x^2	x^4	xy	$x^2 y$	Trend values
1994	-5	100	25	625	-500	2500	97.78
1995	-3	107	9	81	-321	963	110.4
1996	-1	128	1	1	-128	128	126.62
1997	1	140	1	1	140	140	146.44
1998	3	181	9	81	543	1629	169.86
1999	5	192	25	625	960	4800	196.88
Total	0	848	70	1414	694	10160	

Using the calculations shown in above table, we get, $c = 0.45$, $b = 9.91$ and $a = 136.08$, hence the fitted parabolic trend equation is given by

$$\bar{y} = 136.08 + 9.91 x + 0.45 x^2$$

The trend values are calculated for different values of x and also shown in the table 14.11.

Now for the year 2000, the scaled value of x becomes 7, substituting $x = 7$ in fitted trend line, we have the projected price for 2000 which is given by

$$\bar{y}_{2000} = 136.08 + 9.91 \times 7 + 0.45 \times 7^2 = \text{Taka } 227.50$$

The observed values of price and the estimated parabolic trend values are shown in figure 14.10.

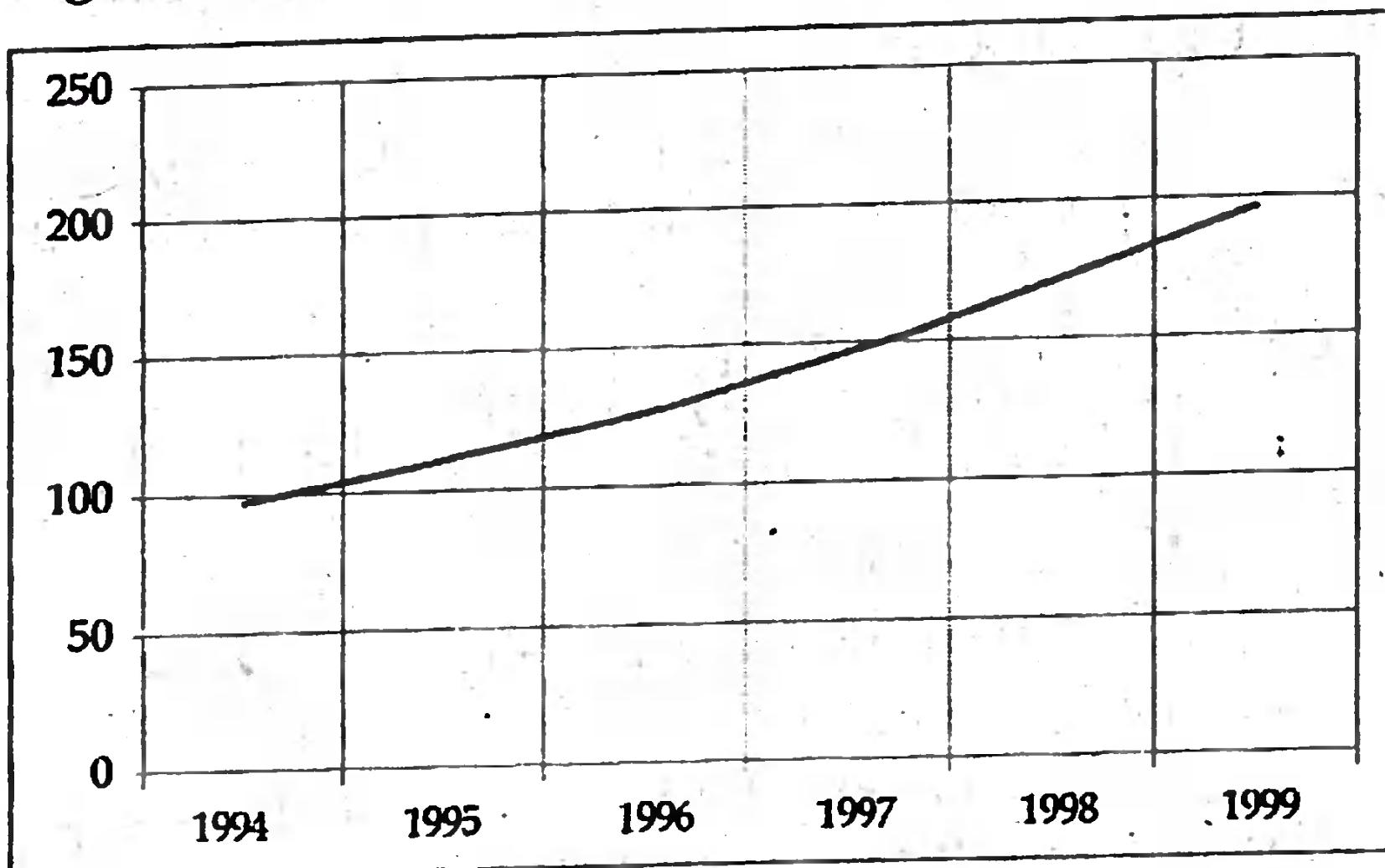


Fig.14.10. Parabolic trend values of price.

Example 14.8.5. The following are the annual profits in thousand of taka in a certain business. Use the method of least squares to determine the trend values of profit. Also forecast the annual profits for 2000 (The same data have been used to fit linear trend line in example 14.7.5).

Year :	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Profit :	60.	72	75	65	80	85	95	101	107	115

Solution. We have to fit the parabola $\bar{y} = a + bx + cx^2$, where y is the profit variable and x is the scaled variable of year defined as $x = \frac{\text{year} - 2003.5}{0.5}$

Here, for coded or scaled x variable, we have $\sum x = 0$ and $\sum x^3 = 0$. Thus, the values of constants a , b , and c are to be computed using the following formulae:

$$a = \frac{\sum y - c \sum x^2}{n}; b = \frac{\sum xy}{\sum x^2} \text{ and } c = \frac{n \sum x^2 y - \sum x^2 \sum y}{n \sum x^4 - (\sum x^2)^2}$$

The calculations needed to compute the values of constants are shown in Table 14.12.

Table 14.12. Calculations for parabolic trend line of profit.

Year	Scaled Time x	Profit y	x^2	x^4	xy	x^2y	Trend values
1999	-9	60	81	6561	-540	4860	63.87
2000	-7	72	49	2401	-504	3528	66.83
2001	-5	75	25	625	-375	1875	70.51
2002	-3	65	9	81	-195	585	74.91
2003	-1	80	1	1	-80	80	80.03
2004	1	85	1	1	85	85	85.87
2005	3	95	9	81	285	855	92.43
2006	5	101	25	625	505	2525	99.71
2007	7	107	49	2401	749	5243	107.71
2008	9	115	81	6561	1035	9315	116.43
Total	0	855	330	19338	965	28951	

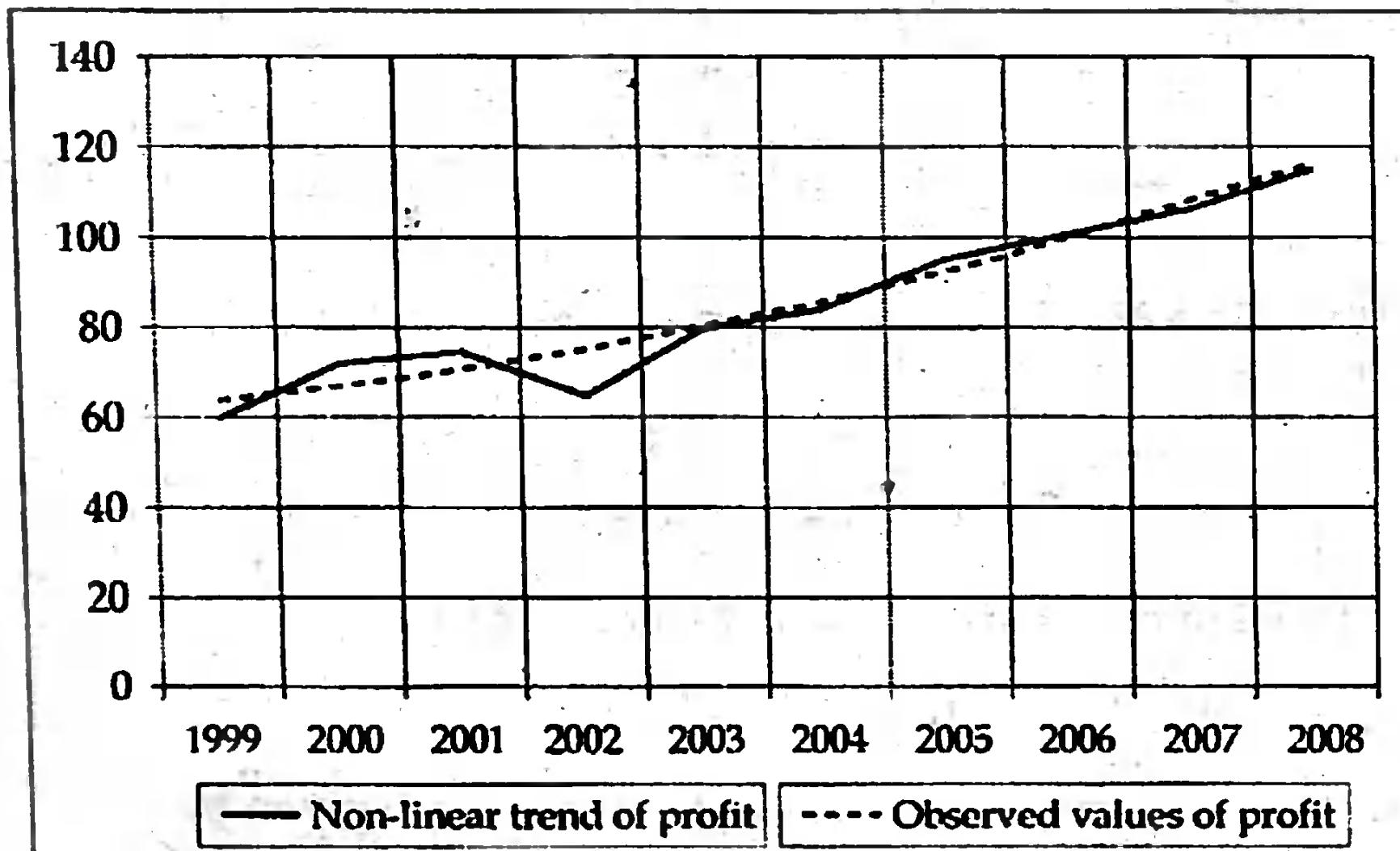
Using the calculations shown in above table, we get, $c = 0.09$, $b = 2.29$ and $a = 82.86$, hence the fitted parabolic trend equation is given by

$$\bar{y} = 82.86 + 2.92x + 0.09x^2$$

Now for the year 2009, the scaled value of x becomes 10, substituting $x = 10$ in fitted trend line, we have the projected profit for 2009 as

$$\bar{y}_{2009} = 82.86 + 2.92 \times 10 + 0.09 \times 11^2 = \text{Taka } 123.00 \text{ thousand (approximately)}$$

The observed values of price and the estimated parabolic trend values are also shown in figure 14.11.

**Fig. 14.11. Parabolic trend of profit.**

Example 14.8.6. The number of students applied in BBA course at a private university increased dramatically from 2005-2010, the data are given below:

Year	2005	2006	2007	2008	2009	2010
No. of students	50	110	350	1020	1950	3710

- i) Develop a linear estimating equation to describe the data
- ii) Develop a second-degree equation
- iii) Estimate the number of students that will be expected to apply in 2012 using both equations
- iv) Comment on the models.

Solution. (i) We have to fit the linear trend equation $y = a + bx$, and where y represents the number of students and x represents the coded year. The values of a , b and c are to be computed using the formula

$$a = \bar{y}, \quad b = \frac{\sum xy}{\sum x^2}$$

Let us construct the following table for necessary calculations of sum of squares and sum of products needed for both requirements.

Table 14.13. Calculations of sum of squares and sum of products for linear and second-degree equation.

(The last two columns of the table are predicted values, to be computed after finding the values of parameters a , b , c using the values obtained in 3rd to 7th columns.)

Year	$x = \frac{\text{Year} - 2007.5}{0.5}$	y	xy	x^2	x^2y	x^4	Prediction by linear trend	Prediction by non-linear trend
2005	-5	50	-250	25	1250	625	-550.97	119.3
2006	-3	110	-330	9	990	81	148.75	14.7
2007	-1	350	-350	1	350	1	848.47	312.26
2008	1	1020	1020	1	1020	1	1548.19	1011.98
2009	3	1950	5850	9	17550	81	2247.91	2113.86
2010	5	3710	18550	25	92750	625	2947.63	3617.9
Total	0	7190	24490	70	113910	1414		

Substituting the values of sum of squares and sum of products obtained from the table in formula, we have

$$a = \bar{y} = 1198.33, \quad b = \frac{\sum xy}{\sum x^2} = \frac{24490}{70} = 349.86$$

Thus, the fitted linear model is $y = 1198.33 + 349.86x$

(ii) The second degree equation $y = a + bx + cx^2$,

We know, the least squares estimates of a , b and c are given by,

$$a = \frac{\sum y - c \sum x^2}{n}; \quad b = \frac{\sum xy}{\sum x^2} \quad \text{and} \quad c = \frac{n \sum x^2 y - \sum x^2 \sum y}{n \sum x^4 - (\sum x^2)^2}$$

Substituting the values of sum of squares and sum of products in formula, we have, $c = 50.26$, $b = 349.86$ and $a = 611.88$.

(iii) In order to estimate the expected number of students to be applied in 2012, we have to use the coded value x of 2012 which is 7. Hence, the estimated numbers of laptops are:

For linear model: $y = 1198.33 + 349.86(7) = 3467$ pieces

For parabolic model: $y = 1198.33 + 349.86(7) + 50.26(7)^2 = 5524$ pieces.

(iv) An increasing trend in the number of students are observed in given data. If we assume that the number of students would be increased by 2012, and hence the increasing trend of number of students will continue till 2012. The parabolic model provides more realistic forecast than that obtained by linear trend model (which is not acceptable, because it is rather less than that used in 2010), hence the performance of parabolic model is better.

14.9. Measurement of Seasonal Variation

We have already discussed how to measure or predict the trend component of a time series data. There are various reasons for studying or analyzing or measuring the seasonal component of a time series data too. Some of the important reasons are mentioned below.

Reasons for Studying Seasonal Component

- i) It allows us to establish the pattern of past changes.
- ii) It is useful to project past patterns for the future.
- iii) Once the seasonal pattern that exists in a time series data has been established, it is possible to eliminate its effect from the data.

In this context the following points are notable:

- i) In an additive time series model, where the cyclical and irregular components are absent, the seasonal component is given by the difference between actual data values and the trend values.
- ii) In a multiplicative time series model, detrending (making data free of trend component) of a time series means the time series which is free from trend component. Detrending of a time series is made dividing actual data series by the estimated trend components, that means.

$$\frac{Y}{T} = S \times C \times I$$

- iii) The process of eliminating seasonal variation from a time series is known as the deseasonalization or seasonal adjustment. Deseasonalization of a time series is made dividing actual data series by the estimated seasonal components, that means.

$$\frac{Y}{S} = T \times C \times I$$

Seasonal Index

Seasonal variations of a time series data are measured in terms of an index, called seasonal index, attached to each period of the time series within a year. For example, for monthly data, there are 12 separate seasonal indices, one for each month, for quarterly data, there are 4 separate indices attached to each quarter of a year. A seasonal index is an average that indicates the percentage of an actual observation relative to what it would be if no seasonal variation in a particular period is present.

The following methods are used to measure the seasonal variation in a time series data:

- i) Method of simple averages
- ii) Ratio-to-trend method
- iii) Ratio-to-moving average method
- iv) Link relatives method

14.9.1. Method of Simple averages. This method is also called average percentage method because this method expresses the data of each month or quarter as a percentage of the averages of the year. The steps involved in this method are summarized below:

- i) The first step in computing seasonal index by simple average method is to add the figures of all years separately for each month or quarter.
- ii) The next step is to find the monthly or quarterly averages which can be obtained by dividing the monthly totals by the number of years. Let the average for 12 months be denoted by $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{12}$.
- iii) The next step is to obtain an average of monthly averages by dividing the total of monthly averages by 12, which is called the grand average. That means, calculate

$$\text{Grand average} = \bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_{12}}{12}$$

- iv) Then, compute seasonal indices for different months by expressing monthly averages as percentage of the grand average as follows

Seasonal index for month i

$$= (\text{Monthly average for month } i / \text{Grand average}) \times 100$$

for all $i = 1, 2, \dots, 12$

- v) If sum of these percentages is not 1200, then monthly percentage so obtained are adjusted by multiplying the index obtained in step (iv) by a correction factor $[1200 \div (\text{sum of 12 indices})]$.

Because, it is important to mention here that the average of the indices should always be 100 that means sum of indices should be 1200 for 12 months, and sum should be 400 for 4 quarterly data.

Example 14.9.1. For the following figures of monthly production of a commodity for the years 1997 to 1999, determine monthly seasonal indices using the method of simple averages.

Month	1997	1998	1999	Month	1997	1998	1999
January	12	23	25	July	20	22	30
February	16	22	25	August	28	28	34
March	18	28	35	September	29	32	38
April	18	27	36	October	33	37	47
May	23	31	36	November	33	34	41
June	23	28	30	December	38	44	53

Solution. Computations of seasonal indices by simple average method for the given data are shown in table 14.14.

Table 14.14. Extraction of Seasonal indices of production.

Month	1997	1998	1999	Monthly Total for 3 years	Monthly Averages for 3 years	Percentage of monthly average of grand average or seasonal index
January	15	23	25	63	21	$(21/30) \times 100 = 70$
February	16	22	25	63	21	70
March	18	28	35	81	27	$(27/30) \times 100 = 90$
April	18	27	36	81	27	90
May	23	31	36	90	30	100
June	23	28	30	81	27	90
July	20	22	30	72	24	80
August	28	28	34	90	30	100
September	29	32	38	99	33	110
October	33	37	47	117	39	130
November	33	34	41	108	36	120
December	38	44	53	135	45	150
Total				1080	360	1200

Here, Monthly average = $1080/12 = 90$, Grand Average = $360/12 = 30$.

Therefore, monthly index for January is $(21/30) \times 100 = 70$, indices for other months are calculated in the same way.

Example 14.9.2. The data on prices (in Taka per kg) of a certain commodity during 1995-1999 for different quarters are shown below:

Quarter	Years				
	1995	1996	1997	1998	1999
I	45	48	49	52	60
II	54	56	63	65	70
III	72	63	70	75	84
IV	60	56	65	72	66

Compute seasonal indices by the simple average method.

Solution. Assuming that the trend is absent in the above data, the difference in the averages of various quarters will be due to seasonal changes. The calculations of quarterly indices by simple average method are shown in table 14.15.

Table 14.15. Calculations of quarterly seasonal indices of prices.

Quarter	Year					Total of 5 years	Average of 5 years	Seasonal indices
	1995	1996	1997	1998	1999			
I	45	48	49	52	60	254	50.8	81.61
II	54	56	63	65	70	308	61.6	98.96
III	72	63	70	75	84	364	72.8	116.95
IV	60	56	65	72	66	319	63.8	102.49
Total						1245	249	400.00

Here, Grand average (average of quarterly averages) = $249/4 = 62.25$.

Seasonal index for quarter I = $(50.8/62.25) \times 100 = 81.61$, for quarter II is $(61.6/62.25) \times 100 = 98.96$, seasonal indices for other quarters are also computed in the same way.

Example 14.9.3. The data below give the average quarterly prices of a commodity for four years.

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
2002	40.3	44.8	46.0	48.0
2003	50.1	53.1	55.3	59.5
2004	47.2	50.1	52.1	55.2
2005	55.4	59.0	61.6	65.3

Calculate seasonal index by the method of simple average.

Solution. Assuming that the trend is absent in the above data, the difference in the averages of various quarters (if there is any) would be due to seasonal changes. Thus, the computations of seasonal indices by the method of simple average are provided in following table.

Table 14.16. Computation of seasonal indices.

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
2002	40.3	44.8	46.0	48.0
2003	50.1	53.1	55.3	59.5
2004	47.2	50.1	52.1	55.2
2005	55.4	59.0	61.6	65.3
Total	193.0	207.0	215.0	228.0
Yearly Average	48.25	51.75	53.75	57.00
Seasonal index	91.57	98.21	102.01	108.18

Here, the average of averages is $\bar{x} = \frac{193.0 + 207.0 + 215.0 + 228.0}{4} = 207.5$.

So, the seasonal index for 1st quarter = $\frac{48.25}{52.96} \times 100 = 91.57$, the seasonal index for other quarters is computed in the same way.

Example 14.9.4. The seasonal indices of the sale of readymade garments in a store are given below:

Quarter	Months	Seasonal Index
I	January to March	96
II	April to June	91
III	July to September	80
IV	October to December	133

If the total sales of garments in the first quarter are worth Tk. 150000, determine how much worth of garments of this type should be kept in stock to meet the demand in each of the remaining quarter.

Solution. Seasonal index for quarter I is given as 96 and its worth is Tk 150000. We have to calculate the proportional values for other quarters if 96 is equivalent to 150000. Using very simple algorithm, the worth for quarter II is calculated as $(150000/96) \times 91 = \text{Taka } 142187.5$. For quarter III it is calculated as $(150000/96) \times 80 = \text{Taka } 125000$ and so on. The estimated stocks of readymade garments are shown in following table.

Table 14.17. Estimated stock of readymade garments for different quarters.

Quarter	Months	Seasonal Index	Estimated stock (Tk)
I	January to March	96	150000.00
II	April to June	91	142187.50
III	July to September	80	125000.00
IV	October to December	133	207812.50

Merits and limitations of method of simple averages

Although this is a very simple method of measuring seasonal variations of time series data, it has some limitations. The merits and limitations of this method are cited below:

Merits :

- i) This method is the simplest of all methods of measuring seasonal variations.
- ii) Simple mathematical knowledge is enough to compute seasonal variation by this method.

Limitations:

- i) This method assumes that no definite trend exists, or trend has a very little impact on time series, which is not always justified.
- ii) The effect of cycles may or may not be eliminated by averaging the observations.

14.9.2. Ratio-to-Trend Method. This method of measuring seasonal indices (also called percentage-to-trend method) is relatively simple and yet an improvement over the simple average method. This is because; in this method it is assumed that seasonal variation for a given period (month or quarter) is a constant fraction of trend. The ratio-to-trend method seemingly isolates the seasonal factor when trend effect is eliminated from the original

time series Y using the ratio $\frac{Y}{T} = S \times C \times I$.

Random elements are supposed to disappear when the ratios are averaged. However, the steps of this method are as follows:

- i) Find the yearly average of time series per month or quarter. For this at first add up the values of all twelve months or four quarters for every year, and then divide the total by 12 or 4 depending on the data given as monthly or quarterly.
- ii) The second step is to compute the trend values for each year applying least squares method.
- iii) The next step is to divide the original data by the corresponding trend values and to multiply these ratios by 100. The values so obtained are free from trend and the problem that remains is to make them again free from cyclical and irregular components.
- iv) Arrange the percentage data values obtained in previous step according to months or quarters for various years.
- v) Find monthly (or quarterly) averages of figures arranged in step (iv) with any one of the usual measures of central tendency such as arithmetic mean, median or mode. Usually arithmetic mean is used in this case.
- vi) The final step is to find the grand average of monthly averages found in step (v). If the grand average is 100, then the monthly averages represent seasonal indices, otherwise, an adjustment is made by multiplying each index by an adjustment factor $[1200 \div (\text{sum of 12 values})]$ to get the final seasonal indices.

Such adjustment is made only to achieve accuracy, but also because when we come to eliminate seasonality from the original data we do not wish to raise or lower the level of data unduly. The logical reasoning behind this method follows from the fact that twelve months moving average can be considered to represent the influence of cycle and trend $C \times T$. If the actual

values for any month are divided by the 12-month moving average entered to that month, presumably cycle and trend are removed. This may be represented by the following expression:

$$\frac{T \times S \times C \times I}{T \times C} = S \times I \text{ which leaves irregular and seasonal influences. Again, if}$$

the ratios for each period of years are then averaged, most random influences will usually be eliminated leaving only the seasonal indices.

Hence, the final output becomes $\frac{S \times I}{I} = S$.

Example 14.9.5. Quarterly sales data (taka in million) in an air-conditioned super market are presented in the following table for a five-year period.

Year	Quarters I	Quarters II	Quarters III	Quarters IV
2002	60	80	72	68
2003	68	104	100	88
2004	80	116	108	96
2005	108	152	136	124
2006	160	184	172	164

Calculate seasonal index for each quarter using the ratio-to-trend method.

Solution. For determining seasonal variations by ratio-to-trend method, at first we have to determine the trend of yearly average data. Calculations to obtain annual trend values from the given quarterly data using method of least squares are shown in table 14.18

Table 14.18. Yearly trend values of sales.

Year	Yearly Total	Yearly Average Col 2 ÷ 4 (y)	x = Year - 2004	x ²	xy	Trend values
2002	280	70	-2	4	-140	64
2003	360	90	-1	1	-90	88
2004	400	100	0	0	0	112
2005	520	130	1	1	130	136
2006	680	170	2	4	340	160
Total		560		10	240	

Here, $\Sigma x = 0$, $\Sigma y = 560$, $\Sigma x^2 = 10$ and $\Sigma xy = 240$

So, $b = \frac{\Sigma xy}{\Sigma x^2}$ and $a = \bar{y} = \frac{\Sigma y}{5}$ which produce $b = 24$ and $a = 112$.

Thus the yearly fitted trend line is given by $y = 112 + 24x$ where the coefficient 24 indicates the yearly increase in sales. Yearly trend values are calculated using different values of x in the fitted trend line and shown in table 14.18.

To calculate quarterly trend values, let us consider the first year 2002. The trend value for this year is 64.3. This is the value for the middle of the year 2002, which is between second quarter and third quarter. Since quarterly increment is $24 \div 4 = 6$, the trend value for the second quarter of 2002 would be $(64 - 6/2) = 61$ and for third quarter it would be $(64 + 6/2) = 67$. Thus the value for first quarter of 2002 would be $61 - 6 = 55$ and that of fourth quarter would be $67 + 6 = 73$. In this way, the quarterly trend values for other years have been calculated and shown in table 14.19.

Table 14.19. Quarterly trend values of sales.

Year	Quarters I	Quarters II	Quarters III	Quarters IV
2002	55	61	67	73
2003	79	85	91	97
2004	103	109	115	121
2005	127	133	139	145
2006	151	157	163	169

After getting the trend values, the given data values in the time series are expressed as percentages of the corresponding trend values in table 7.10c. Thus for the first quarter of 2002, the percentage would be $(60 \div 55) \times 100 = 109.09$, for second quarter it would be $(80 \div 61) \times 100 = 131.15$ and so on. The ratio to trend values for every quarter of each year are calculated and shown in table 14.20.

Table 14.20. Seasonal indices (Ratio-to-trend values).

Year	Quarters I	Quarters II	Quarters III	Quarters IV
2002	109.09	131.15	107.46	93.15
2003	86.08	122.35	109.89	90.72
2004	77.67	106.42	93.91	79.34
2005	85.04	114.29	97.84	85.52
2006	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
Adjusted Seasonal index	92.77×0.99226 $= 92.05$	118.28×0.99226 $= 117.36$	102.92×0.99226 $= 102.12$	89.15×0.99226 $= 88.47$

The total of average of seasonal indices is 403.12, which is more than 400. In order to make this total 400, we have to apply the correction factor $(400 \div 403.12) = 0.99226$. Each quarterly average is multiplied by this factor to get adjusted seasonal index so that its total is 400 as shown in the last row of table.

Example 14.9.6. Consumption of monthly electric power in thousand Kw hours for garments supermarket lighting in Chittagong during 2001-2006 is given below:

Years	Months											
	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct.	Nov	Dec
2001	318	281	278	250	231	216	223	245	269	302	325	347
2002	342	309	299	268	249	236	242	262	288	321	342	364
2003	367	328	320	287	269	251	259	284	309	345	367	394
2004	392	349	342	311	290	273	282	305	328	364	389	417
2005	420	378	370	334	314	296	305	330	356	396	422	452
2006	396	387	360	325	314	270	298	285	311	341	377	362

Calculate the seasonal variations/indices by the ratio-to-trend method.

Solution. For determining seasonal variations by ratio-to-trend method, at first we have to determine the trend of yearly average data. Calculations to obtain annual trend values from the given monthly data using method of least squares are shown in table 14.21.

Table 14.21. Yearly trend values of power consumption.

Year	Yearly Total	Yearly Average Col 2 ÷ 12 (y)	(Year - 2003.5) × 2 = x	x ²	xy	Trend values
2001	3285	273.75	-5.00	25.00	-1368.75	281.03
2002	3522	293.50	-3.00	9.00	-880.50	296.55
2003	3780	315.00	-1.00	1.00	-315.00	312.07
2004	4042	336.83	1.00	1.00	336.83	327.59
2005	4373	364.42	3.00	9.00	1093.25	343.11
2006	4026	335.50	5.00	25.00	1677.50	358.63
Total		1919.00	0.00	70.00	543.33	

From table 14.18 we have, $\Sigma x = 0$, $\Sigma y = 1919.00$, $\Sigma x^2 = 70$ and $\Sigma xy = 543.33$

$$\text{So, } b = \frac{\sum xy}{\sum x^2} \text{ and } a = \bar{y} = \frac{\sum y}{5} \text{ which produce } b = 7.76 \text{ and } a = 319.83.$$

Thus the yearly fitted trend line is given by $y = 319.83 + 7.76x$ where the coefficient 7.76 indicates the yearly increase in power consumption. So, the monthly increment is given by $(7.76 \div 12) = 0.65$ thousand Kw hours. Yearly trend values are calculated using different values of x in the fitted trend line and shown in table 14.21.

To calculate monthly trend values, let us consider the first year 2001. The trend value for this year is 281.03 thousand Kw hours. This is the trend value for the middle of the year 2001 that is between June and July. Since monthly increment is 0.65 thousand Kw hours, the trend value for June of 2001 would be $(281.03 - 0.65/2) = 280.71$ and for July it would be $(281.03 + 0.65/2) = 281.36$. Thus the value for May of 2001 would be $(280.71 - 0.65) = 280.06$ and that of August would be $(281.36 + 0.65) = 282.01$. In this way, the monthly trend values for other years have been calculated and shown in table 14.22.

Table 14.22. Monthly trend values of Power Consumption.

Years	Months											
	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec
2001	277.46	278.11	278.76	279.41	280.06	280.71	281.36	282.01	282.66	283.31	283.96	284.61
2002	292.98	293.63	294.28	294.93	295.58	296.23	296.88	297.53	298.18	298.83	299.48	300.13
2003	308.50	309.15	309.80	310.45	311.10	311.75	312.40	313.05	313.70	314.35	315.00	315.65
2004	324.02	324.67	325.32	325.97	326.62	327.27	327.92	328.57	329.22	329.87	330.52	331.17
2005	339.54	340.19	340.84	341.49	342.14	342.79	343.44	344.09	344.74	345.39	346.04	346.69
2006	355.06	355.71	356.36	357.01	357.66	358.31	358.96	359.61	360.26	360.91	361.56	362.21

After finding the monthly trend values, the given data values in the time series are expressed as percentages of the corresponding trend values in table 14.23. Thus for January of 2001, the percentage would be $(318 \div 277.46) \times 100 = 114.61$, for February of 2001 it would be $(281 \div 278.11) \times 100 = 101.04$ and so on. The ratio to trend values for every quarter of each year are calculated and shown in table 14.23.

Table 14.23. Seasonal indices (Ratio-to-trend values).

Years	Months											
	Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct	Nov	Dec
2001	114.61	101.04	99.73	89.48	82.48	76.95	79.26	86.88	95.17	106.60	114.45	121.92
2002	116.73	105.24	101.61	90.87	84.24	79.67	81.52	88.06	96.59	107.42	114.20	121.28
2003	118.96	106.10	103.29	92.45	86.47	80.51	82.91	90.72	98.50	109.75	116.51	124.82
2004	120.98	107.50	105.13	95.41	88.79	83.42	86.00	92.83	99.63	110.35	117.70	125.92
2005	123.70	111.12	108.56	97.81	91.78	86.35	88.81	95.91	103.27	114.65	121.95	130.38
2006	111.53	108.80	101.02	91.04	87.79	75.35	83.02	79.25	86.33	94.48	104.27	99.94
Total	706.52	639.79	619.34	557.05	521.56	482.26	501.51	533.65	579.49	643.26	689.08	724.27
Average	117.75	106.63	103.22	92.84	86.93	80.38	83.58	88.94	96.58	107.21	114.85	120.71
Adjusted Seasonal Index	Here we do not need any adjustment of averages, because the total of averages is 1199.63 which can be considered as equal to 1200 (requirement for monthly data)											

The total of average of seasonal indices is 1199.63, which is very close to 1200. This fraction of difference from 1200 may remain in the total due to approximation up to 2 decimal points. So, here averages are the required seasonal indices. (However, if it would not happen, that means, if the total would vary from 1200, we would have to multiply each average by the correction factor or adjustment factor as we have done in previous example.)

Merits and limitations of the Ratio-to-trend method

Merits:

- i) This method of measuring seasonal variations is simple to compute and easy to understand.
- ii) Compared with the method of monthly averages this method is certainly a more logical procedure for measuring seasonal variations. It has an advantage over ratio-to-moving average method too because

it has a ratio-to-trend value for each month for which data are given. Thus, there is no loss of data as occurs in case of moving averages.

Limitations:

The main limitation of the ratio-to-trend method is that if there are pronounced cyclical swings in the series, the trend – whether a straight line or a curve – can never follow the actual data as closely as a 12-month moving average does. Consequently, the seasonal variations computed by ratio-to-moving average method may be less biased than one calculated by ratio-to-trend method.

14.9.3. Ratio-to-Moving average Method. The ratio-to-moving average method, also known as the percentage of moving average method, is typically used in order to measure the seasonal variations. This method provides an index that describes the degree of seasonal variation. The index is based on the mean 100, with the degree of seasonality measured by variations away from the base. For example, if we examine the seasonality of tourists at Cox's Bazar, we might find that the winter quarter index, say, 215. This value 215 indicates that 215 percent of average quarterly tourists occur in winter. If total number of tourists recorded in a particular year is 2000, then the average quarterly tourists would be $2000/4 = 500$. Since the winter quarter index is 215, the estimated number of tourists of winter is calculated as $[500 \times (215/100)] = 1075$.

The following steps are followed for determining seasonal variations of a time series data by the method of ratio-to-moving average method:

- i) The first step is to calculate the 4-quarter moving total for the given time series data.
- ii) The second step is to calculate 4-quarter moving average by dividing each of the totals obtained in step (i) by 4.
- iii) In the third step, it is required to center the 4-quarter moving average.
- iv) Next step is to calculate the percentage of the actual value to the moving average value for each quarter in the time series having a 4-quarter entry and list these quotients as 'percent of moving average'.
- v) Arrange the values obtained in step (iv) by quarter. Then calculate the modified mean for each quarter. This is calculated by discarding the highest and lowest values of each quarter and averaging the remaining values. Because, it is assumed that the relatively high or extremely low values occur due to the presence of cyclical or irregular variations. Thus, by eliminating the highest and lowest values from each quarter, extreme cyclical and irregular variations are reduced. Averaging the remaining values does further smoothing. Cyclical and irregular variations tend to be removed by this process, so the modified mean is an index of seasonality component and called the typical seasonal relatives. (Some statisticians prefer to use the median instead of computing the modified mean to achieve the same outcome.)

- vi) In the final step, adjustment of modified mean is done to make the quarterly total of indices 400 or monthly total 1200.

Example 14.9.7. The resort hotel authority wants to study the seasonal pattern of room demanded by its clients to take decision to employ personnel during peak periods. The following table contains the quarterly occupancy that means, the number of guests during each quarter of the last five years.

Year	Quarterly Number of guests			
	I	II	III	IV
2002	1861	2203	2415	1908
2003	1921	2343	2514	1986
2004	1834	2154	2098	1799
2005	1837	2025	2304	1965
2006	2073	2414	2339	1967

Compute the seasonal indices by the method of ratio to moving average method to help the authority.

Solution. At first we have to calculate 4-quarter centered moving averages and then the percentage of actual to moving average values (step i to iv). Then adjustment has to be made considering the modified mean. The calculations up to these percentages are shown in the following table.

Table 14.24. Calculation of centered moving average and percentage of actual values.

Year	Quarter	Occupancy	4-quarter moving total	4-quarter MA	4-quarter centered MA	% of actual values to the MA values.
2002	I	1861	2104.25 2129.25 2159.13 2181.25	114.77 89.61 88.97 107.42	2104.25 2129.25 2159.13 2181.25	114.77 89.61 88.97 107.42
	II	2203				
	III	2415				
	IV	1908				
2003	I	1921	2180.13 2145.63 2070.00 1994.63	115.31 92.56 88.60 107.99	2180.13 2145.63 2070.00 1994.63	115.31 92.56 88.60 107.99
	II	2343				
	III	2514				
	IV	1986				
2004	I	1834	1971.63 1955.88 1965.50 2012.00	106.41 91.98 93.46 100.65	1971.63 1955.88 1965.50 2012.00	106.41 91.98 93.46 100.65
	II	2154				
	III	2098				
	IV	1799				
2005	I	1837	2062.25 2140.38 2193.38 2198.00	111.72 91.81 94.51 109.83	2062.25 2140.38 2193.38 2198.00	111.72 91.81 94.51 109.83
	II	2025				
	III	2304				
	IV	1965				
2006	I	2073				
	II	2414				
	III	2339				
	IV	1967				

These two columns could not be shown because the calculated values correspond to the middle of two quarters. It is hoped that the readers would be able to complete these two columns easily.

The values obtained in above table are arranged according to the quarters of different years as in table 14.25. At this stage the modified means are calculated excluding the two extreme values (smallest and largest values of the quarters). The eliminated values are shown as bold cases.

Table 14.25. Calculation of modified mean (Unadjusted seasonal indices).

Year	Quarterly Number of guests			
	I	II	III	IV
2002			114.8	89.6
2003	89.0	107.4	115.3	92.6
2004	88.6	108.0	106.4	92.0
2005	93.5	100.6	111.7	91.8
2006	94.5	109.8		
Modified total	182.5	215.4	226.5	183.8
Modified mean	91.25	107.70	113.25	91.90

Here the total of modified mean is 404.10 which is not 400 for quarterly data. So, adjustment has to be made in order to make the total 400 and obtain the final seasonal index. The adjustment factor is $(400 \div 404.10) = 0.9899$. The seasonal index are obtained by multiplying each of the modified means by this adjustment factor and shown in following table.

Table 14.26. Adjusted seasonal index for number of guests.

Quarter	Unadjusted index	Adjusted seasonal index
I	91.25	$91.25 \times 0.9899 = 90.3$
II	107.70	106.6
III	113.25	112.1
IV	91.90	91.0
Total	404.10	400.0

It is evident from the seasonal index that the highest guests stay at the hotel in quarter III followed by quarter II.

Example 14.9.8. Find seasonal index from the following table by ratio to moving average method:

Seasons	1999	2000	2001	2002	2003	2004
1st Quarter	40	42	41	45	44	46
2nd Quarter	35	37	35	36	38	37
3rd Quarter	38	39	38	36	38	40
4th Quarter	40	38	42	41	42	39

Solution. The calculation of centered 4-quarterly moving averages and percentage of given values to the moving averages are shown in table 14.27.

Table 14.27. Calculation of centered moving average and percentage of actual values.

Year	Quarter	Occupancy	4-quarter moving total	4-quarter MA	4-quarter centered MA	% of actual values to the MA values
1999	I	40	These two columns could not be shown because the calculated values correspond to the middle of two quarters			
	II	35				
	III	38				
	IV	40				
2000	I	42	It is hoped that the readers would be able to complete these two columns easily			
	II	37				
	III	39				
	IV	38				
2001	I	41				
	II	35				
	III	38				
	IV	42				
2002	I	45				
	II	36				
	III	36				
	IV	41				
2003	I	44				
	II	38				
	III	38				
	IV	42				
2004	I	46				
	II	37				
	III	40				
	IV	39				

The values obtained in above table are arranged as in table 14.23. At this stage the modified means are calculated excluding the two extreme values (smallest and largest values of the quarters). The eliminated values are shown as bold cases.

Table 14.28. Calculation of modified mean (Unadjusted seasonal indices).

Here the total of modified means is 400.96, which can be considered as approximately one. However, for more accurate seasonal index, one can perform the adjustment to make the sum of the seasonal indices exactly one.

The adjusted seasonal index is shown in following table. Here, the adjustment factor is $(400 / 400.96) = 0.9976$. The seasonal index are obtained by multiplying each of the modified means by this adjustment factor as follow:

Table 14.29. Adjusted seasonal index for number of guests.

Quarter	Unadjusted index	Adjusted seasonal index
I	109.91	$109.91 \times .9976 = 109.65$
II	91.96	91.74
III	96.05	95.82
IV	103.04	102.79
Total	400.96	400.00

Merits and Limitations of Ratio-to-Moving average Method.

Merits:

The following are the merits of ratio-to-moving average method.

- i) This method of measuring seasonal variation is considered to be the most satisfactory and more widely used in practice than other methods.
- ii) The index obtained by this method usually does not fluctuate so much as ratio-to-trend method.
- iii) Mathematical methods of avoiding the effects of the business cycle are not usually needed, because, 12-month moving average follows the cyclical course of actual data quite closely. Hence the index obtained by this method is more representative than the ratio-to-trend method.
- iv) Ratio-to-moving average method allows greater flexibility.

Limitations:

We have already mentioned some limitations of moving average. Since ratio-to-moving average method is based on moving averages, so it also bears the same limitations. Hence only limitation of this method is that by this method seasonal variation for all periods can not be obtained. For example, when a 12-month moving average is taken, six months in the top as well as bottom are left out for which we can not calculate seasonal indices.

14.9.4. Link Relative method. Among all the methods of measuring seasonal variation, link relatives method is most difficult one. Link relative

means the percentage change of each period with respect to the immediate previous period to be calculated for each period separately. However, when this method is adopted, the following steps are to be followed.

- i) Calculate link relatives of the given seasonal figures. Link relatives are calculated by dividing the figure of each period by the figure of immediately preceding period and multiplying it by 100. These are called link relatives because they link each period to the preceding one.
- ii) Calculate the average of the link relatives for each season. While calculating average, arithmetic mean may be considered, although median is probably better. Because, the arithmetic mean would give undue weights to the extreme cases which are not primarily due to seasonal fluctuations.
- iii) Convert these averages into chain relatives on the base of first season.
- iv) Calculate the chain relative of the first period on the base of the last period. There will be some difference between the chain relative of the first period and the chain relative calculated by the previous method. This difference will be due to the effect of long-term changes. It is, therefore, necessary to correct these chain relatives.
- v) For correction, the chain relative of the first period calculated by the first method is deducted from the chain relative of the first period calculated by the second method. The difference is divided by the number of periods. The resulting figure multiplied by 1, 2, 3 (and so on) is deducted respectively from the chain relatives of the 2nd, 3rd, 4th (and so on) periods. These would give the corrected chain relatives.
- vi) Express the corrected chain relatives of 2nd, 3rd and 4th quarters as percentage of their averages. These will provide the required seasonal indices by the method of link relatives and the total of seasonal indices would be 400.

The method is illustrated with the following example.

Example 14.9.9. Calculate the seasonal indices by the method of link relatives from the data given below:

Year	Production of rice in million tonnes			
	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2002	60	65	62	69
2003	62	68	65	68
2004	65	70	64	62
2005	70	75	68	67
2006	72	80	70	78

Solution. In order to calculate the seasonal variations for the given quarterly data, at first we have to calculate the link relatives. Then we have to calculate the chain relatives, adjusted chain relatives and finally the

seasonal index. The calculation of link relatives along with chain relatives and seasonal index are shown in table 14.30.

Table 14.30. Calculation seasonal index by the method of Link relatives.

Year	Calculation of Link relatives (L R)			
	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2002		108.3	95.4	111.3
2003	89.9	109.7	95.6	104.6
2004	95.6	107.7	91.4	96.9
2005	112.9	107.1	90.7	98.5
2006	107.5	111.1	87.5	111.4
Total of L R	405.8	544.0	460.6	522.7
Arithmetic Mean	101.45	108.79	92.11	104.55
Chain relatives	100.00	108.79	100.21	104.77
Adjusted Chain Relatives	100.00	107.22	97.07	100.06
Seasonal Index	100.00	105.69	95.68	98.63

Calculation of Link relatives (LR):

Link relative for a quarter = (The figure of that quarter / Previous quarter figure) × 100.

Link relative for 2nd quarter of 2002 is given by $(65/60) \times 100 = 108.3$

For 3rd quarter of 2002 = $(62/65) \times 100 = 95.4$ and so on, similarly,

for 1st quarter of 2006 = $(72/67) \times 100 = 107.5$ and so on.

Calculation of chain relatives:

Chain relative for 1st quarter = 100 by assumption

Chain relative for 2nd quarter = (Average L R for 2nd quarter × C R for 1st quarter) / 100 = $(108.78 \times 100) / 100 = 108.79$

Chain relative for 3rd quarter = (Average L R for 3rd quarter × C R for 2nd quarter) / 100 = $(92.12 \times 108.78) / 100 = 100.21$

Similarly, chain relative for 4th quarter = $(104.55 \times 100.21) / 100 = 104.77$.

Hence, the computed chain relative for 1st quarter becomes $(104.77 \times 101.45) / 100 = 106.29$

The correction factor is calculated as $(106.29 - 100) / 4 = 1.57$

So, Adjusted chain index for 1st quarter = $106.29 - 1.57 = 100$

for 2nd quarter = $108.79 - 1.57 = 107.22$

for 3rd quarter = $100.21 - 2 \times 1.57 = 97.07$

and for 4th quarter = $104.77 - 3 \times 1.57 = 100.06$

Average of chain relatives for three quarters

= $(107.22 + 97.07 + 100.06) / 3 = 101.45$

Finally, the seasonal index for 2nd quarter = $(107.22 / 101.45) \times 100 = 105.69$

Seasonal index for other quarters are also calculated in similar way that means dividing the adjusted chain relatives by 101.45 and multiplying the quotient by 100.

Example 14.9.10. Find seasonal index from the following table by link relatives method.

Seasons	1999	2000	2001	2002	2003	2004
1st Quarter	40	42	41	45	44	46
2nd Quarter	35	37	35	36	38	37
3rd Quarter	38	39	38	36	38	40
4th Quarter	40	38	42	41	42	39

Solution. For the ease of calculation of seasonal index of given data by link relative methods, at first we have to arrange the data as in table 14.31.

The next step is to calculate the link relatives of given values. Then we have to calculate the chain relatives, adjusted chain relatives and finally the seasonal index. The calculation of link relatives along with chain relatives and seasonal index are shown in table 14.32.

Table 14.31. Arrangement of data as Year by Quarter.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1999	40	35	38	40
2000	42	37	39	38
2001	41	35	38	42
2002	45	36	36	41
2003	44	38	38	42
2004	46	37	40	39

Table 14.32. Calculation of seasonal index by the method of link relatives.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1999		87.5	108.6	105.3
2000	105.0	88.1	105.4	97.4
2001	107.9	85.4	108.6	110.5
2002	107.1	80.0	100.0	113.9
2003	107.3	86.4	100.0	110.5
2004	109.5	80.4	108.1	97.5
Total of L R	536.9	507.8	630.7	635.1
Average of L R	107.38	84.63	105.11	105.86
Chain relatives	100.00	84.63	88.95	94.16
Adjusted Chain Relatives	100.00	84.35	88.39	93.32
Seasonal Index	100.00	95.10	99.66	105.22

Calculation of Link relatives (LR):

Link relative for a quarter = (The figure of that quarter / Previous quarter figure) $\times 100$

Link relative for 2nd quarter of 1999 = $(35/40) \times 100 = 87.5$

For 3rd quarter of 1999 = $(38/35) \times 100 = 108.6$

For 1st quarter of 2003 = $(44/41) \times 100 = 107.3$ and so on.

Calculation of chain relatives:

Chain relative for 1st quarter = 100 by assumption

Chain relative for 2nd quarter = (Average L R for 2nd quarter \times C R for 1st quarter)/100 = $(84.63 \times 100)/100 = 108.79$

Chain relative for 3rd quarter = (Average L R for 3rd quarter \times C R for 2nd quarter)/100 = $(105.11 \times 84.63)/100 = 88.95$

Similarly, chain relative for 4th quarter = $(94.16 \times 88.95)/100 = 94.16$

Hence, the computed chain relative for 1st quarter becomes $(107.38 \times 94.16)/100 = 101.11$

The correction factor is calculated as $(101.11 - 100)/4 = 0.28$

So, Adjusted chain index for 1st quarter = $101.11 - 0.28 = 100$
 for 2nd quarter = $84.63 - 0.28 = 84.35$
 for 3rd quarter = $88.95 - 2 \times 0.28 = 88.39$
 and for 4th quarter = $94.16 - 3 \times 0.28 = 93.32$

Average of chain relatives for three quarters = $(84.35 + 88.39 + 93.32)/3 = 88.69$

Finally, the seasonal index for 2nd quarter = $(84.35/88.69) \times 100 = 95.10$

Seasonal index for other quarters are also calculated in similar way that means dividing the adjusted chain relatives by 88.69 and multiplying the quotient by 100. At this stage the total of seasonal index will be 400.

Example 14.9.11. The data below give the average quarterly prices of a commodity for five years. Calculate the seasonal indices by the method of link relatives.

Quarter	Year				
	1996	1997	1998	1999	2000
I	30	35	31	31	34
II	26	28	29	31	36
III	22	22	28	25	26
IV	31	36	32	35	33

Solution. Here data are given as quarter by year. Even then unlike the above example, we can compute the seasonal index using this table instead of arranging them in year by quarter. The computation of seasonal index by link relative method is shown in following table.

Table 14.33. Calculation of seasonal index by the method of link relatives.

Quarter	Calculation of link relatives					Quarterly mean	Chain relatives	Adjusted chain relatives	Seasonal index				
	Year												
	1996	1997	1998	1999	2000								
I		112.90	86.11	96.88	97.14	98.26	100.00	100.00	100.00				
II	86.67	80.00	93.55	100.00	105.88	93.22	93.22	92.30	99.75				
III	84.62	78.57	96.55	80.65	72.22	82.52	76.93	75.09	81.15				
IV	140.91	163.64	114.29	140.00	126.92	137.15	105.50	102.75	111.04				

Here, Computed chain relative for 1st quarter is given by $(105.50/98.26) = 103.67$

And the correction for each quarter is $(103.67 - 100)/4 = 0.92$

Hence, adjusted chain relative for 2nd quarter is $(93.22 - 0.92) = 92.30$, for 3rd quarter $= (76.93 - 0.92 \times 2) = 75.09$ and so on.

Average of adjusted chain relative is 92.54

Finally, seasonal index for 2nd quarter is given by $(92.30/92.54) = 99.75$ and so on.

14.9.4. Deseasonalized values. The time series data, which are free from seasonal variations, are termed as desasonalized data. Since the seasonal variation occurs at particular period over a year, this can affect proper modeling of data. In that case deseasonalized series helps in forecasting business trend more accurately. Deseasonalized values are obtained by dividing the original quarterly prices by the corresponding seasonal price index, that means,

Deseasonalized values for Quarter I = (Original values of quarter I for different years) / (Seasonal index of quarter I) $\times 100$.

For example, consider the seasonal indices obtained in example 14.15, deseasonalized value of quarter I corresponding to 1995 is $(45/81.61) \times 100 = 55.14$, of quarter II of 1995 is $(54/98.96) \times 100 = 54.57$ and so on. The computed deseasonalized values are shown in the following table 14.33a.

Table 14.33a. Quarterly Deseasonalized values of prices.

Quarter	Years				
	1995	1996	1997	1998	1999
I	55.14	58.82	60.04	63.72	73.52
II	54.57	56.59	63.66	65.68	70.74
III	61.56	53.87	59.85	64.13	71.83
IV	58.54	54.64	63.42	70.25	64.40

14.10. Measurement of Cyclical Component

Business cycles play a vital role in economic behavior because these are the most important type of fluctuations in economic data. Hence, business cycles have received much attention in economic literature. However, despite of their importance in economic fluctuations, these are very difficult type of fluctuations to measure. This is because a) successive cycles vary widely in timing, amplitude and pattern, and b) cyclical factors are usually mixed with irregular factors. Hence, it is impossible to construct meaningful cyclical index of curves like trends and seasonal variations. However, the methods, which are used, for measuring cyclical components are as follows:

- i) Residual method
- ii) Reference cycle analysis
- iii) Direct method
- iv) Periodogram analysis
- v) Harmonic analysis

A brief discussion about the simplest method of determining irregular component, the residual method, is provided below.

14.10.1. Residual Method. The simplest and the commonly used method of determining the cyclical components of a time series is the residual method. In this method, the cyclical components are measured as the residue of other components. Sometimes cyclical, irregular or the cyclical movements remain as residuals, hence the name residual method. Thus, this method consists of eliminating trend and seasonal components of a given time series which provides cyclical-irregular movement, symbolically,

$$\frac{T \times S \times C \times I}{T} = S \times C \times I \text{ and } \frac{S \times C \times I}{S} = C \times I$$

These residuals are usually smoothed in order to obtain the cyclical components, which are usually expressed as percentage, and sometimes termed as cyclical relatives.

14.11. Forecasting

Due to increasing competition and complexity in business activities in almost all sectors of business worldwide together with rapid change in demands, expectations and trend towards automation, there is a necessity for every organization to know the key decision variable to be occurred in advance. Forecasting is the term, which refers to projecting the occurrence of uncertain events for a future period analyzing the past and present history of data. This may help the organizations to assess the future plans of actions or strategies. For example, sometimes inventory is ordered without being certain about the future demand, new equipment is purchased

despite this uncertainty, investments are made without knowing profits to be incurred in future (which usually happens in share market), alternative stuff may be employed without knowing the increase in the level of services that can be provided, etc. Forecasting is essential to make reliable and accurate estimates of what will happen in the future in the face of uncertainty. Usually, the decisions are influenced by the chosen strategy with regard to an organization's future priorities and activities. Once decision is made, the consequences are measured in terms of expectation to achieve the desired products or service levels. Although decisions are also influenced by the additional information obtained from the forecasting method used. Hence, the success of any business depends on its future estimates. On the basis of these estimates, a businessman plans his production, stocks, selling markets, expansion of plants, arrangements of additional funds, curtailment of loans, etc.

Forecasting is concerned with two main tasks, first the determination of the best basis available for the formulation of efficient managerial expectations, and second, the handling of uncertainty about the future, so that the implication of decisions become very comprehensible. Forecasting activities can be viewed as a part of the management information system. It also imposes on the control system.

14.11.1. Objectives of Forecasting. Although forecasts are commonly applied to capital investment decisions, strategic planning, product and market planning, etc with an expectation of certain levels of returns and help in decision making, the main objectives or functions of forecasting can be summarized as below. Since, it is impossible to evolve a worthwhile system of business control without acceptable system of forecasting, so no plan of action in business can be created without making forecasts of desired factor.

* Monitoring the continuing progress of action of plans depends on forecasting. The forecast provides a warning system of the critical factors to be monitored regularly because they might drastically affect the future course of action and performance in the plan as well.

14.11.2. Steps in Forecasting. A systematic way of initiating, designing and implementing a forecasting needs the following steps to be followed.

Purpose and policy implication of forecasting: Define objectives and the policies to be achieved that are trying to obtain by the use of forecast. The purpose of forecasting is to make use of the best available present information to guide future activities towards organization's objectives.

Selection of variables: Select the variables of interest such as capital investment, employment level, inventory level, purchasing of new equipment, which are to be forecasted.

Determination of time horizon: At this step determine whether the period of forecast to be made is short term, medium term or long term.

Collection of data: Collect the relevant time series data for the variable to be projected. A single time series data is enough for univariate study, while sometimes, for more efficient forecasting it might be required to collect data on more variable related to the target variable.

Selection of appropriate model for forecasting: This step can be divided into following sub-steps:

At first it is required to study the past behaviour of collected data: In case of time series data it is said that 'Data speaks about itself'. Hence one of the basic principles of all forecasting when historical (time series) data are available is to examine its past behaviour to get a 'Speedometer reading' about the trend of change in data in the past. This "Speedometer reading" constitutes the basis of forecasting. This task can be done by plotting the data in graph.

Once the behaviour is studied, the second step is to postulate various plausible mathematical models in order to explain or describe the past behaviour.

Fit the models using appropriate statistical techniques.

Select the model that best explains the given time series data: If the above-mentioned steps are not followed properly, the forecast will be subject to error. For example, if an attempt is made to forecast business fluctuations without understanding the behaviour of past changes, the forecast will be purely mechanical and subject to error.

Make forecast: Finally, make the forecast using the selected appropriate model and implement the results in business.

If a particular system is used regularly to generate forecasts, then data should be collected in a routine manner so that computations used to make the forecast can be done automatically using a computer.

14.11.3. Requirements of a good forecasting system. A forecasting system to be instrumental in contributing to better management or business decision making, needs to fulfill certain criteria, these are:

- i. It must involve managers or businessmen whose decisions are affected.
- ii. Individual forecasts or group of forecasts have to be specifically relevant to the decisions being taken.
- iii. The forecasts must not claim too much validity or authority

- iv. Implications of various probable errors in the prediction for the organizations need to be thoroughly worked so that management can evaluate the consequences of the probable outcomes.
- v. Management must at least know how badly things could go wrong if all the guesses turned out wrong.

14.11.4. Methods of forecasting. The efficient method of forecasting consists of deciding about the future in terms of past experience and familiarity with problems at hand. The businessmen have been trying to adjust themselves in such a manner as to make the best out of future conditions. For this purpose, the rule of thumb has been widely practiced in business. It should be noted that it is not possible to make 100% accurate future courses. The efficient forecast could be possible only if the influence of the various forces which affect these series such as climate, customs and traditions, growth and decline factors can be identified and eliminated using appropriate tools. However, some of the rational methods used for business forecasting are mentioned below.

- i. Historical analogy method
- ii. Field surveys and opinion poll
- iii. Business barometers
- iv. Extrapolation
- v. Regression analysis
- vi. Time series analysis
- vii. Exponential smoothing
- viii. Econometric models
- ix. Lead-lag analysis
- x. Input-output or end-use analysis.

The mathematical/statistical and popular methods of forecasting are extrapolation, regression analysis, econometric models and time series analysis. A short description of these methods is given below. A brief idea about the exponential smoothing is given below in smoothing section.

Extrapolation: Extrapolation is the simplest but useful method of forecasting. Extrapolation relies on the relative consistency in the pattern of past movements in some time series. Extrapolation is used frequently for sales forecasts and for other estimates when better forecasting methods may not be justified. Selection of appropriate growth curve such as arithmetic trend, semi-log trend, modified exponential trend, logistic or Gompertz curve, that best fits the past movement of the data can be guided by empirical and theoretical considerations.

Regression analysis: The regression approach offers many valuable contributions to the solution of forecasting problems. It is the means by which we select from the many possible or theoretically suggested relationships

between variables in a complex economy. If two variables are functionally related, then knowledge of one will make possible an estimate of the other through regression analysis. For example, if we know that advertising expenditure and sales are correlated, then for a given advertising expenditure we can find the probable increase in sales or vice versa.

Econometric models: The term econometrics refers to the application of mathematical economic theory and statistical procedures to economic data in order to verify 'the economic axioms or theorems, and to establish quantitative results in economics. At present most short-term forecasting uses only statistical methods with little quantitative information. The econometric model is the most formal method of forecasting, since the forecast is based on explicit mathematical models. Theoretically, the model makes possible a wholly mechanical forecast because once values have been estimated for the exogenous variables; the solution of the model gives specific values for the predicted variable. The models like VAR, ARCH, GARCH, ARIMA, MARIMA etc. are popularly used for forecasting economic or business behavior.

Time series analysis: The most popular method of business forecasting is time series analysis. The first step in making estimates for the future consists of gathering information from the past data. The statistical data, which are observed, collected and recorded at successive intervals of time, are usually used for this purpose. For example, if we observe yearly sales of a firm at different points of time, say for 10 years, it will constitute time series data. The data can display ups and downs. There may be several causes for increase or decrease from one year to another such as changes in the tastes and habits of people, growth of population, availability of alternative products, etc. It may be difficult to study all the factors of time series that have led to changes in sales. However, the time series analysis plays vital role in business decision making for the following reasons:

- i) It helps in understanding of past behavior
- ii) It helps in planning future operations
- iii) It helps in evaluating current accomplishments
- iv) It facilitates comparison of different time series

14.12. Smoothing Techniques

Before discussing the techniques for dealing with time series data exhibiting typical patterns of non randomness, it is required to test the randomness of data, then if randomness is present in the data, it is required to smooth out the data to remove randomness. Among several available tests of randomness, run test is particularly easy to carry out. It is a non-parametric test in the sense that no assumption is required to be made about the distribution from which the observations were drawn.

Although this task is beyond the scope of this book however, two other methods of smoothing time series data are discussed below.

One can develop a better forecast using a given time series data, if it is possible to determine which components actually exist in the series. Unfortunately, the existence of the random variation component often hides the other components. One of the simplest ways of removing the random or irregular fluctuation is to smooth the time series. Here two methods of smoothing are discussed, these are (i) moving average method and (ii) simple exponential smoothing method.

14.12.1. Moving average method. The irregular component in some time series may be so large that it sometimes hides the regular variations, thus create difficulty in any visual interpretation of the time series plot. In this case, the actual plot will appear jagged or rough and it is very required to smooth it to achieve a clearer picture of data.

This smoothing can be achieved through the method of moving average which is based on the idea that any large irregular component at any point in time will make a smaller effect if the observation at the point is averaged with its immediate upper and lower neighbours. In this method, the MA computed for a period (say k period) is considered as the forecast value of immediately following period ($k+1$ th period). This concept of smoothing data by this method is illustrated in example 14.12.1.

14.12.2. Simple Exponential Smoothing. We can easily observe two drawbacks of moving average method of smoothing a time series data. First, we do not have moving averages for the first two periods and last two periods (we will not get averages for one observations in either side if three yearly moving averages are considered, for three periods in either side if seven yearly average is considered, for two periods in either side if four yearly average is taken). If the time series has a very few observations, the missing values may constitute an important loss of information. Second, the moving average 'forgets' most of the previous time series, because average is calculated using a section of number of observations which does not contain a set of data of previous period, implying that this set of data has no effect on the average, which is not realistic. For example, in the 5-quarterly moving average the average for the period 4 reflects 2nd, 3rd, 4th, 5th and 6th periods but not affected by period 1. Similarly, the moving average for period 5 is not affected by period 1 and 2. Both of these problems can be addressed by what is called smoothing.

With simple exponential smoothing, the forecast is made up of the last period forecast plus a portion of the difference between the last period's

actual or observe value and the last period's forecast value, algebraically given by,

$$F_t = F_{t-1} + \alpha (y_{t-1} - F_{t-1}) = (1-\alpha) F_{t-1} + \alpha y_{t-1}$$

Where, F_t = current period forecast

F_{t-1} = last or previous period forecast

α = a weight called smoothing constant, ($0 \leq \alpha \leq 1$)

y_{t-1} = last period's observed value of the variable

That means, each forecast is simply the previous forecast plus some correction for actual value of the variable in the last period. If observed value is above the last period forecast, the correction will be positive, if below it will be negative. The correct α -value facilitates scheduling by providing a reasonable reaction to the variable without incorporating too much random variation. An approximate value of α which is equivalent to an arithmetic moving average, in terms of degree of smoothing, can be estimated as $\alpha = 2/(n+1)$. If $\alpha = 1$, there is no smoothing at all so that the forecast for the next time period is exactly the same as the actual value of time series in the current period, that means, $F_t = y_{t-1}$.

The exponential smoothing approach is simple to use and once the value of α is selected, it requires only two pieces of information namely Y_{t-1} and F_{t-1} to calculate F_t . To begin with the exponential smoothing process, let us suppose F_1 equal to the actual value of the time series in period t , which is Y_1 . Hence the forecast for period 2 is written as:

$$F_2 = \alpha Y_1 + (1 - \alpha) F_1, \text{ but since } F_1 = Y_1, \text{ hence, } F_2 = \alpha Y_1 + (1 - \alpha) Y_1 \text{ or } F_2 = Y_1$$

The accuracy of the forecasts can be improved by carefully selecting the value of α . If the time series contains substantial random variability then a small value of α gives better forecast. Otherwise, higher value of α is desirable. However, the accuracy of forecasts can be determined by comparing the forecasted value with the actual or observed value. The forecast error is defined as

$$\text{Forecast Error} = \text{Actual value} - \text{Forecast value}.$$

One popularly used measure of overall forecast error for a model is the mean absolute deviation (MAD), defined as $MAD = \frac{\sum |\text{Forecast Errors}|}{n}$, where standard deviation = 1.25 MAD .

Note that the exponential smoothing method also facilitates continuous updating of the estimate of MAD. The current MAD, is given by

$$MAD_t = \alpha | \text{Actual values} - \text{Forecast values} | + (1-\alpha) MAD_{t-1}$$

Thus, higher values of smoothing constant α make the current MAD more responsive to the current forecast errors.

Example 14.12.1. Consider the following yearly data for steel production from 1979 to 1996 for a country and plot the data and smooth using:

- i) Five-yearly moving average method
- ii) Three-yearly moving average method
- iii) Comment which one is better period of moving average.

Year	Production of steel for 1979-1996				
	Production (in tons)	Year	Production (in tons)	Year	Production (in tons)
1979	299	1985	418	1991	450
1980	366	1986	420	1992	333
1981	380	1987	450	1993	650
1982	390	1988	500	1994	571
1983	400	1989	622	1995	522
1984	419	1990	420	1996	366

Solution. (i) The computed values of five- yearly moving average for this data are tabulated in table 14.34. Here, the MA computed for first 5-period 1979-1983 is considered as the forecasted value for 1984, and so on.

(ii) The computed values of three - yearly moving average for this data are tabulated in table 14.35. Here, the MA computed for first 3-period 1979-1981 is considered as the forecasted value for 1982, and so on.

Table 14.34. Five-yearly Moving average of Production data.

Year	5-yearly Moving Total	5-yearly Moving average	Forecasted or smoothed values	Year	5-yearly Moving Total	5-yearly Moving average	Forecasted or smoothed values
1979				1988	2412	482.4	421.4
1980				1989	2442	488.4	441.4
1981	1835	367.0		1990	2325	465.0	482.0
1982	1955	391.0		1991	2475	495.0	482.4
1983	2007	401.4		1992	2424	484.8	488.4
1984	2047	409.4	367.0	1993	2526	505.2	465.0
1985	2107	421.4	391.0	1994	2442	488.4	495.0
1986	2207	441.4	401.4	1995			484.8
1987	2410	482.0	409.4	1996			505.2

Table 14.35. Three-yearly moving averages.

Year	3-yearly Moving Total	3-yearly Moving average	Forecasted or smoothed values	Year	3-yearly Moving Total	3-yearly Moving average	Forecasted or smoothed values
1979				1988	1572	524.0	429.3
1980	1045	348.3		1989	1542	514.0	456.7
1981	1136	378.7		1990	1492	497.3	524.0
1982	1170	390.0	348.3	1991	1203	401.0	514.0
1983	1209	403.0	378.7	1992	1433	477.7	497.3
1984	1237	412.3	390.0	1993	1554	518.0	401.0
1985	1257	419.0	403.0	1994	1743	581.0	477.7
1986	1288	429.3	412.3	1995	1459	486.3	518.0
1987	1370	456.7	419.0	1996			581.0

From the plot of original data, we have found that 5-yearly moving would smooth the data better. However, for the sake of algebraic comparison of smoothing values obtained by this two method, let us find the SSR and MAD for two smoothing series of values. The computed values of SSR and MAD for two series are shown in following table.

Table 14.36. Computation of SSR and MAD for smoothed data.

Year	Observed values	5-yearly MA	3-yearly MA	Squared residuals for 5-yearly MA	Squared residuals for 3-yearly MA	MAD for 3-yearly MA	MAD for 5-yearly MA
1979	299						
1980	366						
1981	380						
1982	390		348.3		1736.1	41.7	
1983	400		378.7		455.1	21.3	
1984	419	367.0	390.0	2704	841.0	29	52
1985	418	391.0	403.0	729	225.0	15	27
1986	420	401.4	412.3	346	58.8	7.7	18.6
1987	450	409.4	419.0	1648	961.0	31	40.6
1988	500	421.4	429.3	6178	4993.8	70.7	78.6
1989	622	441.4	456.7	32616	27335.1	165.3	180.6
1990	420	482.0	524.0	3844	10816.0	104	62
1991	450	482.4	514.0	1050	4096.0	64	32.4
1992	333	488.4	497.3	24149	27005.4	164.3	155.4
1993	650	465.0	401.0	34225	62001.0	249	185
1994	571	495.0	477.7	5776	8711.1	93.3	76
1995	522	484.8	518.0	1384	16.0	4	37.2
1996	366	505.2	581.0	19377	46225.0	215	139.2
Total				134026	195476.4	1275.3	1084.6

From table 14.36, it is clear that SSR for 5-yearly MA is 134026 and that for 3-yearly MA is 195476 that means SSR for 5-yearly MA is smaller than

3-yearly MA. Again we have the same result using MAD for this two-period MAs. Hence, given data can be smoothed by 5-yearly MA method better.

Example 14.12.2. A firm uses simple exponential smoothing with $\alpha = 0.1$ to forecast demand of their products. The forecast for the week of February 1 was 500 units whereas, actual demand turned out to be 450 units.

- i) Forecast the demand for the week of February 8.
- ii) Assuming the actual demand during the week of February 8 turned out to be 505 units, forecast the demand for the week of February 15. In this way continue forecasting through March 15, assuming that the subsequent demands were actually 516, 488, 467, 554 and 510 units respectively.

Solution. Given, for the week of February 1, $F_{t-1} = 500$, $y_{t-1} = 450$ and $\alpha = 0.1$

- i) so, for week of February 8, $F_t = F_{t-1} + \alpha (y_{t-1} - F_{t-1}) = 500 + 0.1 (450 - 500) = 495$ units.
- ii) for week of February 15, $F_{t+1} = F_t + \alpha (y_t - F_t) = 495 + 0.1 (505 - 495) = 496$ units.

Forecast of the demand up to the week of March 15 are shown in following table.

Table 14.37. Forecast of demand by exponential smoothing method.

Week	Demand y_{t-1}	Old Forecast F_{t-1}	Forecast Error $(y_{t-1} - F_{t-1})$	Correction $\alpha (y_{t-1} - F_{t-1})$	New Forecast (F_t) $F_{t-1} + \alpha (y_{t-1} - F_{t-1})$
Feb	450	500	-50	-5	495
	505	495	10	1	(495+1) = 496
	516	496	20	2	(496+2) = 498
	488	498	-10	-1	497
Mar	467	497	-30	-3	494
	554	494	60	6	500
	510	500	10	1	501

Example 14.12.3. A clinic has used a 9-month moving average forecasting method to predict drug and surgical instrument requirements. The actual demand for one item for the months 24-32 are shown in table below. Using the previous moving average data, convert the figures to an exponential smoothing forecast for month (i) 33 and (ii) 34

Month	24	25	26	27	28	29	30	31	32
Demand	78	65	90	71	80	101	84	60	73

Solution. (i) The moving average of a 9-month period is given by

$$MA = \frac{\sum x}{n} = \frac{78+65+\dots+73}{9} = 78$$

The estimated $\alpha = 2/(n+1) = 2/10 = 0.2$, given the actual demand for month 32 is $y_{32} = 73$, and assuming $F_{32} = 78$.

We have, $F_t = F_{t-1} + \alpha (y_{t-1} - F_{t-1})$

Or, $F_{33} = F_{32} + \alpha (y_{32} - F_{32}) = 78 + (0.2)(73 - 78) = 77$ units.

(ii) Considering the actual demand for the month 33 is 77, that means $y_{33} = 77$, the MA for the 10 months will be considered as the estimated value of F_{33} .

Here, 10-period MA of demand is given by

$$MA = \frac{\sum x}{n} = \frac{78 + 65 + \dots + 73 + 77}{10} = 77.9$$

The estimated value of α for $n = 10$, is given by $\alpha = 2/(n+1) = 2/11 = 0.18$

Hence we have, $F_t = F_{t-1} + \alpha (y_{t-1} - F_{t-1})$

Or, $F_{34} = F_{33} + \alpha (y_{33} - F_{33}) = 77.9 + (0.18)(77 - 77.9) = 78$ units.(approx.)

Example 14.12.4. The following table represents the number of cars sold in the first 6 weeks of the first two months of the year by a particular dealer:

Week	1	2	3	4	5	6
Sales	20	24	22	26	21	22

Smooth or forecast the data using (i) MA method (consider 3-week MA), (ii) Simple exponential method assuming smoothing factor $\alpha = 0.4$ and (iii) Compare the performance of the methods.

Solution. (i) The smoothed data obtained by 3-week MA method are shown in following table:

Table 14.38. Forecasted series of car sold by MA method.

Week	1	2	3	4	5	6
Sales	20	24	22	26	21	22
MA		22	24	23	23	
Smoothed or forecasted data				22	24	23

(ii) We know $F_t = F_{t-1} + \alpha (y_{t-1} - F_{t-1})$ and $\alpha = 0.4$,

F_2 is calculated above as equal to 20, and value of F_3 is calculated as follows:

$$F_3 = (1-\alpha) F_2 + \alpha y_2 = 0.6 F_2 + 0.4 y_2 \text{ since } F_2 = y_1, \text{ we have,}$$

$$F_3 = 0.6 \times 20 + 0.4 \times 24 = 21.6$$

Forecasted values of other weeks are shown in following table:

Table 14.39. Forecasted series of car sold by Exponential smoothing method.

Week	Sales (y_t)	y_{t-1}	$(1-\alpha)F_{t-1} + \alpha y_{t-1}$
1	20		-
2	24	20	20
3	22	24	21.6
4	26	22	21.76
5	21	26	23.456
6	22	21	22.47
7		22	22.28

(iii) Since the number of forecasts by two methods are different, for comparison of the performance of two methods, we have to compute mean squared forecast error MSE or mean squared residual MSR defined as $\sum(y_t - F_t)^2/n$, thus the performance of method which provides with the minimum value of squared forecast error will be considered as better.

Table 14.40. Comparison of methods used for forecasts.

Week	Sales (y_t)	Forecast by MA	Forecast by Exponential smoothing	Forecast Error by MA	Forecast Error by Exponential smoothing	Squared Forecast Error by MA	Squared Forecast Error by Exponential smoothing
1	20				-		
2	24		20		4		16
3	22		21.6		0.4		0.16
4	26	22	21.76	4	4.24	16	17.98
5	21	24	23.456	-3	-2.456	9	6.03
6	22	23	22.47	-1	-0.47	1	0.22
Total						26	40.39
MSE						$26/3 = 8.67$	$40.39/5 = 8.08$

The computed values of MSEs are 8.67 and 8.08 for MA and Exponential smoothing method respectively. Since MSE for exponential smoothing technique is smaller, this approach is considered as better.

14.13. Autocorrelation Co-efficient

The autocorrelation co-efficient describes the association or mutual dependence between values of the same variable but at different time periods. The autocorrelation coefficient provides important information about how a variable relates to itself for a specific time lag. The difference in the period for which a cause-and-effect relationship is established, is termed as 'lag'. While calculating the correlation, the time gap must be considered, otherwise, misleading conclusions may be arrived at. For example, a

decrease or increase in supply of a commodity may not immediately reflect on its price, it may take some lead-time or time lag. This lag of autocorrelation co-efficient helps determine duration of seasonal pattern in a time series data.

The formula for auto-correlation co-efficient at time lag k is given by

$$r_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where, k = length of time lag, n = number of observations.

Example 14.13.1. The monthly sales of a product, in thousands of units, in the last 6 months are given below:

Month	1	2	3	4	5	6
Sales	1.8	2.5	3.1	3.0	4.2	3.4

Compute the autocorrelation co-efficient upto lag 2. What conclusion can be derived from these values regarding the presence of trend and seasonal pattern in the data?

Solution. The calculation of autocorrelation of required lags are shown below:

Time	Sales x_i	x_{i+1}	x_{i+2}	$(x - \bar{x})(x_1 - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})(x_2 - \bar{x})$
1	1.8	2.5	3.1	0.6	1.44	
2	2.5	3.1	3.0	-0.05	0.25	-0.12
3	3.1	3.0	4.2	0	0.01	0
4	3.0	4.2	3.4	0	0	0.12
5	4.2	3.4	-	0.48	1.44	0
6	3.4	-	-	1.03	0.16	0
					3.30	

Here, x_{i+1} = One time lag variable derived from x , and x_{i+2} = Two time lag variable derived from x , $\bar{x} = 3$

$$\text{So, for } k = 1, r_1 = \frac{\sum_{i=1}^5 (x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^6 (x_i - \bar{x})^2} = \frac{1.03}{3.30} = 0.312$$

$$\text{and for } k = 2, r_2 = \frac{\sum_{i=1}^4 (x_i - \bar{x})(x_{i+2} - \bar{x})}{\sum_{i=1}^6 (x_i - \bar{x})^2} = \frac{0}{3.30} = 0.$$

Since the value of r_1 is positive, it implies that there is a seasonal pattern of 6 months duration and $r_2 = 0$ indicates there is no significant change in sales over 6 months.

Question

1. What do you mean by time series? What are its components? Explain them. Also explain the additive and multiplicative models of time series.
2. Define secular trend. Mention some causes of trend in time series data.
3. Mention the methods of measuring trend component of a time series data. Discuss one of them with its merits and demerits.
4. Explain the method of moving averages to determine the trend of a time series. What are the disadvantages of using this method?
5. Define of autocorrelation. Give an example of a single variable with two different time lags.
6. What do you mean by time series data? Describe different components of a time series. What purpose is served by analyzing time series data?
7. Discuss the different methods of determining trend in a time series. What are the relative merits and demerits?
8. Discuss different methods of obtaining measures of seasonal variations. What are the relative merits and demerits?
9. Criticize the use of moving average for determining trend. Establish the effect of eliminating trend by the method of moving averages on the other components of a time series.
10. What do you mean by the seasonal component of a time series data? What are the methods of determining seasonal component of a time series data? Describe the method of determining seasonal component by link relative method.
11. What is the difference between seasonal variation and irregular variation? Describe a suitable method of determining the seasonal variation along with its merits and demerits.
12. Explain the irregular component of a time series data. State the methods of measuring it. Discuss residual method of determining irregular component.
13. State the important factors causing the seasonal variations. Hence, discuss the methods of determining the seasonal variations of a time series.
14. Explain the components of a time series data and mention the methods of determining different components.
15. How does analysis of time series help in making business forecasts?
16. What do you mean by business forecasting? Discuss the steps involved in a forecasting process. Also discuss the commonly used methods of forecasting.

17. What do you mean by smoothing of time series data? Write a note on exponential smoothing. Also state the advantages of this smoothing method.
18. Explain clearly the different components of a time series data and explain any method of isolating trend component in a time series.
19. Describe the seasonal variations and cyclical fluctuations in a time series. What are the common methods used for eliminating seasonal component from a time series data? Explain any one of them along with advantage and disadvantage, if any.
20. What is the difference between seasonal and cyclical fluctuations in a time series data? Explain with an example how seasonal index is useful in planning sales or production for specific period.
21. What purpose is served by analyzing time series data? Explain how the quadratic trend line is estimated? Also give some examples of non-linear trend model.

Applications

22. The sales of a commodity (in tonnes) recorded for every month of a particular year is given below:

280, 300, 290, 280, 270, 240, 230, 220, 210, 200, 195, 190

Fit a trend line by the method of semi-average. Also determine the trend values using free-hand curve method.

23. The sales of sugar in thousand kg for a period of six years are given in following table. Determine the trend component by using (i) semi average method, (ii) least squares method

Year	2002	2003	2004	2005	2006	2007	2008
Sales of sugar	60	72	75	65	80	85	95

Determine the expected sales for 2009. Also eliminate the trend from the data using multiplicative model. What components are left over?

[Ans. $y = 76 + 4.86x$, $y_{2009} = 100.3$]

24. Production of cement (in thousand taka) of a factory recorded for nine consecutive years are as follows:

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Sales of cement	102.3	101.9	105.8	112.0	114.8	118.7	124.5	129.9	134.8

(i) Compute the trend component by semi average method and eliminate the trend values from data using multiplicative model. What components of time series are left?

(ii) Fit a straight line trend by the method of least squares and tabulate the trend values. Also estimate the production for 2009 and 2010.

(iii) What is monthly increase in the production of cement? (Hints: The slope of straight line or regression co-efficient measures the yearly increase in the trend, monthly increase can be obtained by dividing slope value by 12)

$$[y = 116.1 + 4.3x, y_{2009} = \text{Tk } 137.6 \text{ thousand}, y_{2010} = \text{Tk } 141.9 \text{ thousand}, \\ \text{monthly increase} = \text{Tk. } 4.3 \text{ thousand}/12 = \text{Tk. } 0.36 \text{ thousand}]$$

25. The profit (in '00000 taka) of a company for the period 2004-2009 are given in following table. Fit a parabolic curve of the second degree $y = a + bx + cx^2$ to the data.

Year:	2004	2005	2006	2007	2008	2009
Sales:	100	107	128	140	181	192

Also determine the projected profit for 2010.

$$[y = 126.657 + 18.042x + 1.786x^2, y_{2010} = \text{Tk } 227.40 \text{ thousand}]$$

26. The following table shows the number of items (in thousand) sold by a company for the last ten years 2001-2010. Fit the second-degree parabola to the data and estimate the sales of the company in 2011.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Sales	17	20	19	26	24	40	35	55	51	74	79

$$[\text{Ans. } y = 33.97 + 6.282x + 0.603x^2, y_{2011} = \text{Tk } 93.358 \text{ thousand}]$$

27. Fit the second-degree parabola to the following data of profit (in laks) of a company taking 2000 as origin and estimate the profit for 2004.

Year:	1998	1999	2000	2001	2002	2003
Profit:	10	12	18	15	13	16

$$[\text{Ans. } y = 14.86 + 1.339x - 0.482x^2; y_{2004} = 12.50 \text{ laks}]$$

28. Using three-yearly moving average, determine the trend component and plot the original and trend on the same graph paper.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Production	21	22	23	25	24	22	25	26	27	26

$$[\text{Ans. Trend value for 2000} = 22, \dots \text{for 2007} = 26.33]$$

29. The following data show the annual profits in thousands of Taka in a certain business:

Year	2001	2002	2003	2004	2005	2006	2007
Profit	60	72	75	65	80	85	95

Using the method of least squares to fit a straight line to the above data and make an estimate of profits in 2009. [Ans. $y = 76 + 4.86x$, $y_{2009} = 100.3$]

30. Fit a second degree parabola to the following data.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Production	4	8	9	12	11	14	16	17	26

$$[\text{Ans. } y = 11.89 + 20183x + 0.166x^2]$$

31. Determine the trend values for following data using centered 4-yearly moving average.

Year	2001	2002	2003	2004	2005	2006	2007
Sales	30.1	45.4	39.3	42.2	46.4	46.6	49.2

[Ans. 40.6, 42.2, 43.2, 45.1]

32. Consumption of monthly electric power in millions of KW hours for street lighting a big city during 2005-2009 is given below:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2005	300	280	275	240	230	215	225	250	270	300	320	345
2006	350	310	295	270	250	235	240	258	289	320	345	370
2007	370	335	325	298	270	250	260	285	310	350	370	395
2008	400	360	345	310	280	275	285	308	325	365	390	415
2009	425	375	400	335	320	298	315	335	360	360	365	390

Find out the seasonal indices by the method of monthly averages.

33. Assuming the trend is absent; determine the seasonal indices by the method of averages.

Year	Q ₁	Q ₂	Q ₃	Q ₄
2005	3.7	4.1	3.3	3.5
2006	3.7	3.9	3.6	3.6
2007	4.0	4.1	3.3	3.1
2008	3.3	4.4	4.0	4.0

[Seasonal indices: 98.66, 110.74, 95.30, 95.30]

34. Calculate seasonal indices by the method of ratio to trend method from the following data.

Year	Q ₁	Q ₂	Q ₃	Q ₄
2000	36	34	38	32
2001	38	48	52	42
2002	42	56	50	52
2003	56	74	64	62
2004	82	90	88	80

[Seasonal indices: 100.03, 109.55, 103.14, 87.28]

35. Find the seasonal indices from the following table by ratio to moving average method.

Seasons	2003	2004	2005	2006	2007
1st quarter	40	42	41	45	44
2nd quarter	35	37	35	36	38
3rd quarter	38	39	38	36	38
4th quarter	40	38	42	41	42

36. Data for production of wheat in million tones of a country are given below.

Year	Q ₁	Q ₂	Q ₃	Q ₄
2000	60	65	62	69
2001	62	68	65	68
2002	65	70	64	62
2003	70	75	68	67
2004	72	80	70	78

Compute the seasonal variations by the link relative method and ratio to trend method.

[Ans. 100, 106.04, 95.88, 98.08]

37. A sales (in million taka) of a company for 12 months of the year 2009 are given below

Month	1	2	3	4	5	6	7	8	9	10	11	12
Sales	280	300	280	280	270	240	230	230	220	200	210	200

- i) Find the trend line that describes the trend of sales by using semi-average method.
 - ii) Fit a linear trend line by least squares method and estimate sales for January 2010.
 - iii) Forecast the sales for January 2010 using simple exponential smoothing method.
 - iv) Also compute the forecasts for February and March of 2010 stating necessary assumptions.
38. In January of 2009, a city hotel predicted a February demand for 142 rooms occupancy. Actual February demand was 153 rooms. Using a smoothing constant $\alpha = 0.20$, forecast the March demand using exponential smoothing method.

[Ans. 144 rooms]

39. A shoe manufacturer, using exponential smoothing with $\alpha = 0.10$, has developed a October trend forecast of 400 units for a ladies' shoe. This brand has seasonal indexes of 0.80, 0.90 and 1.20 respectively for the last three months October, November and December respectively. Assuming actual sales were 344 units in October and 414 units in November, what would be the seasonalized December forecast?

[Hints: Compute the deseasonalized actual October demand as $344/0.80 = 430$ units, use $F_{t-1} = 400$, then find $F_t = 403$]

40. The following table shows the number of public sector industries failures in a country during the period 1999 to 2005,

Year	1999	2000	2001	2002	2003	2004	2005
No of failures	32	26	30	28	24	22	26

- i) Forecast the number of industries failed in different years using the 4-yearly MA method and computes the mean squared error (MSE).
 - ii) Forecast the number of industries failed using exponential smoothing method and also computes MSE
 - iii) Comment on the performance of the methods.
41. Consider the following time series data:

Week	1	2	3	4	5	6
Value	8	13	15	17	16	9

- i) Develop a 3-week moving average for this time series. What is the forecast for week 7?

- ii) Use $\alpha = 0.2$ to compute the exponential smoothing values for the time series. What is the forecast for week 7?
- iii) Compare the performance of forecast by this two techniques.
42. The following data shows the number of computers sold in a shop in different months of a particular shop:

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Computer sold	52	48	57	60	55	62	54	65	70	80	90	75

- i) Determine trend using the three-month moving average.
 ii) Determine trend using the five-month moving average.
 iii) Which one of these two methods is a better technique?
43. The following data are related to the production of sugar in a factory,

Year	Production ('000 tons)	Year	Production ('000 tons)
1995	17	2001	35
1996	20	2002	35
1997	19	2003	51
1998	26	2004	74
1999	24	2005	79
2000	40		

Fit a trend line to the given data and determine the trend values of production data.

44. Determine seasonal index for the sales data given below using ratio to moving average and link relative methods:

Year	Summer	Monsoon	Autumn	Winter
2004	30	38	34	37
2005	37	41	33	35
2006	37	39	36	36
2007	40	41	33	31
2008	33	44	40	40

45. The number of units produced in factory during the period 1999-2006 are as follows:

Year	1999	2000	2001	2002	2003	2004	2005	2006
Units produced	56	55	51	47	42	38	35	32

- i) Fit a straight line trend and obtain the trend values
 ii) Eliminate the trend and comments on the components have been left in the data.
 iii) What is monthly increase in the number of units produced?
 iv) Estimate the expected units produced in 2008.
46. Use the least squares method to determine the sales for the year 2009

Year:	2003	2004	2005	2006	2007
Sales:	100	110	130	125	160

47. Calculate seasonal indices by using the ratio to trend and link relative methods from the following data:

Year	Summer	Monsoon	Autumn	Winter
2001	26	32	31	33
2002	29	36	33	34
2003	30	38	34	37
2004	37	41	33	35

48. The owner of a land developing company observes that their sales have been increasing over time and collected the following information for the first six months of 2011.

Month	Jan	Feb	March	April	May	June
Sales (in million Tk)	75	99	107	140	165	188

- a) Plot these data
- b) Develop a linear equation that best describes these data
- c) Estimate December sales from the equation.

49. A particular brand of computer has been developed in 1996, and sales have been increasing ever since. The following represent number of computers sold in different years in thousand per year.

Year	1996	1997	1998	1999	2000	2001	2002	2003
Number sold	28.9	36.7	45.2	66.7	80.4	85.7	100.2	125.6

- i. Develop a linear estimating equation for the give data
- ii. Also develop a second degree estimating equation for the given data
- iii. Estimate the number of computers sold in 2005 using both equations.
- iii. If the actual number of computers sold in 2005 is 150.8 thousand, which model would be considered to be the best and why?
- iv. On the other hand, if we assume that number of sales would be decreased in future years due to global financial crisis, which model would you accept as the best?

50. A famous company that specializes in the construction of antipollution filtration devices has recorded the following sales in the last 9 years.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Sales (in lac)	45.2	66.7	80.4	85.7	100.2	125.6	116.8	130.5	135.9

- i) Plot the data in graph
- ii) Would a linear or second-degree equation provide the better prediction of future sales?

51. The number of tables sold by a company since its establishment are given below:

Year	1996	1997	1998	1999	2000	2001	2002	2003
Sold	42	50	61	75	92	111	127	138

- i) Fit a linear trend line for number of tables sold
- ii) Fit a parabolic trend line

- iii) Estimate the number of tables to be sold in 2007.
 - iv) Comment on the performance of the model (use theoretical concept observing the increasing trend and sum of squares of errors as well).
52. The number of workers (in thousand) in a large industry at different times are given below:

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Sold	3.1	3.8	4.2	4.6	5.3	5.7	6.4	7.0	7.9	9.6

- i) Fit a linear trend line for number of workers.
- ii) Fit a parabolic trend line.
- iii) Estimate the number of workers would be employed by the industry in 2012.
- iv) Comment on the performance of the model.

CHAPTER - 15

SAMPLING, SAMPLING DISTRIBUTIONS AND ESTIMATION

15.1. Introduction

In our daily work we usually check a little of some thing to take decision about the whole. The simple example of this might be cooking rice or some other dishes. What do we do in practice? We usually check one or two rice to check the boiling status of rice of whole pan, taste a little to check the sufficiency of salt in some curries, or taste a little to check the sweetness of some sweet dish, etc. In all these cases, if we taste all, nothing will be left for others. Again, a good shopper usually purchases a little of a new commodity before purchasing a chunk to check the quality or demand of it, a seller also purchases a little of some new commodity to check its demand in market, a chemist does the same thing when (s) he tests the percentage and affect of alcohol in any sleeping or coughing medicine, etc. Testing all the product often destroys it and sometimes unnecessary. Hence, to determine the characteristics of the whole, to know the true state of the whole, we have to sample only a portion. Thus, one of the objectives of the fields of statistical analysis is to know the 'True' values of different parameters of the population. However, it is not sometimes possible due to time, cost and other constraints to study the whole population; instead, random samples from the respective population are studied. Hence, there are two ways in which information from a population can be gathered, these are (i) Complete enumeration or census method (ii) sample survey method (sampling is a part of sample survey).

We may be interested in knowing consumers' reaction to a particular product. In this case, we may contact each and every consumer of such product or we can just take a sample of consumers. The former case is known as census method while the later is known as sample survey method.

15.1.1. Types of population. Population is a group of items, units or objects which is under reference of study. The term population is also termed as universe by a number of statisticians and scientists. Workers in a factory, inhabitants of a region, students in a university, number of paddy fields in a district, unit products of an industry, fruits in a shop, rickshaw pullers in a city, employees in an organization, fishes in a river, etc are few examples of population.

Definition. Population. All possible units or items specified by certain characteristics under the targeted study area constitute a population.

Generally, the population consists of a large number of living or non-living units under study. The units or objects of the population vary from survey to survey in the same region or sphere of activity depending on the aims and objectives of the study.

One should keep in mind that statistical population is not only the human population which is usually conceived in literature. It is generally a group or collection of items specified by certain characteristics or defined under certain restrictions.

However, the statistical population can be classified into four major categories. These are (i) Finite population, (ii) Infinite population, (iii) Real population, (iv) Hypothetical population.

- (i) ***Finite population:*** If the number of items or units constituting the population is fixed and limited, it is known as finite population. This type of population usually consists of existing units. For example, the workers in a factory, students in a college, etc.
- (ii) ***Infinite population:*** If the number of items or units constituting the population is not fixed or infinite, it is called an infinite population. The populations of stars in the sky, the population of fishes in a sea, the life time of a bulb etc are some examples.
- (iii) ***Real population:*** A population constituting the items which are all present physically is termed as real population. The populations of products of a factory for a particular time, the population of employees of a garment, the population of inhabitants in a specific area, etc are some example of real population.
- (iv) ***Hypothetical population:*** A population constituting of the units which have not yet happened, but likely to happen, is called a hypothetical population. These types of populations usually result from repeated trials. For example, the population of outcomes results from tossing a coin repeatedly, the population of outcomes results from rolling a die again and again, etc.

15.2. Complete Enumeration or Census Method

Under this method, the relevant data are collected for each and every unit (person, household, shop, factory, businessman, etc, as it is needed) belonging to the population or universe which are of interest in any particular study. That means, when every item of a population is observed or measured or counted, this is called a census or complete enumeration.

This is possible when population is finite and every individual of the population is available or reachable. For example, if you are asked to find the average weight of the students of your class, you can take weight of each student. Then it will be a census. So, census is preferred to sampling in cases where the design of research allows for the inclusion of the whole population if it is accessible and also within manageable limits.

The reasons why a census must be used are:

- i) When every item or units of the population is required to be considered in the study,
- ii) When extreme accuracy of the results of study is needed,
- iii) When crucial decision will have to be made on the basis of the results obtained from the study,
- iv) Moreover, if the population size is small and finite, it is easy to enumerate all units of population.

One of the situations when census must be used and sampling may not be effective is cited below.

Suppose the Government decides to make a particular city free from AIDS and plans to identify the HIV carriers of that city to take necessary steps to prevent the further spread of AIDS in the city, then it is required to test all the inhabitants of that city for presence of HIV i.e. it is required to conduct a census rather than a sample. If a sample is considered in this case, and if one of the HIV carriers is left in the city, there is every possibility of spreading AIDS again in the city. Note that in this case even a new born baby will not be left from testing.

Definition. Census. If the relevant data/information from each and every unit of the targeted population under enquiry is collected it is called census.

Population census, agriculture census, animal census are some examples where census or complete enumerations have been done. In Bangladesh Population Census is done in every ten year. First census in Bangladesh was done in 1974, although it was supposed to be done in 1971.

- **15.2.1. Advantages of census.** In spite of having a number of limitations of conducting a census, one may get some advantages if it is possible to conduct it. However, we have already mentioned the reasons or situations when one should perform census for any enquiry. From that point of view, the most important two advantages of census are (i) details information can be obtained for each and every unit of the population and (ii) greater accuracy can be obtained by using census than that of sample survey.

On the other hand, considering the extreme large effort, money, time and destructiveness (in case of checking lifetime of tube lights) involved in

carrying out a complete enumeration, the idea of collecting information by census method may have to be dropped. The choice left to the researcher is to check the life times through collecting a random sample from the lot, which is known as sampling method. Thus, unless the information is required for all units in the domain of study, one can resort to the method, known as, the sampling method to obtain information or to study the population.

15.3. Sampling

Sampling is the process of selection of individual units of population starting from the formulation of the objective of the study to the collection of individual units using appropriate technique.

The intermediate steps include the selecting target population and sampling units, designing a sampling frame and determining appropriate sample size. A sampling frame is the complete list of sampling units from which the sample is to be selected. Examples of sampling frames are the telephone directories, electoral rolls, list of books in a library, list of students enrolled in a university, list of schools and colleges in a country, list of the employees working in a firm, list of workers in a garments factory, etc. Sometimes these lists are in existence and can be readily obtained from the respective authority. Sometimes these have to be prepared at an extra cost before selection of units is done. Effectiveness of sampling mainly depends on the construction of an appropriate Sampling frame.

The term sampling refers to the process of collection of sample from a population. This term is sometimes used as a synonym of sample survey which means studying the characteristics of a population through a sample.

Definition. Sample. A sample is a representative part of a population.

Definition. Sampling frame. A sampling frame is the complete list of all sampling units of targeted population. It is necessary to prepare a sampling frame before sampling is made.

Definition. Sampling. Sampling is defined as the total process involving in collection of sample from a target population for a particular study.

15.3.1. Purpose of sampling. A sample is not studied for its own sake. The basic objective of its study is to draw inference about the population. In other words, sampling is only the tool which helps us to know the characteristics of the universe or population by examining only a small portion of it. Such values or characteristics obtained from the study of a sample are called statistics (statistic in singular), while their counterparts in respective population are called parameters.

15.3.2. Principles of Sampling. The following are two important principles which determine the possibility of arriving at a valid statistical inference about the features of population or process:

i) *Principle of statistical regularity* : The principle is based on the mathematical theory of probability. According to King 'The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to process the characteristics of large group'. This principle implies that if a sample is taken from a population of interest is likely to possess all the features of the parent population. Thus the random sample is the one in which items are chosen from a population in such a way that each item has an equal probability of being selected in the sample. When the term random sample is used without any specification, it usually refers to a simple random sample, such a sample would be representative of the population, and only this type of sample would provide fairly accurate characteristics of the population. For example, to understand the book buying habit of the students of university, instead of approaching all students, investigator can talk to a randomly selected group of students to draw the inference about buying behaviour of all students in the university.

ii) *Principle of 'Inertia of large number'* : This principle is a corollary of the principle of statistical regularity and plays a significant role in the sampling theory. This principle states that, under similar condition, as the sample size increases, the statistical results are likely to be more accurate and stable. For example, if a fair coin is tossed a small number of times, we may not get equal number of heads or tails, as we expect, but if it is tossed a large number of times, then the chance of getting relative frequency of heads and tails to be equal would be very high, that means the results would be very near to 50% heads and 50% tails. Similarly, if it is intended to study the variation in the hourly production of a machine, say readymade garments, over a number of years and data are randomly collected for ten or twelve hours only, the results would reflect large variation in production due to the deterministic or non-deterministic causes. However, if the data for production are collected for a large number of hours, say 500, it is quite likely that a little variation in the aggregate would be found. This does not mean that the production would remain constant for all hours; rather it implies that the changes in production of the individual hour will be counter balanced and reflect small variation in production for the factories as a whole.

15.3.3. Advantages of Sampling. The advantages of sampling are:

- i) Reduced cost
- ii) Greater speed

- iii) Greater scope
- iv) Greater accuracy

Reduced cost : When data are collected only from a small fraction of the entire population, expenditure is smaller than if the entire group is studied.

Greater speed : The volume of the data to be selected will be obviously smaller than the size of population. Hence it can be collected, tabulated and summarized more quickly with a sample in comparison with the total population. In applied research where urgent answers to the certain problems are needed, this aspect of sampling for its speed gets an additional importance.

Greater scope : In studies where a complete enumeration and census of all units of population are impracticable and the research requires the use of highly trained personnel or specialized equipment, the choice may be between collecting the information by sampling rather than abandoning the research itself. In this case, the surveys using sampling provide greater flexibility and scope. Sampling is also obligatory in the case where population is infinite or countable infinite. It is also highly recommended in case of destruction, for example testing the strength of glass, testing blood, etc.

Greater Accuracy : With the reduction in the volume of work, personnel of higher expertise and training can be employed and a more careful supervision of the field work and processing of the data are possible. Hence sampling may produce results which are more accurate than those which could have been obtained through a complete census. Moreover, sampling is particularly more important in obtaining accurate results about phenomena which are undergoing rapid changes such as opinions about political and social issues, affect of price hike in the market, etc.

15.3.4. Census vs. Sample survey. The difference between a census and a sample survey is tabulated below.

	Census	Sample Survey
1	It is a study which considers all units of the population.	It is a study which considers a part of units of the population.
2	It is useful when the population size is small and finite.	It is useful when the population size is large and/or infinite
3	It is more expensive and more time consuming.	It is less expensive and less time consuming.
4	If the study is performed with trained personnel, the results obtained from census may be more accurate and adequate.	Even if the study is performed with trained personnel, the results obtained from census may not be accurate and adequate.
5	There is possibility of occurrences of only non-sampling errors, if any.	There is possibility of occurrences of both sampling and non-sampling errors.

15.3.5. Requirements of a good sample. If information from a sample data is to be generalized to a population, it is essential that the sample should be representative of that population. A representative sample would be a miniature in all respect of the population from which it has been drawn. An adequate sample is one that contains enough cases to insure reliable results. The basic requirements of a good sample are (i) a good sample should be representative, (ii) it should be adequate or of sufficient size to allow confidence in the stability of its characteristic.

15.4. Methods of Sampling

When a sample is required to be reflected from a population, it is necessary to decide which method of sampling should be applied. The various methods of sampling or sampling designs can be grouped under the heads as random or probabilistic sampling, non-random or non-probabilistic sampling and mixed sampling. If the sampling process is random, the laws of probability can be applied, thus, the pattern of sampling distribution needs to interpret and evaluate the sample. A non-random sample is selected on the basis of other than probability considerations such as expert judgment, convenience or some other criteria. The common methods of sampling are as follows:

- (a) Random or probabilistic sampling methods
 - i. Simple random sampling
 - ii. Stratified random sampling
- (b) Mixed sampling method
 - iii. Systematic sampling
- (c) Non-random sampling or non-probabilistic sampling methods
 - iv. Quota sampling
 - v. Judgment sampling
 - vi. Convenience sampling
 - vii. Snowball sampling

Brief descriptions of some of the sampling methods are provided below:

15.4.1 Simple random Sampling. Simple random sampling refers to the sampling technique in which each and every item of the population has an equal chance of being included in the sample. Thus simple random sampling is a method of selecting n units out of a population of size N by assigning equal probability to all units, or a sampling procedure in which all possible combinations of n units that may be formed from the population of size N have the same chance of being a sample. That's why it is also sometimes referred to as unrestricted random sampling.

This method is appropriate when the population size is not too large and population units are homogeneous with respect to the characteristics of interest. This method is very easy to use.

If a unit is selected and noted and then returned to the population before the next drawing is made and this procedure is repeated n times, it gives rise to a simple random sample of n units and this procedure is called a simple random sampling with replacement. On the contrary, if this procedure is repeated until n distinct units are selected and all repetitions are ignored, then the procedure is called a simple random sampling without replacement.

For example, suppose there are 500 students in a class. We have to draw a sample of size 50. Let us think that the gender of the students will not hamper the objective of the study. The following steps are taken to draw a random sample of size 50:

- i) At first collect the list of the students from the academic office of the School
- ii) Assign a three-digit number starting from 001 against each successive student. Suppose the numbers are 001 to 500.
- iii) Then draw 50 numbers following any column or row of random number table available in any book on sampling. Usually the repetition of any random number will not be allowed, i.e., the sample is drawn without replacement or any student will not appear in the sample twice.
- iv) Select the students corresponding to the number obtained from the random number table. These 50 students will constitute the sample of size 50.

15.4.1.1. Methods of obtaining a simple random sample. To ensure the randomness of selection, the following methods are adopted for collection of the data by simple random sampling method.

Lottery Method. This is a very popular method of taking a random sample. Under this method, all items of the population are numbered or named on separate sheet of papers of identical size, colour and shape. The sheets are then folded and mixed up in an urn. A blindfold selection is then made of the number of sheets required to constitute the desired size of sample. The selection of items thus depends entirely on chance. For example, if we want to take a sample of twelve dealers from the population of 300 dealers of a company, it is required to write the names of dealers on 300 separate papers of same size, shape and colour, fold these papers, mix them thoroughly and then make a blindfold selection of 12 papers, which would provide the required sample of twelve dealers.

Random number table method. The lottery method is quite burdensome if the size of population is large. An alternative and the most efficient method of drawing a simple random sample is using the Table of Random Numbers. The table of random numbers have been prepared by Kendall and Smith, Fisher and Yates, and Tippett, and constitute a very convenient and most objective method of random selection. The table consists of 41600 digits taken from census reports and combined by fours to give 10400 four-figure numbers. From the members of population already numbered from 1 to N, the required number of units is selected from one of these tables in any convenient and systematic way. These tables are so prepared that all the ten numbers from 0-9 have an equal chance of being selected. If we examine these tables, we will see that whether we go down a column or across a row, there is no distinct pattern. These numbers are computer generated and are truly random..

In random number sampling each element of the population is assigned a number, for example, for a population of size 400, the numbers like 001, 002,, 400 are usually assigned. Once this has been done, one can use the tables for random sampling. Although the tables of random numbers are available in most of the books on statistics on sampling theory, for the sake of explanation, a sample of random numbers is provided below:

3905	9796
0946	9133
0106	6465
1840	9779
7056	3015
9736	5661
9915	5686
5614	7123
5477	6629
5701	8733

Let us illustrate the method with an example. In order to obtain a sample of 10 students from 400 students who have already been assigned numbers from 001 to 400. Using the first three columns (because our population size is three-digit) we get the following three-digit numbers 390, 094, 010, 184. However, in the first column we have only four numbers within the range 001-400, others are ignored as they do not lie in the range 001 to 400. When the end of the table reached one can start again at the top with the next three unused digits along the top row (597 in this case). Once we have selected our sample of 10 numbers, we have to go to the population and select the corresponding students. Although we started at the top of the table and read downwards, it is better to start at a randomly selected place in the table and any direction can be considered for selection of a random sample.

15.4.1.2. Sampling without replacement. When all the sampling units are considered as distinct from one another and the unit drawn is not put back or replaced in the population before another unit is drawn is known as the sampling without replacement. In this case, there is no possibility of appear the same unit more than once and the size of population does not remain the same with every unit drawn. If there are N different units, n units can be selected from these N units one by one without replacement in N_{P_n} ways if the order of the units are important. On the other hand, if there are N different units, n units can be selected from these N units one by one without replacement in N_{C_n} ways if the ordered are not important which is equivalent to the selection of n units at a time from N units. That is, if there are N distinct units in the population and a sample of size n is to be drawn, the number of distinct samples of size n that can be drawn from the N units is given by the combinational formula

$$N_{C_n} \text{ or } \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Suppose a population contains 4 units denoted by A, B, C and D, if a sample of size 3 is drawn from this population, there will be $\binom{N}{n} = \binom{4}{3} = 4$ distinct samples which are : ABC ABD ACD BCD

Here, the same letter is not repeated in any sample and the order of the occurrence of the letter has been ignored. For example, ABC and ACB are identical, and so one is considered in the sample.

The probability of each drawing in case of sampling without replacement is shown in following table.

Table. Probabilities of different element in sampling without replacement.

Position of elements in the sample	Number of possible choices	Probability
1st	N	$1/N$
2nd	$N-1$	$1/(N-1)$
3rd	$N-2$	$1/(N-2)$
.	.	.
.	.	.
($n-1$)th	$N-n+2$	$1/(N-n+2)$
n th	$N-n+1$	$1/(N-n+1)$

Two important things need to be noted in sampling without replacement, firstly, with the drawing of every unit, the possible number of choices goes on decreasing by one every time, thus implying that every time, we are

sampling from a population of different size. Secondly, the probability of selection of every unit is different, although known. In a population of 100, the first unit to be drawn has 100 choices with a probability of 1/100; the second unit to be drawn has 99 choices with a probability of 1/99, and so on.

15.4.1.3. Sampling with Replacement. The random sampling in which case a unit that has been drawn, is put back or replaced, and can be drawn again, is called a sampling with replacement or unrestricted sampling. In this case, the same unit of population has the possibility of reappearing in the sample more than once. If there are N distinct units in the population and a sample of size n is to be drawn, the total number of samples will be N^n , however, the number of distinct samples of size n that can be drawn from the N units is given by the combinational formula:

$$\frac{(N+n-1)!}{n!(N-1)!}$$

For example, suppose a sample of size 3 is to be drawn from the population of four units A, B, C and D using the method of sampling with replacement.

In this case, the number of possible samples will be $\frac{(4+3-1)!}{3!(4-1)!} = \frac{6!}{3! \times 3!} =$

$\frac{120}{6 \times 6} = 20$ with no two groups being identical, which are listed below:

AAA	BBB	CCC	DDD	AAB	AAC	AAD	ABB	BBC	BBD
ACC	BCC	CCD	ADD	BDD	CDD	ABC	ABD	ACD	BCD

In both cases, samples are selected using Random Number Table.

However, if we consider all possible occurrences of samples (identical groups with different positions of sampling units of draw), there will be $4^3 = 64$ total samples. For example, AAB can occur in three possible ways AAB, ABA, BAA, similarly ABC can occur in six possible ways such as ABC, ACB, BCA, BAC, CAB, CBA and all such arrangements will constitute the 64 (4 for the same sampling units, $12 \times 3 = 36$ for two same units and one different unit, $4 \times 6 = 24$ for all distinct sampling units) possible samples.

15.4.2 Stratified Random Sampling. This is a form of random sampling in which the population is divided into groups or categories that are mutually exclusive so that no individual or item can belong to two groups. The groups are formed with the principle that within the group the items are homogenous with respect to some criteria (for example, sex, age, religion, income level, etc) and between the groups they are heterogeneous. These groups are called strata (in singular- stratum). Within each of these strata a simple random sample is selected. If the same proportion of each stratum is

taken, then each stratum will be represented in the correct population in the overall result.

In stratified sampling, the population of N units is sub-divided into k sub-populations, called strata, the i th sub-population having N_i units ($i = 1, 2, \dots, k$). These sub-populations are non-overlapping so that they comprise the whole population such that

$$N_1 + N_2 + \dots + N_k = N$$

In order to draw a sample of size n from the whole population, a sample of size n_i is drawn from i th stratum such that $n_1 + n_2 + \dots + n_k = n$. Hence, the procedure of taking simple random sample from every stratum separately is known as the stratified random sampling.

In case of proportional allocation, a sample is taken from each stratum as

the proportional to the stratum size, i.e. $n_i \propto N_i$ or, $n_i = \left[\frac{n}{N} \times N_i \right]$

For example, suppose in a large class of 500 students, where there are 300 male and 200 female students. So, there are two strata, which are mutually exclusive. In order to take a sample of 50 students, we can allot the sample in each stratum as proportional to stratum size. Thus, $\left[\frac{50}{500} \times 300 \right] = 30$ male students out of 300 and 20 female students out of 200 will represent the students of all gender proportionally and we can be sure of getting a balanced sample of male and female.

Stratified sampling procedure is appropriate when

- i) The sample size is large and respect to some criteria and it is possible to divide the whole population into some mutually exclusive groups.
- ii) It is required to estimate the characteristics for different strata separately.

The advantages of a stratified random sampling are as follows:

- i) Stratification ensures adequate representation of various groups of population, which may be of some interest or importance,
- ii) Stratification also ensures selection of a better cross-section of population than that under un-stratified sampling,
- iii) stratification brings a gain in precision in estimation of a characteristic of a population when there are clear strata present, because as a result of stratification, the strata variances will be as small as possible,
- iv) Estimates of population characteristics for different strata can be obtained separately by this sampling procedure.

However, this method has some inherent disadvantages as well. For instance,

- i) If the strata are not clearly defined or if the stratification can not be done properly, this procedure may lead to inaccurate results, and
- ii) This method is not suitable if population size is small.

15.4.3. Systematic Sampling. Instead of going through the laborious process of choosing sample randomly from a list that is not necessarily assumed to be random, it would be much simpler if we could select only the first unit randomly with the help of random numbers, and the rest of the units are selected automatically or systematically according to some pre-designed pattern. Then this type of sampling is known as the systematic sampling. In this case, sample is selected at regular intervals from an ordered list of sampling units. In order to select a sample of size n from a population of size N , let $N = nk$, where n is the number groups and k is the number of items in each group, then in this method first unit is selected randomly from first k units randomly listed in the sampling frame. Let this is the r th unit of first group of k units, and then every r th unit is selected from each of the subsequent $n-1$ groups, i.e. $(k+r)$ th item will be second member of sample, $(2k + r)$ th member will be third member of sample, and so on. In this way sample of size n is selected.

- The procedure of systematic selection is easier and more convenient than simple random sampling. It provides more even spread of the sample over the population list and hence leads to a greater precision. The dependence or linkage of one member of the sample on the previous one makes the process different from simple random sampling method, in which selection of every member is independent of the other. That's why method is sometimes termed as a Quasi-random sampling or mixed sampling.

This method of sampling is appropriate when the population is too large for simple random sample, or if a quick sample is to be selected where chance of being a member of sample for all units is not a matter. It is especially useful for the population with more or less definite periodic trend. For example, weekly sales, 12-monthly rainfall, quarterly remittance, etc.

The main advantage of this sampling is its simplicity of selection, operational convenience and even spread of the sample over the population. The second advantage is that because of its simplicity of drawing sample, it is very useful for large samples.

The serious disadvantage of systematic sampling lies in its use with populations having unforeseen periodicity which may substantially contribute bias to the estimate of the parameter, or if the list itself is biased then serious error may arise in estimation. Again, it does not provide with a random sampling, it is only random if ordered list of population is truly random.

15.4.4. Quota Sampling. The chief characteristic of simple, stratified and systematic sampling is that known probability is associated with the selection of every individual of the sample that means, the sample is random or quasi-random. Sometimes non-random sampling methods are also used when it is not possible to use a random sampling, particularly, when the whole population is not known.

Quota sampling is an example of non-probability sampling. It involves the selection of sample units within each group or quota, on the basis of the judgment of interviewers rather than on calculable chance of being included in it. Interviewer is given considerable freedom in choosing the individual cases. Quota sampling is a method in which an interviewer is instructed to interview a certain number of respondents with specific characteristics. The quotas are selected before sampling takes place and they are chosen so that they reflect the known population characteristics. Age, sex and social class are the three universally used quota controls.

It is useful when the number of sampling units is pre-fixed for groups of population of the same characteristics.

This method is extensively used in opinion survey, for example product satisfaction opinion, polling opinion, etc. Suppose, a company wants to know the customers' opinion regarding the quality of their product, and decides to take opinion about quality from 100 female and male consumers of apparently young and old age as per the following table:

Age Group	Sex	Number
Young (Age group 20-40 years)	Male	25
	Female	10
Old (Age group above 40 years)	Male	40
	Female	25

The number corresponding to each age/sex group is the quota for respective group.

The investigator can collect the opinion just by asking an indefinite number of customers one by one standing in the exit of a market or by house-to-house survey. In this case, the investigator will start asking the customers coming out of the market after shopping regardless of the age and sex. Once the investigator finishes taking information from 25 young male customers, s/he will not try to know the opinion of any more young male customer. For this the investigator has to ask an indefinite number of male customers, because, some of the customers may not use the particular product or some may not respond properly. Similarly, the investigator needs to ask the opinion to an indefinite number of young female customers to fulfill the quota of 10 females. Same procedure is to be followed in case of other groups of customers.

Advantage:

- i) It enables the fieldwork to be done quickly because a representative sample can be achieved with a small sample size.
- ii) Costs are kept to a minimum level
- iii) Administration is relatively easy

Disadvantage:

- i) It is not possible to estimate the sampling error, because the process is not a random process.
- ii) The interviewer has to choose the respondents and may not be able to judge the characteristics easily.
- iii) Non-responses are not recorded in this method
- iv) The process does not allow for an easy supervision of the field worker, hence the correctness of the data collected remain doubtful.

15.4.5. Judgment Sampling. In this method of sampling, the choice of sample items depends exclusively on the judgment of investigator. The investigator exercises his judgment in the choice of sample items and includes those items in the sample which he thinks are most typical of the population with regard to the characteristics under investigation. For example, if a sample of twenty workers is to be selected from a factory having 100 workers for analyzing their spending habits, the investigator would select twenty workers, who in his/her opinion represent the factory.

Advantage:

The only advantage of this type of sampling is that it is very easy to select sample according to the judgment of investigator. There is no need of further query for this. One investigator can justify the choice of sampling units in his own way. However, the success of this method completely depends on the excellence in judgment.

Disadvantage:

This method possesses a number of disadvantages, some of important limitations are mentioned below:

- i) The method is not at all a scientific method; hence the results may be considerably affected by the personal prejudice or bias.
- ii) This method provides the quick estimation
- iii) This method involves the risk that the investigator may establish foregone conclusions by including those sampling units which conform his preconceived notions.
- iv) There is no objective method for determining the sample size or likelihood of sampling error, which is considered as a big defect of this method.

Applications:

Although the principles of sampling theory are not applicable to judgment sampling, this method is often used in solving many types of economic and business problems, such as :

- i) It is used when sample size is small; in such case simple random sample may miss the more important elements, whereas judgment selection would certainly include them in the sample.
- ii) In solving everyday business problems and making public policy decisions, executives and public officials are often in hurry and can not wait for probability sampling. In this situation this is the only practical method.
- iii) Judgment sampling may be used to conduct pilot survey. In any case, the reliability of sample results in judgment sampling depends on the quality of the sampler's expert knowledge or judgment.

14.4.6. Convenience Sampling. The method of convenience sampling is also known as chunk or portion. A chunk is a fraction of one population taken for investigation because of its convenient availability. Thus a chunk is selected neither by probability nor by judgment, but by the convenience. A sample obtained from a readily available list, such as telephone directories or automobile registrations (not the complete list of these), is a convenience sample and not a random sample, even if the sample is drawn at random from the list.

Advantage:

Like judgment sampling, this method is also very simple and convenient for analyzing and obtaining quick results.

Disadvantage:

Since this type of sample is a convenient part of whole population, it can hardly be representative of population; hence one can not be sure whether the sampling units included in this type of sample are representative of the target population.

Applications:

Formerly this method was frequently used in public opinion surveys when interviewers stopped near the railway station or the bus stop or in front of the office nearby building to interview people. However, accountants still use this sampling method to analyze or audit accounts. This is also useful in making pilot survey - questions may be tested and preliminary information may be obtained from the mass before the final sampling design is constructed.

14.4.7. Snowball Sampling. The term 'snowball' comes from the analogy of a snowball, which begins small but becomes bigger and bigger as it rolls

downhill. The 'snowball sampling' has been used to describe a sampling procedure in which the sample goes on becoming bigger and bigger as the observation or study proceeds. For example, an opinion survey is to be conducted on smokers of a particular brand of tobacco. At the first stage, we may pick up a few persons who are known to us or can be identified to be the smokers of that brand. At the time of interviewing them, we may obtain the names of other persons known to the first stage respondents. Thus, the respondents of a stage are serving as informants for the identification of more respondents and sample goes on increasing.

Snowball sampling, which is generally considered to be non-probabilistic, can be converted into probabilistic by selecting the subjects randomly within each stage. For a non-probabilistic sample, some methods such as quota sampling can be used at each stage.

15.5. Sampling and Non-sampling Errors

As we have mentioned earlier, the results obtained from a sample will not obviously be exactly same as the results obtained from a population. The term error means the difference between the value of a sample statistic and that of corresponding population parameter. A number of factors may be responsible for this error. In accordance with the sources, this error is classified into sampling and non-sampling error.

15.5.1. Sampling Error. Error or variation among sample statistics due to chance, that means, the differences between each sample and the population, and among several samples, which are solely due to elements happened to choose for the sample. Sampling errors arise due to the fact that a particular method of sampling is used in selecting the items from the population which may not be correct method. Hence, the complete enumeration will not possess any sort of sampling 'error, because in this case, the whole population is studied and no question of drawing sample arises. Sampling errors are again of two types, such as biased and unbiased errors.

Definition. Sampling Error. The error due to drawing inference about the population on the basis of a sample is termed as sampling error.

Biased error arises due to faulty process of selection, faulty work during the collection of information and faulty method of analysis. Faulty process of selection may arise in a number of ways such as, deliberate selection of a sample, conscious or unconscious fault in the selection of random sample, substitution, non-response, etc. Faulty work during collection of information may include - poorly designed questionnaire, ill-trained interviewer, failure of a respondent's memory, errors in measurements, etc. Again, bias in analysis may arise, particularly, due to faulty method of

analysis such as from improper use of statistical measurements, improper selection of models, etc.

15.5.2. Non-sampling Error. When a complete enumeration of units in the population is made, one would expect that it would give rise to data free from any error. Unfortunately, it is not so in practice. For example, it is very difficult to avoid errors of observation or measurement. Again, in the processing of data, tabulations errors may be committed affecting the final results. Such errors are termed as non-sampling errors, because, they are due to factors other than the inductive process of inferring about the population from a sample. Thus, the data obtained in an investigation by complete enumeration, although free from sampling error, non-sampling errors can occur at every stage of planning and execution of the census or survey. This type of errors can arise due to a number of causes, such as defective methods of data collection and tabulation, faulty definition, incomplete coverage of population or sample, inappropriate questionnaire, etc. However, some of the major sources of non-sampling error can be pointed out as follows:

- i) Data specification being inadequate and inconsistent with respect to the objective of the study, whatever the study method is, census or survey.
- ii) Omission or duplication of units due to imprecise definition or boundaries of area units, incomplete or wrong identification of units, or faulty methods of enumeration.
- iii) Defective frame, faulty selection of sampling units. Inaccurate or inappropriate questionnaire, methods of interview, definition or instruction may also cause non-sampling error.
- iv) Lack of trained and experienced investigators,
- v) Lack of adequate inspection and supervision of primary staff
- vi) Errors due to non-response, that means, incomplete coverage in respect of units,
- vii) Errors in data processing operations such as coding, punching, certification, tabulation, etc.
- viii) Errors committed during presentation or printing of tabulated results
- ix) Errors in scrutiny of primary or basic data

Definition. Non-Sampling Error. The possible error which may arise at any stage of investigation, either in census or in sampling, is termed as non-sampling error. This type of error arises due to faulty questionnaire, due to non-response, due to faulty tabulation method, etc.

However, the non-sampling error tends to increase with the sample size, while sampling error decreases with increase of sample size. In case of complete enumeration, non-sampling errors and in case of sample survey,

both sampling and non-sampling errors require to be controlled and reduced to a level at which their presence does not distort the final results.

15.6. Sampling Distribution

One of the major objectives of the field of statistical analysis is to know the true or actual values of different parameters of the population under study. In this case, the ideal situation is to consider the entire population in determining these values. Sometimes that is not feasible due to cost, time, labour and other constraints. Hence we use a random sample to estimate the parameters of a population. This is done with the belief that characteristics of this random sample represent the similar characteristics of the population from which sample are taken. But in practice, the results obtained from a sample may or may not represent the population. This point can be illustrated with the following example.

Suppose that a large company owns a number of sales stations of its products. In order to promote the sales of its product, the company has advertised that any person purchasing more than 20 or more items of its product from any of its station will receive a television as a gift. The promotional period is limited to 15 days. However, due to the promotional costs and purchase cost of televisions, the company has decided to continue the promotion for a period of another 15 days, only if the average daily sales at each station are at least 1000 items. Fifteen of its stations are randomly selected during the first promotional period for survey regarding the daily sales and found that the average daily sales per station were 1015 items. The company has to make a decision as to whether it should continue the promotion based on this statistical data. In this case, the initial reaction would be to assume that since the sample mean of sales is in excess of 1000, and it represents the population mean, the condition set by the company is satisfactory and hence it should continue the promotional campaign. This decision would depend upon how close the sample mean is to population mean. But, it is not possible to determine this closeness from a single sample of fifteen stations. The sample mean could be close to population mean or it could be quite different.

Since it is not possible to determine the proximity of sample mean and the population mean from the given information, one can use the concept of sampling distribution to bring the value of sample mean close to population mean.

Again, the only information available to us is the observations in the sample and we must use some function of the sample values to estimate the parameters of the population. Such a function is called a statistic. A statistic is based on a random sample and since a number of samples of fixed size

can be drawn from a given population, the values of any statistic calculated from each sample will obviously vary from one sample to another. Each sample will lead to one value of the statistic and the totality of these values obtained from all possible samples of a given size constitutes a sampling distribution of statistic. That probability distribution of all possible values of a statistic is called a sampling distribution.

Suppose, all possible random samples of size n are drawn with replacement from a population of size N , and the mean values for all samples are computed. If the possible values of mean are arranged in the form of a distribution along with their probabilities, it is called sampling distribution of mean. That means the sampling distribution is the probability distribution of all possible values of a given statistic obtained from all the distinct possible samples of equal size drawn from a population. Thus, the form of sampling distribution depends on the nature of parent population and the size of the sample.

Parameter. The unknown constant or any function of them that appear in the mathematical specification of a population is known as parameter.

Any numerical quantity calculated from the population data is also called parameter.

Statistic. Any function of a random sample is known as statistic.

Any numerical quantity calculated from sample is also called statistic.

Sampling distribution. The probability distribution of a statistic derived from all possible random samples of a given population is called sampling distribution.

15.7. Concept of Standard Error

Suppose we wish to know about average daily working hours of the workers of a large industry with the help of sample. It is possible to take several samples of workers of particular size from the population. If we calculate average daily working hours for all samples, it would be highly unlikely that all of these sample means would be the same; some variability in the samples means would be observed. This variability in the sample statistics results from sampling error due to chance error in sampling process. This variability occurs because there are differences between each sample and the population, and among the selected samples. The standard deviation of the distribution of sample means measures the extent to which we expect the means from the different samples to vary because of chance error in the sampling process. Thus, the standard deviation of a statistic is called its standard error. For example, standard deviations of sample mean, sample

median, sample proportion etc. are known as the standard errors of mean, median and proportion, etc.

Standard Error. The positive square root of the variance of a statistic (sample mean, sample median, sample proportion, etc) is known as the standard error of the statistic.

Suppose, X_1, X_2, \dots, X_n constitute a random sample of size n from a normal population with mean μ and variance σ^2 , \bar{X} denotes the mean of the random sample, then the expectation and standard error of sample mean are given by μ and $\frac{\sigma}{\sqrt{n}}$ respectively.

15.7.1. Importance of standard error. Standard error plays an important role in statistical inference, such as estimation and test of hypothesis. All of the test statistics are defined based on standard error of the statistic. Its importance in estimation is cited below:

- i) Standard error gives an index of the precision of the estimate of the parameter.
- ii) Standard error enables us to determine the probable limits within which the population parameter may expect to lie.

15.8. Central Limit Theorem

The central limit theorem states that 'Regardless of the shape of the population, the distribution of the sample means approaches the normal probability distribution as the sample size increases.' In practice, the sample size of 30 or more is considered adequate for this purpose. However, regardless of the sample size, the sampling distribution would be normal, if the original population is normally distributed.

Statement of Central limit theorem. Let X_1, X_2, \dots, X_n be a random sample from a population having mean μ and variance σ^2 . Let \bar{X} be the sample mean, then central limit theorem states that as n becomes large the sampling

distribution of $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ approaches to the standard normal distribution

whatever may be the form of the distribution.

That means, $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

It can be seen that the possible values of sample means tend to be close to the population mean, and according to the central limit theorem, the distribution of these sample means tend to be approximately normal for a sample size larger than 30.

Remarks.

1. Central limit theorem holds only if the mean and variance of the distributions from which the random sample drawn exist.
2. If the random sample has been drawn from the normal population then the sampling distribution of the $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is exactly $N(0,1)$ for any sample size n .

15.9. Some Important Sampling Distributions

The important sampling distributions are

- i) χ^2 - distribution (Chi-square distribution)
- ii) t-distributions
- iii) F-distributions
- iv) Distribution of sample mean
- v) Distribution of difference between two sample means
- vi) Distribution of sample proportion.

These distributions play important roles in test of hypothesis. All these distributions are derived from the normal distribution. Now we shall give a brief survey of these distributions. The applications of these statistics are mentioned in chapter-16.5.

15.9.1. Chi-square (χ^2) distribution. The sum of squares of n independent standard normal variates is called chi-squares with n degrees of freedom. Let Z_1, Z_2, \dots, Z_n be n independent standard normal variables, then chi-square denoted χ^2 is defined as

$$\chi_n^2 = \sum_{i=1}^n Z_i^2.$$

However, if X_1, X_2, \dots, X_n are n independently and identically distributed random variables each of which is normally distributed with mean μ and variance σ^2 , then $\chi_n^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2$ is distributed as χ_n^2 with n df.

The probability density function of χ^2 with n degrees of freedom is

$$f(\chi^2) = \frac{1}{2^{n/2} \Gamma(\frac{n}{2})} e^{-\frac{\chi^2}{2}} (\chi^2)^{n/2-1}; \chi^2 > 0$$

Important Properties of χ^2 distribution

- i) The distribution contains only one parameter which is the degree of freedom of the distribution.
- ii) The mean of the distribution is n and the variance is $2n$.
- iii) The mode of the distribution is $n-1$.
- iv) It is positively skewed distribution for smaller values of n ; the distribution becomes symmetrical as n tends to infinity.

15.9.2. Student's t-Distribution. Let X_1, X_2, \dots, X_n be random sample from a normal distribution with mean μ and variance σ^2 , then \bar{x} is normally distributed with mean μ and variance σ^2/n . Now, if the estimators of μ and variance σ^2 are given by $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ respectively,

Then the statistic t is defined as $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ which follows Student's t distribution with $n-1$ degrees of freedom (df).

A continuous random variable t is said to have a t-distribution with n df if its probability density function is given by

$$f(t) = \frac{1}{\sqrt{n} B(1/2, n/2)} \left(1 + t^{2/n}\right)^{-\frac{n+1}{2}}; -\infty < t < \infty$$

15.9.2.1. Properties of t distribution.

- i) The distribution has only one parameter which is the degree of freedom of the distribution.
- ii) The distribution symmetric about mean zero and variance is $n/(n-2)$ and all odd order moments of t-distribution are zero.
- iii) Since, the distribution is symmetric at mean $t = 0$, hence, the mean, median and mode are all zero.
- iv) If the degree of freedom increases, t-distribution tends to normal distribution. Actually, t-distribution tends to normal distribution when $n > 30$.

Remarks. t-tests are called small sample tests $n < 30$.

19.9.3. F- Distribution. If $X_{11}, X_{21}, \dots, X_{1n_1}$ be a random sample of size n_1 drawn from a normal population with mean μ_1 and variance σ_1^2 , and $X_{21}, X_{22}, \dots, X_{2n_2}$ be another random sample of size n_2 drawn from a normal population with mean μ_2 and variance σ_2^2 . Let \bar{x}_1 and s_1^2 are the estimators of μ_1 and σ_1^2 , and \bar{x}_2 and s_2^2 are the estimators of μ_2 and variance σ_2^2 respectively, defined by $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$, $s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2$,

$$\bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}, s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2$$

Thus, $\chi_1^2 = \frac{(n_1 - 1)s_1^2}{\sigma_1^2}$ is a χ^2 -variate with $(n_1 - 1)$ df

and $\chi_2^2 = \frac{(n_2 - 1)s_2^2}{\sigma_2^2}$ is a χ^2 -variate with $(n_2 - 1)$ df ..

Since the two samples are independent, these χ^2 -variates are also independent. Thus, the ratio of two independent chi-squares divided by their respective degrees of freedom is called F-variate and it is defined as :

$$F = \frac{\chi_1^2 / n_1 - 1}{\chi_2^2 / n_2 - 1} = \frac{s_1^2}{s_2^2}$$

Which follows Snedecor's F with $n_1 - 1 = \gamma_1$ and $n_2 - 1 = \gamma_2$ degrees of freedom.

The density function of F with γ_1 and γ_2 df is given by

$$f(F) = \frac{\left(\frac{\gamma_1}{\gamma_2}\right)^{\frac{\gamma_1}{2}}}{\beta\left(\frac{\gamma_1}{2}, \frac{\gamma_2}{2}\right)} \frac{F^{\frac{\gamma_1-1}{2}}}{\left(1 + \frac{\gamma_1}{\gamma_2} F\right)^{\frac{\gamma_1+\gamma_2}{2}}}; \quad F \geq 0$$

Remark. The sampling distribution of F-statistic does not involve any population parameters and depends only on the degrees of freedom.

Properties of F-distribution

- i) The distribution contains two parameters which are the degrees of freedom of the distribution.
- ii) The mean and variance of F-distribution are

$$\text{Mean} = \frac{\gamma_2}{(\gamma_1 + \gamma_2)} \text{ and}$$

$$\text{Variance} = \text{var}(F) = \frac{2\gamma_2^2(\gamma_1 + \gamma_2 - 2)}{\gamma_1(\gamma_1 - 2)^2(\gamma_2 - 4)}; \quad \gamma_1 > 2 \text{ and } \gamma_2 > 4.$$

- iii) The mode of the distribution is $\text{Mode} = \frac{\gamma_2(\gamma_1 - 2)}{(\gamma_2 + 2)\gamma_1}$ which less than unity is always. It means mode of the distribution exists if $\gamma_1 > 2$.
- iv) The distribution is positively skewed.

15.9.4. Sampling distribution of sample mean. Let X_1, X_2, \dots, X_n be a random sample from a population with mean μ and variance σ^2 . Then the sample mean is $\bar{X} = \frac{1}{n} \sum X_i$.

Mean of sample mean. It can be easily shown that the expectation of the sample mean is equal to the population mean. That is

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) = \frac{n\mu}{n} = \mu.$$

Hence, the mean of the sampling distribution of sample means is the population mean. This is an important result of random sampling and indicates the protection that random samples provide against unrepresentative samples. A single sample mean could be larger or smaller than the population mean. However, on average, there is no reason for us to expect a sample mean that is either higher or lower than the population mean.

Variance of sample mean. We know that the variance of a linear combination of independent random variables is the sum of the linear coefficients squared times the variance of the random variables. It follows that

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n\right) = \sum\left(\frac{1}{n}\right)^2 \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

And the corresponding standard error of mean is given by $\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

which shows that the variance or standard error of mean decreases as the sample size n increases. The above results are also true for all possible samples of size n drawn with replacement from a finite population of size N .

Theorem 15.2. If all possible random samples of size n are drawn with replacement from a finite population of size N with mean μ and standard

deviation σ , then the sampling distribution of the mean \bar{X} follows a distribution with mean μ and standard deviation σ/\sqrt{n} .

Remarks. According to central limit theorem, for large n , $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

follows a standard normal variate with mean zero and variance one. Now we shall prove the theorem with the help of an example.

Example 15.9.1. Suppose a population consists with four values 0, 1, 2, 3. Draw all possible of size 2 with replacement and show that the sample mean follows the above the theorem.

Solution. The population mean is $\mu = \frac{0+1+2+3}{4} = 1.5$.

$$\begin{aligned}\text{Population variance } \sigma^2 &= \frac{1}{N} \sum (x - \mu)^2 \\ &= \frac{1}{4} [(0 - 1.5)^2 + (1 - 1.5)^2 + (2 - 1.5)^2 + (3 - 1.5)^2] \\ &= \frac{2.25 + 0.25 + 0.25 + 2.25}{4} = \frac{5}{4} = 1.25.\end{aligned}$$

All possible samples and their means are shown in the table given below:

Sample No.	Sample	\bar{x}	Sample No.	Sample	\bar{x}
1	0, 0	0	9	2, 0	1.0
2	0, 1	0.5	10	2, 1	1.5
3	0, 2	1.0	11	2, 2	2.0
4	0, 3	1.5	12	2, 3	2.5
5	1, 0	0.5	13	3, 0	1.5
6	1, 1	1.0	14	3, 1	2.0
7	1, 2	1.5	15	3, 2	2.5
8	1, 3	2.0	16	3, 3	3.0

The sampling distribution of \bar{X} is :

\bar{x}	f	$p(\bar{x}) = \frac{f}{k}$
0	1	1/16
0.5	2	2/16
1.0	3	3/16
1.5	4	4/16
2.0	3	3/16
2.5	2	2/16
3.0	1	1/16

Here k is the number of samples. $\frac{f}{k}$ is the relative frequency and the ratio's are the probabilities for different values of \bar{X} .

Mean of \bar{X} is given by

$$\begin{aligned} E(\bar{X}) &= \sum \bar{x}_i p(\bar{x}_i) = 0 \times 1/16 + 0.5 \times 2/16 + 1 \times 3/16 + 1.5 \times 4/16 \\ &\quad + 2 \times 3/16 + 2.5 \times 2/16 + 3 \times 1/16 \\ &= \frac{1+3+6+6+5+3}{16} = \frac{24}{16} = 1.5 = \mu = \text{population mean.} \end{aligned}$$

$$\begin{aligned} \text{Variance of } \bar{X} &= \sigma^2(\bar{X}) = \sum (\bar{x} - \mu)^2 p(\bar{x}) \\ &= (0 - 1.5)^2 \times \frac{1}{16} + (0.5 - 1.5)^2 \times \frac{2}{16} + (1 - 1.5)^2 \times \frac{3}{16} + (1.5 - 1.5)^2 \times \frac{4}{16} \\ &\quad + (2 - 1.5)^2 \times \frac{3}{16} + (2.5 - 1.5)^2 \times \frac{2}{16} + (3 - 1.5)^2 \times \frac{1}{16} \\ &= \frac{1}{16} (2.25 + 2 + 0.75 + 0.75 + 2 + 2.25) = \frac{10}{16} = \frac{5}{8} = \frac{1.25}{2} = \frac{\sigma^2}{n}. \end{aligned}$$

Theorem 15.9. 3. If all possible random samples of size n are drawn without replacement from a finite population of size N with mean μ and standard deviation σ , then the sampling distribution of the mean \bar{X} follows a distribution with mean μ and standard error $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$.

Remarks. It is to be noted that the sampling distribution of the sample mean follows the above theorem if n observations are taken from N population observations one by one without replacement or n observations are taken at a time from the N observations. In first case the total number of samples is $k = N(N-1)(N-2)\dots(N-n+1)$. In second case the total number of samples is N_{C_n} . Although the total number of samples in first case is $n!$ times the total number of samples in second case, but the distribution of sampling mean remains the same.

Example 15.9.2. Suppose a population consists with four values 0, 1, 2 and 3. Draw all possible of size 2 without replacement and show that the sample mean follows the above the theorem..

All possible samples and their means are shown in the table given below:

Sample No.	Sample	\bar{x}	Sample No.	Sample	\bar{x}
1	0, 1	0.5	7	1, 0	0.5
2	0, 2	1.0	8	2, 0	1.0
3	0, 3	1.5	9	3, 0	1.5
4	1, 2	1.5	10	2, 1	1.5
5	1, 3	2.0	11	3, 1	2.0
6	2, 3	2.5	12	3, 2	2.5

The sampling distribution of \bar{X} is

\bar{x}	f	$p(\bar{x}) = \frac{f}{k}$
0.5	2	1/6
1.0	2	1/6
1.5	4	1/3
2.0	2	1/6
2.5	2	1/6

Here k is the number of samples. $\frac{f}{k}$ is the relative frequency and the ratio's

are the probabilities for different values of \bar{X} . Here k = 12

$$\text{Mean of } \bar{X} = E(\bar{X}) = \sum \bar{x}_i p(\bar{x}_i)$$

$$= 0.5 \times 1/6 + 1.0 \times 1/6 + 1.5 \times 1/3 + 2 \times 1/6 + 2.5 \times 1/6$$

$$= \frac{0.5 + 1 + 3 + 2 + 2.5}{6} = \frac{9}{6} = 1.5$$

$$\text{Variance of } \bar{X} = \sigma^2(\bar{X}) = \sum (\bar{x} - \mu)^2 p(\bar{x})$$

$$= (0.5 - 1.5)^2 \times \frac{1}{6} + (1 - 1.5)^2 \times \frac{1}{6} + (1.5 - 1.5)^2$$

$$\times \frac{1}{3} + (2 - 1.5)^2 \times \frac{1}{6} + (2.5 - 1.5)^2 \times \frac{1}{6}$$

$$= \frac{1}{6} (1 + 0.25 + 0 + 0.25 + 1) = \frac{2.5}{6} = \frac{5}{12}$$

$$= \frac{5}{12} \left(\frac{4-2}{4-1} \right) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

The term $(N - n)/(N - 1)$ is often termed as finite population correction factor. This formula can be made simpler to use by the fact that we

generally deal with very large populations, which can be considered as infinite, so that the population size N is very large and sample size is very small, then $(N-n)/(N-1)$ would approach 1, then we can use the formula

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

We have the same probability distribution of the sample mean if we take $4C_2 = 6$.

The possible samples, sample means are

Sample No.	Sample value	Values of Sample mean \bar{X}
1	0, 1	0.5
2	0, 2	1.0
3	0, 3	1.5
4	1, 2	1.5
5	1, 3	2.0
6	2, 3	2.5

The sampling distribution of the sample mean is

The sampling distribution of \bar{X} is

\bar{x}	f	$p(\bar{x}) = \frac{f}{k}$
0.5	1	1/6
1.0	1	1/6
1.5	2	1/3
2.0	1	1/6
2.5	1	1/6

Properties of sampling distribution of sample mean. Let \bar{X} denotes the sample mean of a random sample of n observations from a population with mean μ and variance σ^2 , then

- The sampling distribution of \bar{X} has mean $E(\bar{X}) = \mu$.
- The sampling distribution of \bar{X} has standard error $\frac{\sigma}{\sqrt{n}}$.
- However, if the sample size n is not small compared to the population size, N , and the population size is finite and the sampling is drawn without replacement, then the standard error of \bar{X} is given

$$\text{by } \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

iv. If the parent population distribution is normal, thus sampling distribution of \bar{X} is also normal, then the random variable $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows a normal distribution with mean 0 and standard deviation 1.

However, if the sample size n is small (< 30) and the population variance σ^2 is not known, it is to be estimated using the formula $s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$ and the standard error of the mean is given by s/\sqrt{n} , then the mean \bar{X} will not follow normal distribution, rather it will follow t-distribution, and the corresponding statistic is defined as $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ which is distributed as Student's t with $n-1$ degrees of freedom. Again, if sample size is large, this t tends to standard normal distribution.

Let us illustrate the concept of sampling distribution with different examples.

Example 15.9.3. The MBA class has a total of 60 students. Their average score in statistics after final term was 70 with a standard deviation of 8. A sample of 36 students is taken at random from this class. Calculate the standard error of the mean for this sample.

Solution. Here $N = 60$, $n = 36$ and $\sigma = 8$, since the sample size is not a small portion of population size, the standard error of mean is given by

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{8}{\sqrt{36}} \sqrt{\frac{60-36}{60-1}} = 0.85.$$

Example 15.9.4. A large bag contains some counters, 60% of the counters have the number 0 on them and 40% have the number 1. A random sample of 3 counters is taken from the bag, find the sampling distribution of sample mean and sample mode.

Solution. Let X be the number of counters which can take value 0 and 1. The distribution of population is given by

Values of $X : x$:	0	1
$p(x)$:	0.6	0.4

And the population mean is $E(X) = 0 \times 0.6 + 1 \times 0.4 = 0.4$.

The possible samples of size 3 are

$(0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)$

$$p(\bar{X} = 0) = (0.6)^3 = 0.216 \text{ for } (0, 0, 0) \text{ case}$$

$$p(\bar{X} = 1/3) = 3 \times 0.4 \times 0.6^2 = 0.432 \text{ for } (1, 0, 0), (0, 1, 0), (0, 0, 1) \text{ cases}$$

$$p(\bar{X} = 2/3) = 3 \times 0.4^2 \times 0.6 = 0.288 \text{ for } (1, 1, 0), (1, 0, 1), (0, 1, 1) \text{ cases}$$

$$p(\bar{X} = 1) = (0.4)^3 = 0.064 \text{ for } (1, 1, 1) \text{ case}$$

The sampling distribution of sample mean \bar{X} is

$\bar{X} : \bar{x}$	0	1/3	2/3	1
$p(\bar{x})$:	0.216	0.432	0.288	0.064

Thus $E(\bar{X}) = 0 \times 0.216 + 1/3 \times 0.432 + 2/3 \times 0.288 + 1 \times 0.064 = 0.4$ which is exactly same as the population mean μ .

Similarly, from the list of possible samples it is clear that mode can take two values 0 or 1. Hence the sampling distribution of mode (M) is given by,

M :	0	1
P(M) :	0.648	0.352

[Hints: $p(\text{Mode} = 0) = (0.6)^3 + 3 \times 0.4 \times 0.6^2 = 0.648$ for first four cases

$p(\text{Mode} = 1) = 3 \times 0.4^2 \times 0.6 + 0.064 = 0.352$ for last four cases]

Example 15.9.5. Suppose a baby sitter has 5 children under her supervision with average age of 6 years. The individual ages of five children are $X_1 = 2$, $X_2 = 4$, $X_3 = 6$, $X_4 = 8$, $X_5 = 10$ years. If a sample of 2 children is selected at random, find the sampling distribution of mean age.

Solution. Here we can consider ${}^5C_2 = 10$ samples in place of $5 \times 4 = 20$ samples. The possible samples, sample means and their probabilities are given below:

Sample No.	Sample Individual	Sample age	Mean age	Probability
1	X_1, X_2	(2, 4)	$\bar{x} = 3$	$p(\bar{x} = 3) = 1/10 = 0.1$
2	X_1, X_3	(2, 6)	$\bar{x} = 4$	$p(\bar{x} = 4) = 1/10 = 0.1$
3	X_1, X_4	(2, 8)	$\bar{x} = 5$	$p(\bar{x} = 5) = 1/10 = 0.1$
4	X_1, X_5	(2, 10)	$\bar{x} = 6$	$p(\bar{x} = 6) = 1/10 = 0.1$
5	X_2, X_3	(4, 6)	$\bar{x} = 5$	$p(\bar{x} = 5) = 1/10 = 0.1$
6	X_2, X_4	(4, 8)	$\bar{x} = 6$	$p(\bar{x} = 6) = 1/10 = 0.1$
7	X_2, X_5	(4, 10)	$\bar{x} = 7$	$p(\bar{x} = 7) = 1/10 = 0.1$
8	X_3, X_4	(6, 8)	$\bar{x} = 7$	$p(\bar{x} = 7) = 1/10 = 0.1$
9	X_3, X_5	(6, 10)	$\bar{x} = 8$	$p(\bar{x} = 8) = 1/10 = 0.1$
10	X_4, X_5	(8, 10)	$\bar{x} = 9$	$p(\bar{x} = 9) = 1/10 = 0.1$

The probability distribution of the sample mean, referred to the sampling distribution of mean, is given by

Sample mean (\bar{x})	3	4	5	6	7	8	9
$p(\bar{x})$	0.1	0.1	0.2	0.2	0.2	0.1	0.1

Here the population mean is

$$\mu = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6.$$

$$\begin{aligned}\text{Variance } \sigma^2 &= \frac{1}{5} [(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2] \\ &= \frac{16+4+0+4+16}{5} = \frac{40}{5} = 8\end{aligned}$$

The mean of the sample mean =

$$\begin{aligned}E[\bar{X}] &= 3 \times .1 + 4 \times .1 + 5 \times .2 + 6 \times .2 + 7 \times .2 + 8 \times .1 + 9 \times .1 \\ &= .3 + .4 + 1.0 + 1.2 + 1.4 + .8 + .9 = 6.0.\end{aligned}$$

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \sum (\bar{x} - \mu)^2 p(\bar{x}) = 9 \times .1 + 4 \times .1 + 1 \times .2 + 0 + 1 \times .2 + 4 \times .1 + 9 \times .1 \\ &= 3.0 = \frac{8}{2} \left(\frac{5-2}{5-1} \right) = \frac{\sigma^2}{2} \left(\frac{N-n}{N-1} \right).\end{aligned}$$

Example 15.9.6. Suppose the experiences in years of six employees are given as 2, 4, 6, 6, 7, 8. Find the sampling distribution of means for random samples of size 2.

Solution. The number of samples of size two is ${}^6C_2 = 15$. The years of experience of possible 15 samples of 2 employees are listed below along with sample means

Sample No.	Sample	Sample mean	Sample No.	Sample	Sample mean
1	2, 4	3.0	9	4, 8	6.0
2	2, 6	4.0	10	6, 6	6.0
3	2, 6	4.0	11	6, 7	6.5
4	2, 7	4.5	12	6, 8	7.0
5	2, 8	5.0	13	6, 7	6.5
6	4, 6	5.0	14	6, 8	7.0
7	4, 6	5.0	15	7, 8	7.5
8	4, 7	5.5			

Thus the sampling distribution of sample means from the employees population for sample of size 2 is

Sample Mean (\bar{x})	Probability of \bar{x}
3.0	1/15
4.0	2/15
4.5	1/15
5.0	3/15
5.5	1/15
6.0	2/15
6.5	2/15
7.0	2/15
7.5	1/15

In this case too, it can be easily shown that the mean of sampling distribution is exactly same as the population mean $\mu = 5.5$.

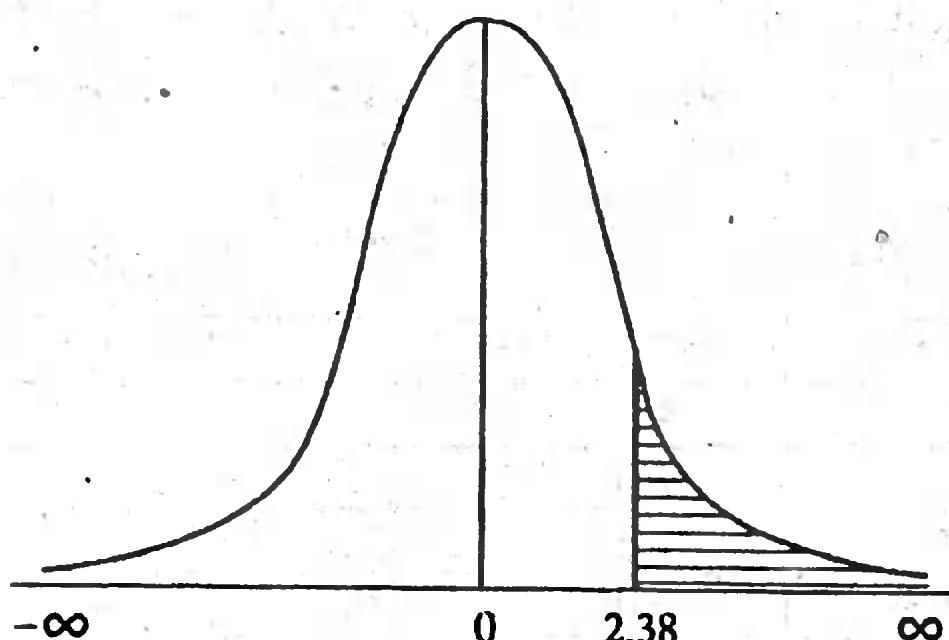
Example 15.9.7. The time between two arrivals in a queuing process of cars in a busy road is normally distributed with a mean 2 minutes and standard deviation 0.25 minutes. If a random sample size 36 of such cars is taken, what is the probability that the sample mean will be greater than 2.1 minutes?

Solution. Since the population is normally distributed, therefore, the sampling distribution of the sample mean will follow a normal distribution

with mean $\mu_{\bar{x}} = 2$ and standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.25}{\sqrt{36}} = 0.042$.

Thus, the sampling distribution of \bar{X} is given by $\bar{X} \sim N(2, 0.042^2)$

Therefore the probability that the sample mean will be greater than 2.1 minutes is given by

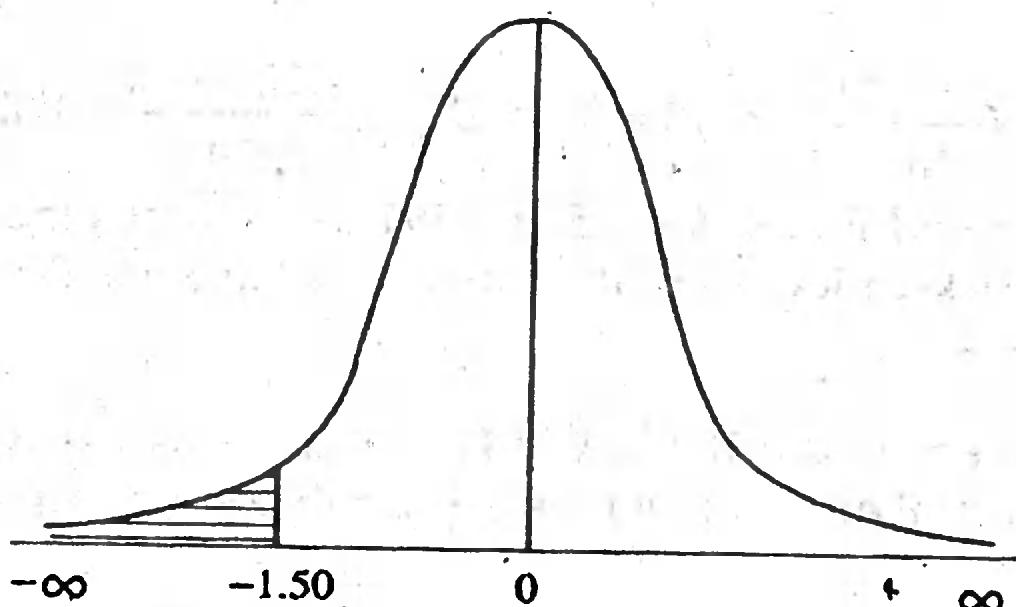


$$P(\bar{X} \geq 2.1) = P(Z \geq 2.38) = 1 - P(Z < 2.38) = 1 - 0.9913 = 0.0087.$$

Example 15.9.8. A spark plug manufacturer claims that the lives of its plugs are normally distributed with mean 36000 miles and standard deviation 4000 miles. A random sample of 16 plugs had an average life of 34500 miles. If the manufacturer's claim is correct, what is the probability of finding a sample mean of 34500 or less?

Solution. To compute the probability, we need first to obtain the standard error of the sample mean, which is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{16}} = 1000.$$



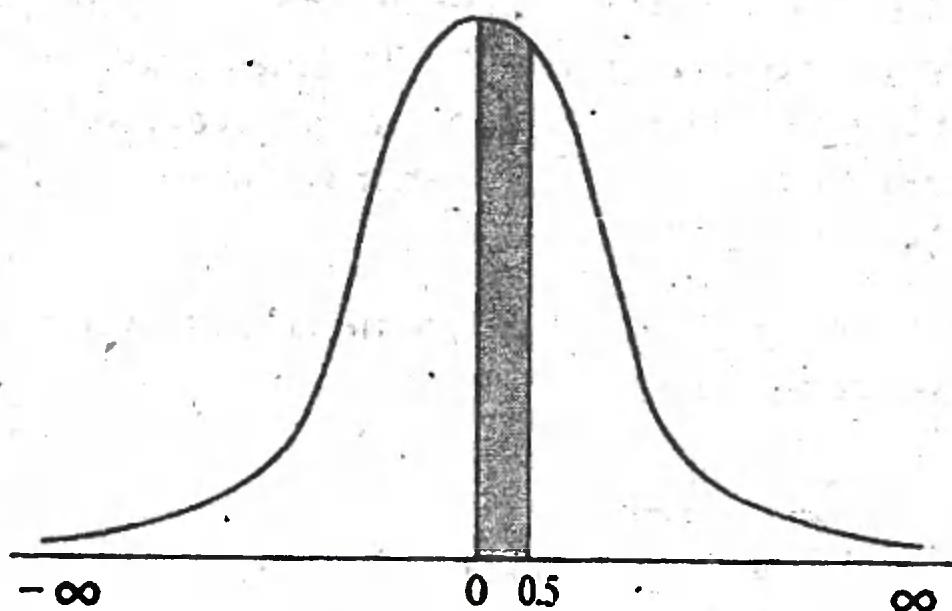
Hence, the desired probability is

$$P(\bar{X} < 34500) = P\left(Z < \frac{34500 - 36000}{1000}\right) = P(Z < -1.50) = 0.0668$$

which suggests that if the manufacturer's claims – $\mu = 36000$ and $\sigma = 4000$ – are true, then a sample mean of 34500 or less has a small probability.

Example 15.9.9. The weights of packets of cosmetics are normally distributed with mean 120 pounds and standard deviation 10 pounds. (i) What is the probability that the weight of any packet chosen at random is between 120 and 125 pounds? (ii) If a random sample of 25 packets is taken, what is the probability that the mean of this sample will be between 120 and 125?

Solution. (i) Let X represents the weights of the packets, given that $X \sim N(120, 10^2)$. We have to find $P(120 < X < 125)$. Using the standardized normal distribution formula $Z = \frac{X - \mu}{\sigma}$.

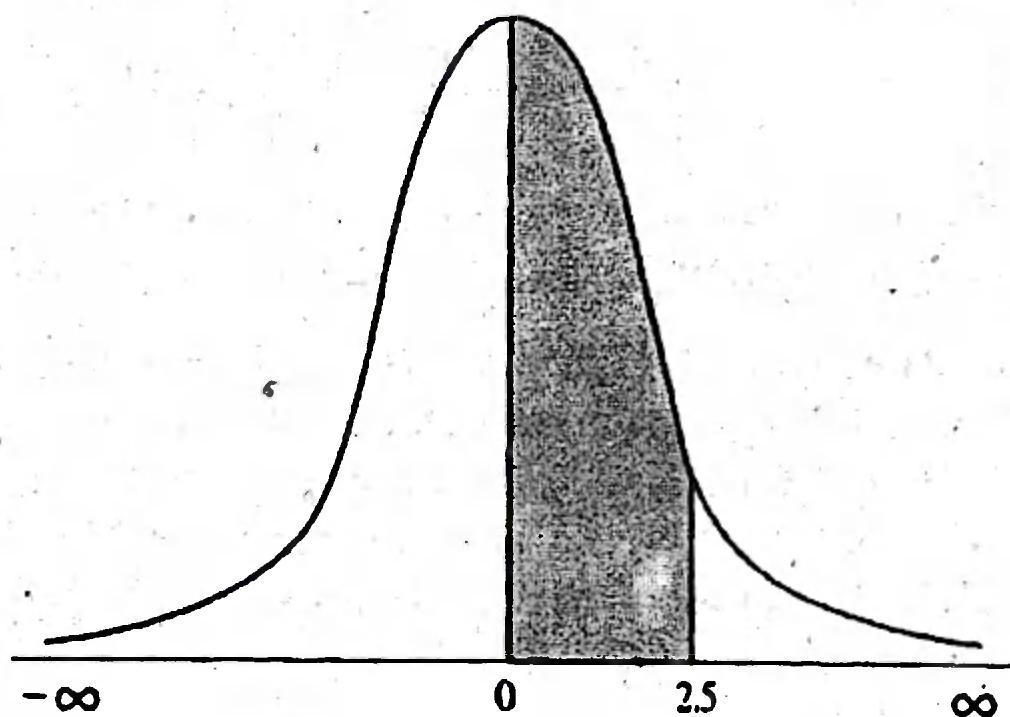


We have, $Z = \frac{125 - 120}{10} = 5/10 = 0.5$ and $Z = \frac{120 - 120}{10} = 0$, thus, in terms

of Z we have to find $P(0 < X < 0.5) = 0.1915$, this means there is 19.15 % chance that a packet picked up at random will have weight between 120 and 125 pounds.

(ii) Here we have to find $P(120 < \bar{X} < 125)$, that means we have to use the standardized normal distribution for sampling distribution of sample mean defined as

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$



$$\begin{aligned} \text{So, we have } P(120 < \bar{X} < 125) &= P\left[\frac{120 - 120}{10 / \sqrt{25}} \leq Z \leq \frac{125 - 120}{10 / \sqrt{25}}\right] \\ &= P(0 < Z < 2.5) = 0.4938. \end{aligned}$$

Which shows there is a chance of 49.38% that the sample mean will be between 120 and 125 pounds? It is also clear from here that the chance of a

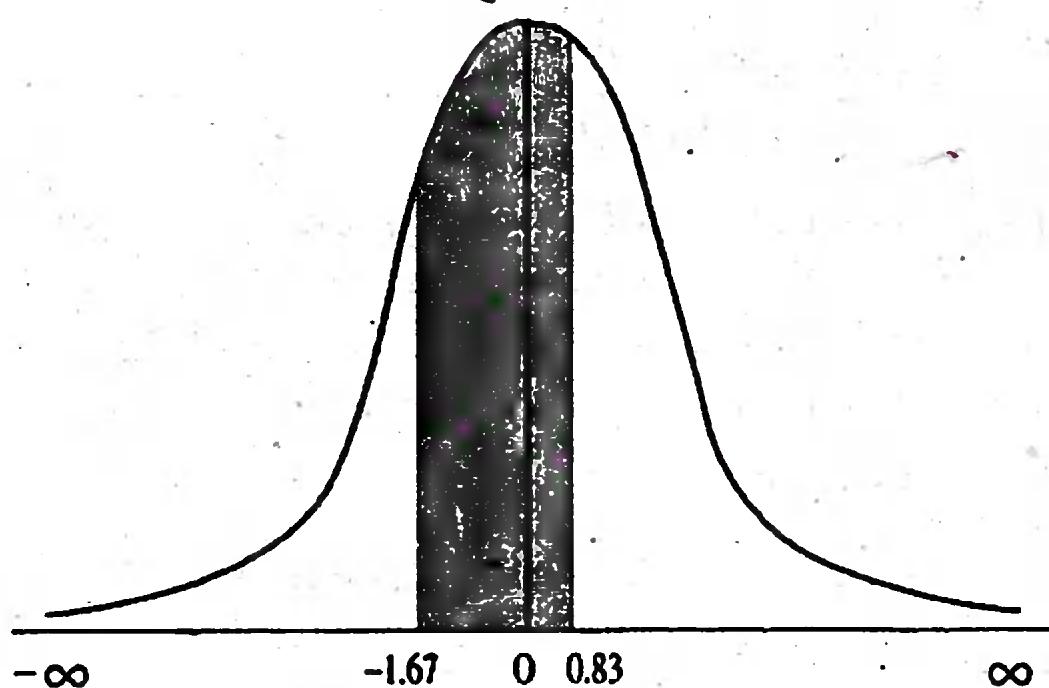
sample mean being between 120 and 125 pounds much higher than the probability of an individual packet having weight between 120 and 125 pounds and the probability of it increases as the sample size increases.

Example 15.9.10. A bank calculates that its individual savings deposits are normally distributed with mean Tk. 2000 and a standard deviation of Tk. 600. If the bank takes a random sample of 100 accounts, what is the probability that mean deposit will lie between Tk. 1900 and Tk. 2050?

Solution. Let X be the individual savings deposit, given that $X \sim N(2000, 600^2)$. We have to find $P(1900 < \bar{X} < 2050)$. Using the standardized normal distribution for sampling distribution of sample mean we have,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Here, standard error of \bar{X} is given by $\frac{\sigma}{\sqrt{n}} = \frac{600}{\sqrt{100}} = 60$,



So, the required probability is

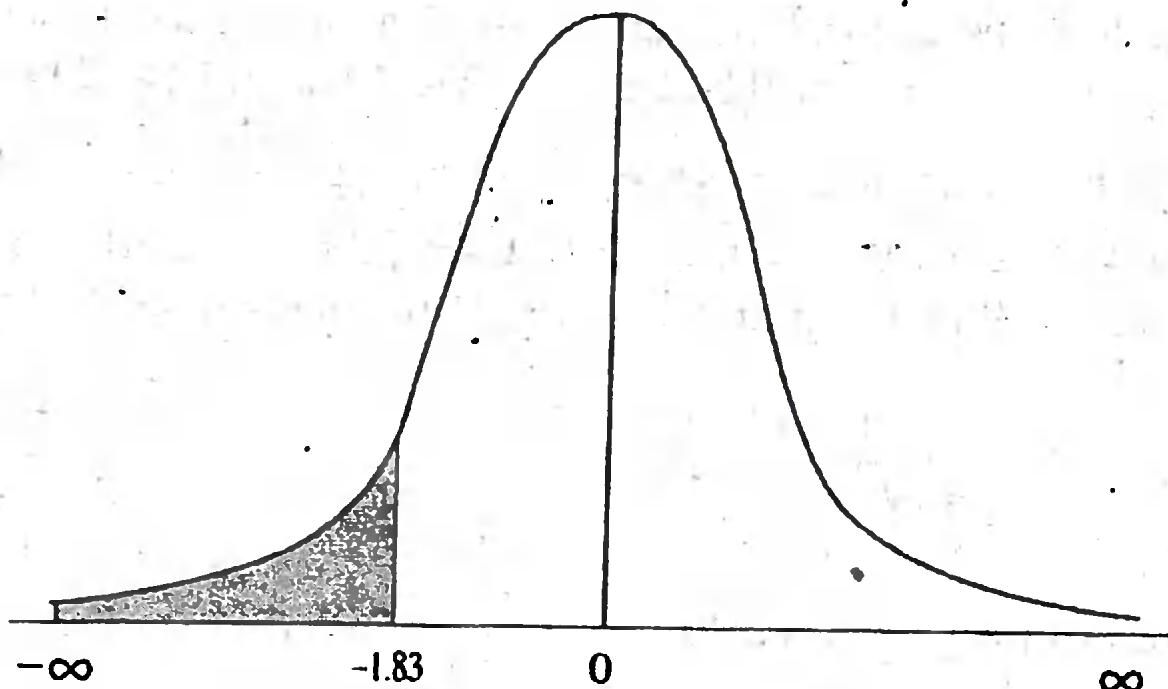
$$\begin{aligned} P(1900 < \bar{X} < 2050) &= P\left(\frac{1900 - 2000}{60} \leq Z \leq \frac{2050 - 2000}{60}\right) \\ &= P(-1.67 < Z < 0.83) \\ &= P[Z < 0.83] - P[Z < -1.67] = 0.7967 - 0.0475 = 0.7492. \end{aligned}$$

Example 15.9.11. Suppose that the annual percentage salary increases for the managers of a large industry are normally distributed with mean 12.2% and standard deviation 3.6%. A random sample of nine managers is obtained from this population and the sample mean computed. What is the probability that the mean will be less than 10%?

Solution. Given $\mu = 12.2$, $\sigma = 3.6$ and $n = 9$

Let \bar{X} denote the sample mean, and computing standard deviation of sample mean we have,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.6}{\sqrt{9}} = 1.2, \text{ thus, we have to compute,}$$



$$P(\bar{X} < 10) = P\left(Z < \frac{10 - 12.2}{1.2}\right) = P(Z < -1.83) = 0.0336$$

15.9.5. Sampling Distribution of the difference between two means.

Suppose $X_{11}, X_{12}, \dots, X_{1n_1}$ be a random sample of size n_1 taken from a normal population with mean μ_1 and variance σ_1^2 , and $X_{21}, X_{22}, \dots, X_{2n_2}$ be another independent sample of the same size n_2 taken from normal population with mean μ_2 and variance σ_2^2 . Let \bar{X}_1 and \bar{X}_2 be the sample means of two samples respectively defined by $\bar{X}_1 = \frac{1}{n_1} \sum X_{1i}$ and

$$\bar{X}_2 = \frac{1}{n_2} \sum X_{2i}.$$

(a) When the population variances are known, the sampling distribution of $(\bar{X}_1 - \bar{X}_2)$ follows exactly normal distribution with

i) Mean = $(\mu_1 - \mu_2)$ and

ii) Standard error = $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ whatever may be the values of n_1 and n_2 .

(b) When σ_1 and σ_2 are not known, and but $n_1 > 29$ and $n_2 > 29$ are sufficiently large, then standard error of the difference between two sample

means can be estimated by $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ where s_1^2 and s_2^2 are the variances obtained from two samples respectively. In this case, the sampling distribution of $(\bar{X}_1 - \bar{X}_2)$ follows approximately normal distribution with

i) Mean = $(\mu_1 - \mu_2)$ and

ii) Standard deviation = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ whatever may be the form of the parent populations from which the samples are drawn.

(c) When the sample sizes are small and population variances are equal but

unknown, then standard error of $(\bar{X}_1 - \bar{X}_2)$ is estimated by $s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$ is the pool estimate of the equal population variances and s_1^2, s_2^2 are the sample variances of two samples.

Then the sampling distribution of $(\bar{X}_1 - \bar{X}_2)$ follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom and the variate t is defined as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In all of the above cases, we have assumed that the parent distribution is normal. However, if the parent distribution is not normal and if the sample size is large, then by the virtue of central limit theorem, the distribution of the difference between two means follow normal distribution with respective mean and variance.

Example 15.9.12. Strength of wire produced by company A has a mean of 4500 kg and a standard deviation of 200 kg. Company B has a mean of 4000 kg and a standard deviation of 300 kg. If 50 wires of company A and 100 wires of company B are selected at random and tested for strength, what is the probability that the sample mean strength of A will be at least 600 kg more than that of B?

Solution. For the sampling distribution of the difference between two means, we know the mean value of the difference between two sample means is given by $4500 - 4000 = 500$.

and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{200^2}{50} + \frac{300^2}{100}} = 41.23$.

Thus, the desired probability is given by

$$\begin{aligned} P(\bar{X}_1 - \bar{X}_2 > 600) &= P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{600 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) \\ &= P\left(Z > \frac{600 - 500}{41.23}\right) \\ &= P(Z > 2.43) = 0.0075 \end{aligned}$$

Therefore, the probability that the sample mean strength of the wire produced by company A will be at least 600 kg more than that of B is given by 0.0075.

Example 15.9.13. A man buys 200 electric bulbs of each of two well known brands taken at random from stock for testing purposes: He finds that brand A has a mean life of 2560 hours with a standard deviation of 90 hours and brand B has a mean life of 2650 hours with standard deviation of 75 hours. Find the probability that average life of brand A is 110 hours less than that of brand A.

Solution. For the sampling distribution of the difference between two means, we know the mean value of the difference between two sample means is given by $(\mu_1 - \mu_2) = 2560 - 2650 = -90$ hours.

$$\text{and standard error } \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{90^2}{200} + \frac{75^2}{200}} = 8.248$$

Thus, the required probability is given by

$$\begin{aligned} P(\bar{X}_1 - \bar{X}_2 < -100) &= P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < \frac{-100 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) \\ &= P\left(Z < \frac{-110 + 90}{8.248}\right) = P(Z < -2.41) = 0.0080 \end{aligned}$$

15.9.6. Sampling distribution for proportion. Suppose a variable has two categories which follows binomial distribution with parameters n and π , and suppose a random sample of size n is taken from the population, where P is the proportion of a particular category of the variable. We know the mean and variance of distribution are $n\pi$ and $n\pi(1 - \pi)$ respectively. The

standard error of estimated p is given by $\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$.

Usually a large sample is considered for finding the distribution of sample proportion, so since the sample size is large, by the virtue of central limit theorem,

$$Z = \frac{P - \pi}{\sigma_p} = \frac{P - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \text{ is standard normal variate.}$$

That is Z defined above follows standard normal variate with mean zero and variance unity.

Example 15.9.14. It is known that 65% of items of lot are defectives, (i) what is the probability that a simple random sample of size 100 items will reveal that the proportions of defectives items to be 60% or less? (ii) how would this probability change if the sample size is increased to 500?

Solution. (i) The problem states that the population proportion of defectives items is 65%. This also means that if all possible samples of size 100 are taken from the population, then, the various sample proportions would be normally distributed with average proportion 65%, that means $E(P) = \pi = 0.65$, and the standard error of proportion P is

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} = \sqrt{\frac{0.65(1 - 0.65)}{100}} = 0.0477.$$

$$\text{Now we have to find } \Pr(P \leq 0.60) = \Pr\left(Z < \frac{0.60 - 0.65}{0.0477}\right) = \Pr(Z \leq -1.05) \\ = 1 - \Phi(1.05) = 1 - 0.8531 = 0.1469, \text{ which is the required probability.}$$

(ii) Again, when the sample size is increased to 500, then with the same parameter value of $\pi = 0.65$, the value of σ_p would be .0213, and that of Z defined above would be -2.35 , so the required probability is given by $\Pr(p \leq 0.60) = \Pr(Z \leq -2.35) = 1 - \Phi(2.35) = 0.0094$.

Example 15.9.15. 45% of the workers working in a garments factory are married. If a sample of size 200 of workers is selected at random, what is the probability that the proportion of married worker in this sample would be between 40% and 48%?

Solution. Distribution of sample proportion of workers of the factory would follow normal distribution with average proportion $\pi = 0.45$ and standard error,

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.45(1-0.45)}{200}} = 0.035.$$

Now, we have to find the $\Pr(0.40 < P < 0.48) = \Pr(-1.43 < Z < 0.86)$

$$= \Phi(0.86) - \Phi(-1.43) = 0.8051 - 0.0764 = 0.7287.$$

Example 15.9.16. A random sample of 250 homes was taken from a large population of older homes to estimate the properties of homes with unsafe wiring. If, in fact, 30% of the homes have unsafe wiring, what is the probability that sample proportion will be between 25% and 35% homes with unsafe wiring?

Solution. For the given problem, we have $\pi = 0.30$, $n = 250$.

We can compute the standard error of the sample proportion, p as

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.30(1-0.30)}{250}} = 0.029.$$

Thus, the required probability is $\Pr(0.25 < P < 0.35)$

$$\begin{aligned} &= \Pr\left(\frac{0.25-\pi}{\sigma_p} < \frac{P-\pi}{\sigma_p} < \frac{0.35-\pi}{\sigma_p}\right) \\ &= \Pr\left(\frac{0.25-0.30}{0.029} < Z < \frac{0.35-0.30}{0.029}\right) = \Pr(-1.72 < Z < 1.72) \\ &= P[Z < 1.72] - P[Z < -1.72] = \Phi(1.72) - \Phi(-1.72) \\ &= 0.9573 - 0.0427 = 0.9146. \end{aligned}$$

Example 15.9.17. It has been found that 43% of business graduates believe that a course in business ethics is very important for imparting ethical values to students. Find the probability that more than one-half of a random sample of 80 business graduates believe this fact.

Solution. We are given $\pi = 0.43$, $n = 80$.

The standard error of sample proportion P is calculated as

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.43(1-0.43)}{80}} = 0.055.$$

$$\text{Thus, we have to find, } \Pr(P > 0.50) = P\left(\frac{P - \pi}{\sigma_p} > \frac{0.50 - \pi}{\sigma_p}\right)$$

$$= \left(Z > \frac{0.50 - 0.43}{0.055} \right) = P(Z > 1.27) = 0.1020.$$

That means, the probability of having one-half of the sample believing in the value of business ethics courses is approximately 0.1

15.10. Concept of Estimation

Statistical inference is a branch of statistics which is concerned with uncertainty in decision making. The concept of statistical inference basically relates the sample characteristics to the population characteristics. The ultimate interest of statistical inference lies drawing conclusions about the population on the basis of available information in a sample. For example, the telephone company may be interested in estimating the average length of an international call in order to make decisions regarding the extent of cable to be put or satellites to be launched. In this case, it is not possible to take the entire population of international callers under study and analysis. Every one makes estimates in daily life. A businessman estimates the profit of the day in advance at the early morning in order to take decision about the investment of next day; an insurance agent estimates the income to be incurred from the premium of clients at the beginning of month, etc. An office goer takes decision on whether to wait, walk or run depending on estimation of time at hand. Managers use estimates because in all but the most trivial decisions, they must make rational decisions without complete information and with a great deal of uncertainty about what to do in future. For example, a credit manager attempts to estimate the creditworthiness of prospective customers from a sample of their past payment habits, an industrial entrepreneur attempts to estimate the future course of interest rates by observing the current behavior of those rates, etc. Hence, estimation plays a vital role in proper decision making in business.

15.10.1. Estimator and estimate. Any sample statistic that is used to estimate a population parameter is called an estimator. For example, the sample mean \bar{X} can be used as an estimator of the population mean μ , and the sample proportion P can be used as an estimator of population proportion π .

Definition. Estimator. An estimator is a sample statistic used to estimate a population parameter.

Definition. Estimate. An estimate is a specific observed numerical value of a statistic used to estimate a parameter.

Again, when a specific numerical value is observed for an estimator, it is called an estimate. For example, suppose mean sales of a shop for a month is observed from a sample of a week's sales as TK. 20000, if this specific value of mean sales is used to estimate the mean sales of the shop, then it would be an estimate.

15.10.2. Types of Estimation. It is possible to make two types of estimations of a population parameter; these are (i) point estimation and (ii) interval estimation.

Point Estimation. In case of point estimation the parameter of a population is estimated at a particular point. For example, observing the first quarter's credit of a bank, the manager can say that our average credit of certain year will reach to TK. 30 million; again, observing the early first week's sales of a retail shop, a shopkeeper can say that the average daily sales of this month will be TK. 20,000, etc.

Definition. Point estimate. A point estimate is a single number that is used to estimate an unknown population parameter.

However, it is also possible that such an estimate may be made by an assistant instead of shopkeeper himself. Assistant might say that the average daily sales of this month will be TK. 22000. In that case the question arises regarding which one is better estimate of average sales. The better estimator is that which possesses maximum of the desirable criteria of a good estimator. These criteria are described below in brief.

Desirable properties of a good estimator

The desirable criteria of a good estimator are unbiasedness, consistency, efficiency and sufficiency. A brief description of the criteria is provided below.

- Unbiasedness:** We know, the value of the sample average would depend on the values of the sample and may differ from sample to sample. An estimator is said to be unbiased if its expected value is equal to the parameter. For example, a sample mean \bar{X} is said to be an unbiased estimator of population mean μ if $E(\bar{X}) = \mu$ (since sampling distribution is a probability distribution, average is referred to the expected value).
- Consistency:** Consistency refers to the effect of sample size on the accuracy of the estimator. A statistic is said to be consistent estimator of the population parameter, if it approaches the parameter as the sample

size increases. Thus, sample mean \bar{X} is said to be a consistent estimator of population mean μ if $\bar{X} \rightarrow \mu$ as $n \rightarrow \infty$.

- c) **Efficiency:** An estimator is considered to be efficient, if its value remains stable from sample to sample. Hence, the best estimator is the one which would have the least variance from sample to sample taken randomly from the same population. For example, sample mean is the best estimator of population mean than median or mode because variance of sample mean is less than that of median or mode.
- d) **Sufficiency:** An estimator is said to be sufficient if it uses all the information about the population parameter contained in the sample. For example, the statistic sample mean uses all the sample values for its computation while mode and the median do not, hence, sample mean is a better estimator in this sense.

However, a point estimator is sometimes insufficient, because it may or may not be closed to respective parameter. If the shopkeeper's point estimate is wrong, it does not indicate any information about the extent of its error or deviation from true value, although this deviation can be reduced by increasing the sample size. Again, point estimate does not specify the confidence regarding its closeness to the parameter. However, this amount of error and confidence can be accomplished by second type of estimation, called, interval estimation.

Confidence Interval Estimation.

A confidence interval estimator for a population parameter is a rule for determining a range or an interval based on sample information that is likely to include the parameter. This type of estimation indicates the error in two ways, by the extent of its range, and by the probability of the true population parameter lying within that range. In this case, the manager would use some interval for estimation of his average credit, such as between TK. 25 million to TK. 35 million, and can say that it is very likely (with a confidence of 90%, say) that the exact credit will fall within this interval, or the shopkeeper would say that it is very likely (with a confidence of 95%, say) that the daily sales amount would be between TK. 19500 and TK. 20500.

The confidence interval is a probabilistic statement where the probability that the estimated values of the parameter for repeated samples would lie between an interval is known or the percent of samples that the sample mean will lie in the computed interval is known. This probability or percent is called the confidence level or confidence co-efficient. For example, in case of 95% confidence interval the value 95% or 0.95 is known as level of confidence while $100(1 - 0.95)\% = 5\%$ is known as level of significance,

similarly, in case of 99% confidence interval the value 0.99 is known as level of confidence and $100(1 - 0.99)\% = 1\%$ is known as the level of significance. A brief concept of this level of significance, usually denoted by α , is discussed in the next chapter. The smaller value of the interval is called lower confidence limit and the larger value is called upper confidence limit.

Definition. Interval estimation. An interval estimate is a range of values used to estimate a population parameter.

Definition. Confidence Level or co-efficient. The percent of the samples the true mean will lie in the specified interval estimate.

Definition. Confidence Interval. An interval estimate of the parameter with a specific probability or level of confidence.

However, the choice of the level of confidence to use in a particular situation will depend on the problem involved but a value of 95% is commonly used if no other value is specified. For example, suppose 50 samples of size 100 were taken from a population and 95% confidence intervals for μ were calculated for each sample. This would give 50 different confidence intervals, each based on one of the 50 different values of sample mean. The problem for the statistician is that he or she never knows whether the confidence interval just computed is one that would contain the parameter or not. However, it can be said that in 95% cases his or her interval would be right. Confidence interval may be interpreted in different ways. Three types of possible interpretations of confidence interval for population mean are illustrated below.

Case I. Confidence interval for population mean when a sample is drawn from a normal population with known variance σ^2 for all possible sample size

For any given confidence coefficient $(1 - \alpha)$, confidence interval for population mean μ is given by $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. The expression for confidence

interval is sometimes written as $\bar{x} \pm ME$, where, ME, the margin of error, is given by $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

For example, 95% confidence interval for population mean μ is given by $\bar{x} \pm z_{0.025} \frac{\sigma}{\sqrt{n}}$ that means, $X_1 = \bar{x} - 1.96 se(\bar{x}) = \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}$ (note that here $\alpha = 0.05$), and $X_2 = \bar{x} + 1.96 se(\bar{x}) = \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$.

Case 2. Confidence interval for population mean when a sample is drawn from any population with unknown variance when the sample size is large.

In this case 95% confidence interval for μ is

$$X_1 = \bar{x} - 1.96 \frac{s}{\sqrt{n}} \text{ and } X_2 = \bar{x} + 1.96 \frac{s}{\sqrt{n}}.$$

Here population variance σ^2 is estimated by $s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$. Here, X_1

and X_2 are known as lower and upper confidence limits respectively, and the figure 95% or 0.95 is called the confidence coefficient..

Note that $P(Z > 1.96) = 0.025$ that means $z_{0.025} = 1.96$, and $P(Z < -1.96) = 0.025$ that means $-z_{0.025} = -1.96$ are respectively used as the lower and upper critical values of Z for 95% confidence interval. In this way if a 90% confidence interval is to be constructed, ± 1.645 are to be used as upper and lower critical values respectively, because, $P(Z > 1.645) = 0.05$ and $P(Z < -1.645) = 0.05$.

This confidence interval for μ may be interpreted in any of the following ways.

- i) If all the samples of size n are taken from the population, then on an average 95% of the sample would indicate the population mean would lie between X_1 and X_2
- ii) If a random sample of size n is taken from a given population, then the interval (X_1, X_2) contain the population mean with confidence coefficient 0.95.
- iii) If a random sample of size n is taken from a given population, we can be 95% confident in our assertion that the population mean would lie around the sample mean in the interval bounded by the values X_1 and X_2 .

Width of confidence interval

Sometimes we are interested in the width of a confidence interval, which is simply defined as the difference between the upper confidence limit and the lower confidence limit. There are three factors that affect the width, namely, the value of standard deviation, the size of sample n and the level of confidence required.

Definition. Width of a confidence interval. Width of a confidence interval is difference between the upper confidence limit and the lower confidence limit.

A higher level of confidence (say 99%) will give a greater width than a lower level of confidence (say 90%) and it is suggested to take the advantage of high confidence against greater width during calculation of confidence interval.

Thus, the width for $100(1 - \alpha)\%$ CI for population mean is given by

$$\left\{ \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} - \left\{ \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Case III. Confidence interval for population mean (for small sample) :

Again, if the sample size is small and population variance is unknown, then $100(1 - \alpha)\%$ confidence interval for population mean is given by

$$\bar{x} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \text{ or } \bar{x} \pm ME$$

$$\text{where, } ME = t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \text{ and } s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

Case IV. Confidence interval for population proportion (for large sample) :

Let P denotes the observed proportion of 'successes' in a particular sample of n observations from a population with a proportion of successes π . Then, if n is large enough, say, $n\pi(1 - \pi) > 9$, a $100(1 - \alpha)\%$ confidence interval for the population proportion is given by

$$P - z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}} < \pi < P + z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$$

Or, $P \pm ME$, where ME , the margin of error, is given by

$$ME = z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$$

$$\text{Hence, lower limit is given by } P_1 = P - z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$$

$$\text{and upper limit is given by } P_2 = P + z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}.$$

Examples 15.10.1. The sponsor of a television program targeted at the children's market wants to find out the average period of time children spend watching television. A random sample of 100 children indicated the average time spent by them per week to be 27.2 hours. From the previous experience, the population standard deviation of the weekly extent of watching television is known to be 8 hours. Calculate 95% confidence interval for population mean watching time if it is considered to be adequate for taking decision about average period of watching television.

Solution. The distribution of period of watching television is not known, but the sample is quite large, hence, by the virtue of central limit theorem, the sample mean will be approximately normally distributed with mean μ and standard error $\frac{8}{\sqrt{100}} = 0.8$, that means $\bar{x} \sim N(\mu, 0.8^2)$

We know the 95% confidence interval for population average is given by $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$, here $\bar{x} = 27.2$, $\sigma = 8$, $n = 100$, so the confidence limits are

$$X_1 = \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} = 27.2 - 1.96 \frac{8}{\sqrt{100}} = 25.63 \text{ and}$$

$$X_2 = \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} = 27.2 + 1.96 \frac{8}{\sqrt{100}} = 28.77$$

Hence, it can be concluded with 95% confidence that on an average a child spends between 25.63 hours and 28.77 hours per week after watching television.

Again, if we compute 90% confidence interval for average watching period, we would have to use ± 1.645 instead of ± 1.96 , because $P(Z > 1.645) = P(Z < -1.645) = 0.05$. Thus, 90% confidence interval for population mean watching time would be (25.88, 28.52).

Here, the confidence width for 95% confidence interval is $(28.77 - 25.63) = 3.14$ hours, while for 90% confidence interval it is $(28.52 - 25.88) = 2.64$. So the width is higher for higher confidence level.

Example 15.10.2. Gasoline prices rose dramatically during the recent months of this year. A study was conducted with 24 truck, the fuel consumption, in miles per gallon, for these 24 trucks were

15.5	21.0	18.5	19.3	19.7	16.9	20.2	14.5
16.5	19.2	18.7	18.2	18.0	17.5	18.5	20.5
18.6	19.1	19.8	18.0	19.8	18.2	20.3	21.8

- i) Estimate the population mean fuel consumption for the trucks with 90% confidence.
- ii) Later it has been found that the actual population mean consumption is 20 miles per gallon. What conclusion might the statistician draw about his sample?

Solution. (i) Assuming the data has been obtained from a normal population with mean μ and variance σ^2 . Since the sample size is small and population variance σ^2 is unknown, the 90% confidence interval is given by

$$\bar{x} \pm t_{n-1, 0.1} \frac{s}{\sqrt{n}}$$

Here, $\bar{x} = 18.68$, $s = \sqrt{\frac{1}{n-1} \sum (x - \bar{x})^2} = 1.69526$, $t_{n-1; \alpha/2} = t_{23, 0.05} = 1.714$,

Thus, 90% confidence interval for μ is

$$\begin{aligned} \bar{x} &\pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \\ &= 18.68 \pm (1.714) \times (0.3460) = 18.68 \pm 0.5930 = (18.087, 19.273) \end{aligned}$$

(ii) Since the actual mean consumption is more than the estimated upper confidence limit, it can be concluded that the sample was not representative, somehow the less gasoline consumed trucks were selected as sample.

Example 15.10.3. A student of a private university who wants to buy a new model car for his personal use. He has decided that he will buy a Toyota of 2010. He randomly selected 100 sales advertisements from the local newspapers over a period of six months and found that average car price in this sample was \$4500 (expressed in dollar to make the figures smaller). He also knows that the standard deviation of such new model car prices is \$520. Establish a 95% confidence interval for the true average price for all new model cars in this category.

Solution. As we know the 95% confidence interval for the population mean is given by the points X_1 and X_2 defined as $X_1 = \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}$ and $X_2 =$

$$\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}.$$

Here, $n = 100$, $\bar{x} = 4500$, $\sigma = 520$, thus the values of X_1 and X_2 become

$$X_1 = 4500 - 1.96 \frac{520}{\sqrt{100}} = 4398.08 \text{ and } X_2 = 4500 + 1.96 \frac{520}{\sqrt{100}} = 4601.92$$

[Note that if it is required to calculate 90% confidence interval for population mean, then the change would be made only in critical values of z , thus, the values of X_1 and X_2 would be $X_1 = 4500 - 1.645 \frac{520}{\sqrt{100}} = 4414.46$

$$\text{and } X_2 = 4500 + 1.645 \frac{520}{\sqrt{100}} = 4585.54]$$

Example 15.10.4. The breaking strains of reels of string produced at a factory have a standard deviation of 1.5 kg. A sample of 100 reels from a certain batch were tested and mean breaking strain was found as 5.30 kg.

- Find a 95% confidence interval for the mean breaking strain of string.

- b. The manufacturer becomes concerned if the lower 95% confidence limit falls below 5 kg. A sample of 80 reels from another batch gave a mean breaking strain of 5.31 kg, will the manufacturer be concerned?

Solution. The distribution of breaking strain is not known, but the sample is quite large, hence, by the virtue of central limit theorem, the sample mean will be approximately normally distributed with mean μ and standard error

$$\frac{1.5}{\sqrt{100}} = 0.15, \text{ that means } \bar{x} \sim N(\mu, 0.15^2).$$

Here, $\bar{x} = 5.30$ and for 95% confidence interval, the value of z is 1.96, so.

- a) 95% confidence interval for population mean breaking strain is given by

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 5.30 \pm 1.96 \times 0.15 = (5.006, 5.594).$$

- b) If mean breaking strain for another sample of size is found as 5.31, the lower 95% confidence limit would be

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} = 5.31 - 1.96 \frac{1.5}{80} = 4.98$$

hence the manufacturer would be concerned, because, this estimated limit is beyond the CI obtained for first batch.

Example 15.10.5. A survey was made on the opinion of mobile phone users of a particular operator whether the service provided by them is sufficient to satisfy a customer. 1200 users were randomly selected and 780 of them answered in affirmative. Construct a 99% confidence interval for the corresponding true proportion of affirmative answers regarding such opinion of all users.

Solution. We know, the 99% confidence interval for population proportion is given by

$$P \pm z_{0.005} \sqrt{\frac{P(1-P)}{n}}$$

Here, the sample proportion of users who are in favor of satisfactory service is $n = 1200$, so, $P = 780/1200 = 0.65$, and $z_{0.005} = z_{0.0025} = 2.58$ (from standard normal table)

Substituting the values of n , P and $z_{0.005}$, we have,

$$P_1 = 0.65 - 2.58 \sqrt{\frac{0.65(1-0.65)}{1200}} = 0.614 \text{ and } P_2 = 0.65 + 2.58 \sqrt{\frac{0.65(1-0.65)}{1200}} = 0.686$$

Hence the 99% confidence interval for π is given by (0.614, 0.686).

Example 15.10.6. Management wants to estimate the proportion of the corporation's employees who favour a modified bonus plan. From a random sample of 344 employees it was found that 261 were in favour of this particular plan. Find a 90% confidence interval estimate of the true population proportion that favours this modified plan.

Solution. Suppose π denotes the true population proportion and p denotes the sample proportion, then $100(1 - \alpha)\%$ confidence interval for the population proportion is given by $P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$.

Here, $n = 344$, so, $p = 261/344 = 0.759$ and $z_{0.05} = 1.645$. Therefore, the required confidence interval is given by

$$P_1 = 0.759 - 1.645 \sqrt{\frac{0.759(1-0.759)}{344}} = 0.721 \text{ and}$$

$$P_2 = 0.75 + 1.645 \sqrt{\frac{0.759(1-0.759)}{344}} = 0.797$$

15.11. Determination of Sample Size

The need for determination of sample size is very important for any survey in business or management where either the standard error is known on the basis of past experience or where a given absolute level of accuracy is desired. The appropriate sample size depends on the various factors related to the subject under investigation, such as, time aspect, cost aspect, degree of accuracy, etc. However, if the sample size is too large, more money and time will be spent and even then the results may not be adequate. Also a valid conclusion may not be arrived if the sample size is too small, therefore, need of proper size of sample is significant. The following two considerations are usually kept in mind in determining the appropriate sample size.

- i) The size of the sample should increase as the variations in the individual items increases.
- ii) The greater the degree of accuracy desired, the larger should be the sample size.

15.11.1. Sample size for given width of confidence interval. If width of a estimated confidence interval for a for a given sample size is seemed to be larger for any decision making, then it is possible to determine the minimum sample size to be required to achieve a given minimum width of

confidence interval for a given confidence level which might think to be more realistic.

For example, we know, the width of $100(1 - \alpha)\%$ confidence interval for population mean is given by $2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Thus, if it is desired that the width should not exceed a given value, say, k , then, the required minimum sample size may be computed using this restriction as

$$2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq k, \text{ or, } n > \left[2z_{\alpha/2} \frac{\sigma}{k} \right]^2$$

15.11.2. Sample size for estimating the population mean. We know the $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}} = \bar{x} \pm E,$$

where E is the margin of error (the difference between the sample mean and the population mean, to be desired by the investigator).

So, $E = z_{\alpha/2} \sigma_{\bar{x}}$, which can also be written as

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ or } \sqrt{n} = z_{\alpha/2} \sigma / E \text{ which gives, } n = [z_{\alpha/2} \sigma / E]^2$$

Where z is to be set by the level of confidence interval and σ , the value of the population SD may be actual or estimated from the past experience, or estimated by using s which is from either a previous sample or a pilot survey.

5.11.3. Sample size for estimating the population proportion. As in case of mean, we have, $100(1 - \alpha)\%$ confidence interval for population proportion is given by

$$p \pm z_{\alpha/2} \sigma_p = p \pm E \text{ where } E = z_{\alpha/2} \sigma_p.$$

Where $E = P - \pi$ the difference between the sample proportion p and the population proportion P .

$$\text{Now, } E = Z \sigma_p = z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}} \text{ Or. } n = [z_{\alpha/2}^2 P(1-P)/E^2]$$

Here the value of parameter P may be taken from the past experience, a previous sample or a pilot survey.

Example. 15.11.1. What should be size of the sample from a set of 2000 accounts if the standard deviation of default as per past experience was 2.6

when a 95% confidence is desired and sample mean should not differ by more than half from the population mean.

Solution. The desired confidence interval is $\bar{x} \pm 0.5$ thus $E = 0.5$.

At 95% confidence interval $\alpha = 0.05$, so, $z_{0.025} = 1.96$, and $\sigma = 2.6$ (given), Substituting these values in the formula we get,

$$n = [z_{0.025} \sigma / E]^2 = [1.96 \times 2.6 / 0.5]^2 = 103.86 \text{ or } 104.$$

Example 15.11.2. The length of metal rods produced by an industrial process are normally distributed with a standard deviation of 1.8 mm. Based on a random sample of nine observations from this population, the 99% confidence interval for mean has been found as $194.65 < \mu < 197.75$. Now suppose that a production manager believes that the interval is too wide for practical use and instead requires 99% confidence interval extending no further than 0.50 mm on each side of the sample mean. How large a sample is needed to achieve such an interval?

Solution. Here ME is given as 0.50, $\sigma = 1.8$ and since $\alpha = 1\%$ or 0.01, $z_{\alpha/2} = z_{0.005} = 2.576$, so the required sample size is computed as

$$n = [z_{0.005} \sigma / E]^2 = \left[\frac{2.576 \times 1.8}{0.5} \right]^2 = 86$$

Therefore to satisfy the manager's requirement, a sample of at least 86 observations is needed.

Example: 15.11.3. The sales manager of a large manufacturing company wants to check the inventory records against the physical inventories by a sampling study. He indicates that i) the maximum sampling error should not be more than 10% above or below the true proportion of inaccurate records, ii) the level of confidence interval is 95%, and iii) the proportion of the inaccurate records is estimated as 25% according to the past experience.

We find the sample size as follows:

The confidence interval is $P \pm 10\%$, thus, $E = 0.10$.

At 95% confidence interval $Z_{\alpha/2} = Z_{0.025} = 1.96$.

The estimated population proportion is $P = 25\% = 0.25$ and $Q = 1 - P = 0.75$

Substituting the values in formula we get,

$$n = (1.96)^2 \times (0.25) \times (0.75) / (0.10)^2 = 72.03 = 72.$$

Example 15.11.4. Suppose the university authority wishes to know if graduate admission personnel who viewed scores on standardized exams as very

important. In a sample of 142 observations, 78 answered 'very important'. Suppose, instead it must be ensured that a 95% confidence interval for the population proportion extends no further than 0.06 on each side of sample proportion. How large a sample be needed for this purpose?

Solution. It is given that $P = 78/142 = 0.55$, $E = 0.06$ and $z_{\alpha/2} = 1.96$, thus the number of sample observations needed is

$$n = (z_{\alpha/2})^2 \frac{P(1-P)}{E^2} = (1.96)^2 \times \frac{0.55(1-0.55)}{(0.06)^2} = 266.78 \text{ or } 267$$

Example 15.11.5. A random sample of 120 electrical bulbs are tested and the mean duration are found as 78.7 hours. The standard deviation of duration of bulbs is 11.5 hours. (i) Find a 95% CI for μ , (ii) The authority thinks that the width of the interval is too large to take decision, so determine the sample size to conform that the width of this interval should not exceed 2.5 hour.

Solution: (i) The distribution of duration of electric bulbs is not known, but the sample is quite large, hence, by the virtue of central limit theorem, the sample mean will be approximately normally distributed with mean μ and

standard error $\frac{\sigma}{\sqrt{n}} = \frac{11.5}{\sqrt{120}} = 1.05$, that means $\bar{x} \sim N(\mu, 1.05^2)$.

We know the 95% confidence interval for population average is given by

$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$, here $\bar{x} = 87.8$, $\sigma = 11.5$, $n = 120$, so the confidence limits are

$$X_1 = \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} = 87.8 - 1.96 \times 1.05 = 76.64, \text{ and}$$

$$X_2 = \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} = 87.8 + 1.96 \times 1.05 = 80.76$$

Hence, the 95% confidence interval for true mean is (76.64, 80.76), that means, the width of this interval is $(80.76 - 76.64) = 4.12$ hours.

(ii) We know, the width of 95% confidence interval estimation of population mean is given by

$$2 \times 1.96 \times \frac{\sigma}{\sqrt{n}}, \text{ and for the given problem, it is } 2 \times 1.96 \times \frac{11.5}{\sqrt{n}}.$$

We have to calculate n so that $2 \times 1.96 \times \frac{11.5}{\sqrt{n}} \leq 2.5$, or, $n > 325.15$

That, for achieving the width of interval estimation 2.5 hours, the sample size should be at least 326.

Example 15.11.6. A random sample of 25 medium size mangoes is taken from a big stockiest shop. Suppose the weights of packets are normally distributed with standard deviation 2.5 kg. The sample mean was found as 17.8 kg..

- Find 99% confidence interval for the population mean weight μ
- What sample size is required to obtain a 99% confidence interval of width at most 1.5 kg?

Solution. Here, $\bar{x} = 17.8$, $n = 25$, $\sigma = 2.5$, so the standard error of sample mean is given by $\frac{2.5}{\sqrt{25}} = 0.50$, that means $\bar{x} \sim N(\mu, 0.5^2)$. Again, for 99% confidence interval, the value of z is $z_{0.005} = 2.5758$, so.

- (a) 99% confidence interval for μ is given by

$$\bar{x} \pm 2.5758 \frac{\sigma}{\sqrt{n}} = 17.8 \pm 2.5758 \times 0.50 = (16.51, 19.09).$$

- (b) Width of 99% CI for μ is given by

$$2 \times 2.5758 \times \frac{\sigma}{\sqrt{n}} = 2 \times 2.5758 \times \frac{2.5}{\sqrt{n}} = \frac{12.879}{\sqrt{n}}$$

According to the question, we have to find n so that

$$\frac{12.879}{\sqrt{n}} \leq 1.5, \text{ or } n > 73.719, \text{ hence, } n = 74.$$

Thus, the required sample size is 74.

Questions

- Define census and sample survey. What are the differences between them?
- Explain the purpose and necessity of sampling or explain the need of sampling as compared to complete enumeration.
- Briefly discuss some of the reasons why a sample is chosen instead of testing the entire population.
- Define what is meant by a census. By referring to specific examples, state the reasons when census is unavoidable. Also state when the census is not suggested.
- State the importance of sampling in business. Explain the principles on which sampling is to be undertaken.
- Define sampling units and sampling frame with example. Also discuss in which situation sampling is inevitable.

7. Discuss the various methods of sampling along with the requisite of a good sample.
8. What is sampling? Discuss the well-known methods of probability sampling and non-probability sampling.
9. What are the techniques of sampling? Explain simple random sampling technique and discuss the method of drawing a simple random sample using random number table.
10. What do you understand by simple random sampling? Discuss how can you draw simple random sample of size 10 from a population of size 100.
11. Explain the purpose of stratification in carrying out a sample survey. Explain the concept of proportional allocation in context to stratified random sampling.
12. Distinguish between probability sampling and non-probability sampling with examples. Also explain why systematic sampling is called a mixed probability sampling?
13. When is stratified random sampling method preferred? Explain the procedure of drawing a stratified random sample using proportional allocation. Also discuss the advantage of this sampling method over simple random sampling method.
14. How does a stratified random sampling differ from simple random sampling? What is the advantage of stratified sampling over simple random sampling?
15. Explain the procedure of conducting quota sampling. How does it differ from judgment sampling or convenient sampling?
16. What do you mean by sampling distribution? How does it differ from a probability distribution?
17. State the names of some important sampling distributions known to you. For a sample of size n from normal population with mean μ and variance σ^2 , find the distribution of sample mean.
18. Explain what is meant by the term 'sampling' and point out the advantages of taking sample over census.
19. Explain i) Why a sample might be preferred to a census? ii) What do you understand by a sampling frame?
20. Write brief notes on i) simple random sampling ii) stratified random sampling iii) systematic sampling iv) quota sampling v) judgment sampling vi) convenience sampling.
21. State two circumstances when you would consider using i) systematic sampling, ii) stratified sampling and iii) quota sampling.
22. State the advantages and disadvantages of systematic sampling over other sampling methods. Also state the situation when systematic sampling is preferred.

23. What do you mean by sampling error? How does it differ from non-sampling error? What are their sources and how they can be controlled?
24. Explain the concept of standard error. Discuss the role of standard error in large sample theory. How does standard error differ from standard deviation?
25. Distinguish between sampling error and non-sampling error. What are their sources and how these can be controlled?
26. What do you mean by sampling distribution? Write notes on i) Chi-square statistic, ii) t- statistic, iii) F-statistic.
27. State the properties of i) Chi-square distribution, ii) t- distribution, iii) F- distribution.
28. Stating the necessary assumptions, write down the sampling distribution of i) a sample mean ii) difference between two sample means, iii) a sample proportion
29. What do you mean by estimation. Distinguish between point estimation and interval estimation.
30. How can you compute the confidence limits for the population mean with known variance.
31. What are the criteria of a good estimator? How does interval estimation differ from point estimation?
32. How can you determine the sample size for in case of estimating i) population mean ii) population proportion?
33. How can you determine the required sample size to estimate population mean for a given level of confidence and width of confidence interval?

Exercise

34. Suppose a box contains 30% 50 paisa coins and the rest 100 paisa coins. Of two coins are selected randomly from the box, find the sampling distribution of sample mean of coins. Also find the standard error of mean.
35. Draw a random sample three numbers from the first five natural numbers, and show that the sample mean is an unbiased estimate of population mean.
36. A factory makes safety harnesses for climbers and has an order to supply 1000 harnesses. The buyer wishes to know at what load the harness will break. Suggest a reason why a census would not be used for this purpose.
[Ans. Harnesses might be destroyed during testing process.]
37. A company wishes to do consumer market research using a certain city. Suggest a suitable sampling frame and describe in detail a way of selecting a sample of 400 people aged over 18.

38. A survey is to be done on adult population of a certain small village, the population of which is, say, 5000. An ordered list of villagers is available. A sample 50 people are required for a particular study. Suggest a sampling method with justification.
[Ans: Systematic sampling could be used since the population size is large enough for a simple random sampling, assuming that ordered list is purely random.]
39. In a marketing survey the sales of cigarettes in a variety of stores are to be investigated. The stores consist of small temporary tea-shop selling tea and tobacco, shops that sell cigarettes and other related products, and shops that sell cigarettes and other stationary products. Suggest the most suitable method of selecting sample for this purpose. Also explain how would you conduct the sample survey.
40. A random sample of size 25 is obtained from a normal population with mean 100 and variance 81,
- Write down the sampling distribution of sample mean.
 - Find the probability that sample mean is more than 102.
 - Find the probability that the sample mean lies between 98 and 107.
 - Find 99% confidence interval for population mean.
 - Find the required sample size to obtain a width of confidence interval no more than 3.
41. A random sample of size 60 is obtained from a normal population with mean 80 and variance 36,
- Write down the sampling distribution of sample mean
 - Find the probability that sample mean is more than 81.5
 - Find the probability that the sample mean lies between 78 and 82
 - Find 90% confidence interval for population mean
 - Find the required sample size to obtain a width of confidence interval no more than 2.
42. A random sample of size 60 is obtained from a normal population with mean 80 and variance 36,
- Write down the sampling distribution of sample mean
 - Find the probability that sample mean is more than 81.5
 - Find the probability that the sample mean lies between 78 and 82
 - Find 90% confidence interval for population mean
 - Find the required sample size to obtain a width of confidence interval no more than 2.
43. A sample of size 9 is taken from a normal distribution with variance 36, the sample mean is 128,
- Find a 95% CI for population mean μ of the distribution and interpret.
 - Find a 99% CI for population mean μ of the distribution and interpret.

[Ans. i) 124.08, 131.92 ii) 122.85, 133.15]

44. A sample of size 25 is taken from a normal distribution with standard deviation 4, the sample mean is 85, i) find a 90% CI for population mean μ of the distribution and interpret ii) find a 99% CI for population mean μ of the distribution and interpret. ii) Also compute the size of samples in both cases required to obtain a width of maximum 1.5.
 [Ans. i) 83.68, 86.32 ii) 83.43, 86.57]
45. A normal distribution has standard deviation 15. Estimate the sample size required if the following confidence intervals for the mean should have width less than 2 i) 90% ii) 95% and iii) 99%
 [Ans. i) 609 ii) 865 iii) 1493]
46. Suppose that we have a population with proportion $P = 0.40$, and a random sample of size $n = 100$ drawn from the population
 i) What is the probability that the sample proportion is greater than 0.45?
 ii) What is the probability that the sample proportion is less than 0.29?
 iii) What is the probability that the sample proportion is between 0.35 and 0.51?
 iv) Find 95% confidence interval for population proportion and hence compute the width of confidence interval.
47. Clearly state the meaning of the statement $P(1.35 \leq \mu \leq 15.67) = 0.99$

Applications

48. Suppose that the shopping times for customers at a local store are normally distributed. A random sample of 16 shoppers in the local grocery store had a mean time of 25 minutes. Assume $\sigma = 6$ minutes and the shopping time is normally distributed, find the standard error of mean, margin of error and width for a 95% confidence interval for the population mean μ .
 [Ans, SE = 1.5, ME = 2.94, width = 5.88]
49. A process produces bags of refined sugar. The weights of the contents of these bags are normally distributed with SD 1.2 ounces. The contents of 25 bags had a mean weight of 19.8 ounces. Find the 99% confidence interval for the true mean weight for all bags of sugar produced by the process.
 [Ans. 19.18, 20.42]
50. The director of an electronic company is interested to estimate the mean expenditure of customers on electrical appliances. A random sample of 80 customers was questioned and found the average expenditure as Tk 47 thousand with a standard deviation of 16.5 thousand. Find a 95% confidence interval for true mean expenditure. Suppose the director is not satisfied with the confidence interval, because it is too high, so he wishes to know how large a sample would be required to obtain a 95% confidence interval of total width no more

than 4 thousand. Find the smallest size of sample that will satisfy this desire. [Ans. 43.38, 50.62, 262]

51. The lifetime of bulbs produced by a particular company have a mean of 1200 hours and a standard deviation of 400 hours. The population distribution is normal. Suppose that you purchase nine bulbs which can be regarded as a random sample from the manufacturer's output,
- What is the distribution of sample mean?
 - What is the probability that on average those nine bulbs have average life of less than 1050 hours?
 - Find 99% confidence interval for population mean
 - Suppose, the authority is not satisfied with the interval so obtained, it wishes to have an interval estimation so that the width is no more than 3 hours. Estimate the sample size required.
52. An overnight delivery service claims that on an average 95 percent of all mail is delivered before noon the following day. Random sample of 400 deliveries are selected. What proportion of the sample will have
- Between 95 percent and 97 percent of deliveries before noon the next day.
 - More than 92% of the deliveries before noon the next day.
 - Determine 90% confidence interval for population proportion.
 - Also determine the sample size required to obtain a width of no more than 0.05.
53. A local bank has 2000 depositors with 40 percent of these depositors having current as well as savings accounts. The rest have only the current accounts. A random sample of 400 such accounts has been selected.
- What is the probability that the sample proportion of depositors with both accounts will be between 0.40 and 0.43?
 - What is the probability that the sample proportion of depositors with both accounts will be more than 0.38?
 - Compute a 99% confidence interval for population proportion.
 - Also determine the sample size required for 99% confidence interval with width less than 0.03.
54. A random sample of 16 housewives has an average body weight of 52 kgs and standard deviation of 3.6 kgs,
- Find the probability that a randomly selected housewife will have weight more than 55 kg.
 - Find the probability that a randomly selected housewife will have weight between 50 kg and 26 kg
 - Compute the standard error of mean weight and 99% confidence interval for population mean weight.
55. Batteries of manufacturer A have a mean lifetime of 1400 hours with standard deviation of 200 hours, while those of manufacturer B have a

mean lifetime of 1200 hours with a standard deviation of 100 hours. If a random sample of 125 batteries of each manufacturer are tested, what is the probability that manufacturer A's batteries will have a mean lifetime which is at least i) 160 hours more than manufacturer B's batteries and ii) 250 hours more than the manufacturer B's batteries?

[Ans. i) 0.9772, ii) 0.0062]

56. Suppose the following information have been gathered from two samples drawn from two independent populations of hourly number of same products of two machines;

	Machine A	Machine B
Sample size	28	35
Average hourly product	62	74
Variance	144	225

Compute the 95% confidence interval for difference of population average products of two machines and comment.

57. The strength of wire produced by company A has a mean of 3500 kg and a standard deviation of 150 kg. Company B has a mean of 3000 kg and a standard deviation of 200 kg. If 50 wires of company A and 100 wires of company B are selected at random and tested for strength, what is the probability that the sample mean strength of A will be at most 600 kg less than that of B?
58. Two methods of performing a certain task in a manufacturing plant, method A and method B, are under study. The variable of interest is the length of time required to perform the task. It is known that the variance of method A is 9 minutes squared and variance of method B is 12 minutes squared. A simple random sample of 35 employees performed the task A and independent random sample of 35 employees performed task B. The average time required by the first group to complete the task was 25 minutes and the average time for the second group was 23 minutes. What is the probability that the difference between the sample means is at least 3 hours?
59. A random sample of 250 homes was taken from a large population of older homes to estimate the proportion of homes with unsafe wiring. If 30% of the homes have unsafe wiring,
- What is the probability that the sample proportion will provide between 25% and 30% of homes with unsafe wiring?
 - Compute 95% confidence interval for population proportion.
 - Estimate the sample size required to obtain width of interval maximum.

[Ans. i) 0.9146 ii) 0.24, 0.36 iii) 505]

CHAPTER - 16

TESTS OF HYPOTHESIS

16.1. Introduction

In most of the situations, it is very difficult to study the whole population. The value of population parameter is usually unknown, and one objective of sampling is to estimate its value. The choice of appropriate statistics depends on which population parameter is of interest to statistician. Any inference drawn about the population is based on sample statistic. Now the question arises, whether the sample statistic is a representative value of respective population parameter or whether there is any significant difference between the parameter and statistic to some extent. This matter can be ensured by the test of hypothesis. For example, a reputed toothpaste producer claims that average weight of its big size toothpaste is 140 gm. The consumers association of Bangladesh (CAB) can verify this claim by the following steps:

- i) Collecting a random sample of toothpastes
- ii) Determining the average weight of toothpastes and
- iii) Performing a test of hypothesis about the mean weight.

Definition. Hypothesis Testing. The process that enables a decision maker to draw an inference about population characteristics by analyzing the difference between the value obtained from sample and the hypothesized value of parameter is called hypothesis testing.

The tests of hypothesis involve all of the above-mentioned three steps. The second step is simply the determination of sample statistic corresponding to the parameter, which is of interest to the researchers.

16.2. Concepts of Hypothesis Testing

The process of hypothesis testing can be compared with criminal jury trial. In a jury trial it is assumed that the criminal is innocent, and the jury will decide that a person is guilty only if there is very strong evidence against the presumption of innocence. This criminal jury trial process for choosing between guilt and innocence possesses:

- i. Rigorous procedures for presenting and evaluating evidence
- ii. A judge to enforce the rules
- iii. A decision process that assumes innocence unless there is evidence to prove guilt beyond a reasonable doubt.

It is to be noted here that this process will fail to convict a number of people who are, in fact, guilty. But if a person's innocence is rejected and the person is found guilty, we have a strong belief that the person is guilty. In testing the hypothesis, the sample collected from population acts as evidence of trial.

However, the following concepts are related to this type of decision.

16.2.1. Hypothesis and Statistical hypothesis. Every moment, we use different types of statements regarding different events of our life. Statistics is not concerned with all types of statements. Statistics usually deals with statements, which have some relation with uncertainty. Thus, any statement about any aspect of a phenomenon is considered as hypothesis. Tomorrow will be a sunny day, s/he is the president of some association, the firm is not running well, etc. are examples of the statements about tomorrow's weather, chief of an association, state of the firm, respectively. Hence these are simply hypothesis. However, in attempting to take decision regarding some characteristics of population on the basis of sample, it is necessary to make some assumption regarding parameters of the population, such assumptions which may or may not be true, are called statistical hypothesis. Hence, statistical hypothesis is a claim or statement (belief or assumption) about unknown feature (distribution or parameter) of a population. For example, average monthly sale of a store is Taka 10,000, average hourly production of a machine is 200 units, proportion of defective products produced by a certain machine is less than other and so on.

Hypothesis. Any statement about any phenomenon is termed as hypothesis.

Statistical Hypothesis. Statistical hypothesis is a statement about population characteristic that can be tested on the basis of sample data.

A pious person will go to the heaven is not a statistical hypothesis since it cannot be proved with the statistical data. It does not mean that it has no value. It is a faith.

In statistical tests of significance, two mutually exclusive hypotheses are to be used; these are null hypothesis and alternative hypothesis.

16.2.2. Null hypothesis. The approach of statistical hypothesis testing starts with a statement complement to the original claim. The hypothesis about the parameter of a population such as mean μ , the variance σ^2 , or the proportion π which is formulated for sole purpose of rejecting or nullifying it, is called null hypothesis. Hence, null hypothesis is a statement about no difference between the parameter and statistic. Null hypothesis is denoted by H_0 . For example, if we want to decide whether a given coin is not fair, we formulate the null hypothesis that the coin is fair, i.e. $H_0: \pi = 0.5$.

Null Hypothesis. According to R A Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. Null hypothesis is denoted by H_0

16.2.3. Alternative hypothesis. The alternative hypothesis is a logical opposite statement of null hypothesis. If null hypothesis is rejected (or actually it is false), then some alternative form of parameter should be true. Thus, any hypothesis that differs from a given null hypothesis is called an alternative hypothesis. Alternative hypothesis is denoted by H_A or H_1 . For example, if the null hypothesis about population mean is $H_0 : \mu_0 = 55$, an alternative might be $H_A : \mu_1 \neq 55$, or $\mu_1 < 55$ or $\mu_1 > 55$.

Alternative Hypothesis. The hypothesis, which is true if the null hypothesis is false is called alternative hypothesis. Alternative hypothesis indicates the type of test (left, right, or two-tail). It is denoted by H_1 or H_A

Once the test has been carried out, the final conclusion is always given in terms of the null hypothesis. We either "Reject H_0 in favour of H_1 " or "Do not reject H_0 "; we never conclude "Reject H_1 ", or even "Accept H_1 ". However, if we conclude, "Do not reject H_0 ", this does not necessarily mean that the null hypothesis is true, it only suggests that there is no sufficient evidence against H_0 or in favour of H_1 . Rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Clear and precise formulation of the null and alternative hypothesis is the first and foremost step in the test of significance. Without formulation of this hypothesis one cannot proceed to perform any test of significance. Once a null hypothesis is formulated, it is required to formulate an alternative to this that is to be considered if there is not enough evidence to accept null hypothesis. Hence, role of null and alternative hypothesis in a test of significance lies in taking decision against or in favor of the postulated presumption regarding population parameter. The decision will be taken in favor of that hypothesis which will be more evident from the test for a given level of significance. Thus the null and alternative hypothesis in a test of significance enables us to take decision whether the postulated presumption is justified or not. Moreover, all hypothesis testing are done under the assumption that the null hypothesis is true. The decision in a test of significance is based on null hypothesis and the type of test (left, right, or two-tail) is based on the alternative hypothesis.

16.2.4. Simple hypothesis. A hypothesis is said to be a simple hypothesis if it completely specifies the distribution of the population from which the sample has been considered. In this case, the information about all the parameters of population distribution is known. For example, if a coin is tossed 50 times (that means n of binomial distribution is 50) to determine if

the coin is fair one, the null hypothesis to be formulated is $H_0: \pi = 0.50$, which is a simple hypothesis because it specifies the population distribution completely. Again for testing $H_0: \mu = \mu_0$ of a normal distribution if the population variance σ^2 is known, it is a simple hypothesis.

Simple Hypothesis. The hypothesis, which completely specifies all the parameters of the related population, is called simple hypothesis.

16.2.5. Composite hypothesis. On the other hand, if a hypothesis does not specify the population distribution completely, it is called a composite hypothesis. In the above coin tossing example, if $n = 50$ is not specified, $H_0: \pi = 0.50$ would be a composite hypothesis, because, there is a number of distributions all with $\pi = 0.50$. In this case, n is called a nuisance parameter. Similarly, even if we know $n = 50$, but the hypothesis to be tested is defined as $H_0: \pi \neq 0.50$ or $H_0: \pi > 0.50$, it would be composite hypothesis, because value of π is not specified by a single value. Again for testing $H_0: \mu = \mu_0$ of a normal distribution if the information about σ^2 is not given, it is a composite hypothesis. In this case, the parameter σ^2 is known as nuisance parameter.

Composite hypothesis. The hypothesis, which does not completely specify the parameters, is called a composite hypothesis.

16.2.6. Errors in decision-making. In any decision-making, if the decision is not correct, the decision maker may commit error in two mutually exclusive ways, termed as type I error and type II error.

Table 16.1. Errors in decision-making

Decision	State of Nature	
	H_0 is True	H_0 is False
Reject H_0	Type I Error	Correct decision
Fail to reject H_0	Correct decision	Type II Error

Type I error. The error of rejecting the null hypothesis when it is in fact true is called type I error. That means, type I error occurs when null hypothesis is wrongly rejected. A type I error is also known as first kind of error.

Type II error. The error of accepting the null hypothesis when it is false is called type II error. Type II error occurs when null hypothesis is not rejected wrongly. A type II error is also known as second kind of error.

A type I error is often considered to be more serious, and therefore more important to avoid than a type II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0. While the exact probability of a type II error is generally unknown.

If we do not reject the null hypothesis, it may still be false (a type II error) as the sample may not be representative enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis).

For any given set of data, type I and type II errors are inversely related; the smaller the risk of one, the higher the risk of the other.

Although it is desirable to keep the both types of errors at minimum level, but unfortunately in practice it is not possible. Hence the probability of type I error is kept fixed (to be considered at the beginning of testing procedure) and then try to minimize the probability of type II error.

16.2.7. Level of significance. In testing a given hypothesis, the maximum probability with which we would be willing to take risk of rejecting a hypothesis when it should be accepted, is called the level of significance of the test. This probability is denoted by α , generally specified before any sample is drawn so that the results obtained will not influence the choice of the decision maker. Since *Type I is the more serious error* (usually) that is the one we concentrate on. We usually pick to be very small such as 0.05, 0.01 or in some cases 0.001. It is to be noted here that alpha (α) is not a Type I error, alpha is the *probability of committing a Type I error* and beta (β) is the *probability of committing a Type II error*.

Level of significance: The probability of committing a type I error is called the level of significance. In other words, it is the total area under critical region. Symbolically, $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$. It is also known as size of a test.

$1 - \alpha = P(\text{accept } H_0 \mid H_0 \text{ is true})$ is called the confidence coefficient.

Power of a test: The complement of the probability of type II error is called the power of a test. That means, the probability of rejecting a false null hypothesis is the power of a test. In other words, the probability of correct decision is power of a test. Symbolically, Power of a test = $1 - \beta = 1 - P(\text{accept } H_0 \mid H_0 \text{ is false}) = P(\text{reject } H_0 \mid H_0 \text{ is false})$.

Interpretation of level of significance. Generally a significance level of 0.05 or 0.10 is considered, although other values are also used. Thus if the level of significance is 0.05, it will mean that there are about 5 samples out of 100 that would direct to reject the hypothesis when it should be actually accepted. So $(1 - 0.05) = 0.95$ is the probability of accepting null hypothesis when it is true, i.e. there is 95% confidence in taking the right decision. In such case it is said that the hypothesis has been rejected at 5% level of significance, which again means that the probability of wrong decision is 0.05.

16.2.8. One tailed and two tailed test. A one-tailed test is a test, which is concerned about possible deviation of the value of the parameter in only one direction from the specified value defined in the null hypothesis, while a two-tailed test is a test, which is concerned about the possible deviation of the parametric value in both directions. These are also called a one-sided alternative or two-sided alternative. In this case, the parameter can take any value other than the value specified by null hypothesis. For example, $H_0 : \mu = 0$, against $H_1 : \mu \neq 0$ is a two tailed test, while, $H_0 : \mu = 0$, against $H_1 : \mu > 0$, or $H_0 : \mu = 0$, against $H_1 : \mu < 0$ is a one tailed tests.

A left – tailed test. When the rejection region is in the left tail of the distribution of the test statistic, the test is called a left-tailed test. If the null hypothesis is $H_0 : \mu_0 = 0$, then the alternative hypothesis will be $H_1 : \mu < 0$.

A right – tailed test. When the rejection region is in the right tail of the distribution of the test statistic, the test is called a right-tailed test. If the null hypothesis is $H_0 : \mu_0 = 0$, then the alternative hypothesis will be $H_1 : \mu_0 > 0$.

Alternative hypotheses defined in the above two tests are called one –sided alternatives.

A Two- tailed test: When the rejection region is equally divided in the left and right tails of the distribution of the test statistic, the test is called a two-tailed test. The alternative hypothesis defined in this test is called two – sided alternatives. If the null hypothesis is $H_0 : \mu_0 = 0$, then the two – sided alternative hypothesis is defined by $H_1 : H_1 : \mu \neq 0$.

In the test of significance, the decision whether the use of a one-tailed or a two-tailed test is appropriate depends on the objective of the test, i.e. it is chosen on the basis of the direction of claimed deviation of the parameter.

Suppose that a manufacturer of ballpoint pen, whose machine produces on average 1000 pens per hour, is planning to purchase a new machine. The authority will not buy a new machine unless it is definitely proved as superior. For this he would test the hypothesis that the new machine is no better than the existing machine or similar to the machine now available in the market against the alternative that the new machine is superior with respect to the hourly average production of existing machine. In other word, it is required to test the null hypothesis $H_0 : \mu = 1000$ against the alternative $H_1 : \mu > 1000$ and buy the new machine only if the null hypothesis is rejected. Such an, alternative test will result in a one tailed test with the critical region in the right tail.

Similarly, consider a drug, on an average five doses of which is enough in order to get cured a certain disease. A company introduces a new drug and claims that in average only two doses of this new drug will need to get cure

of the same disease. In order to verify the companies claim one has to formulate $H_0: \mu = 5$ against the alternative $H_1: \mu < 5$. The company's claim will be true if the null hypothesis is rejected. Such an alternative test will result in a one tailed test with the critical region in the left tail.

Again, suppose one wishes to test the inequality of income of two populations. In this case, the deviation of the income of one population to other may happen in any of the two sides, i.e. income of one population may be either more or less than other population. In such case, it is required to consider a two-tailed test.

16.2.9. Test Statistic. The decision about test of a hypothesis or the acceptance or rejection of a hypothesis is based on the statistical evidence from sample data. In testing procedure, it is desirable to select an appropriate statistic to be computed from sample data depending on the assumptions or available information or nature of population from which the sample has been drawn.

Test statistic. The statistic, which is used to provide evidence about the rejection or acceptance of null hypothesis, is called test statistic. The decision about the rejection or acceptance of a null hypothesis is taken comparing the observed (calculated) and theoretical (tabulated) value of test statistic.

16.2.10. Critical Region and Acceptance region. Sample space of experiment, which corresponds to the area under the sampling distribution curve of the test statistic, is divided into two mutually exclusive regions, such as acceptance region and rejection or critical region. If the value of the test statistic falls within the rejection region, the null hypothesis is rejected, and if the value of the test statistic falls within the acceptance region, we fail to reject it. In case of one-tailed test the critical region, the area of which is exactly equal to the level of significance, lies entirely on one extreme end of the curve depending on whether the test is right or left tailed, while in case of two-tailed test area of rejection is divided into two regions lie at both ends of the curve, area of each of these regions is usually exactly half of level of significance.

Definition. Critical region. The set of possible values of the test statistic, which provides evidence to contradict with null hypothesis and lead to the rejection of null hypothesis is called critical region. The set of values of the test statistic that support the alternative hypothesis and lead to rejecting the null hypothesis is called the rejection region.

Definition. Acceptance region. The set of values of the test statistic, which provides evidence to agree with the null hypothesis and lead to the acceptance of null hypothesis is called acceptance region.

Critical regions for different types of alternatives are displayed in following figures.

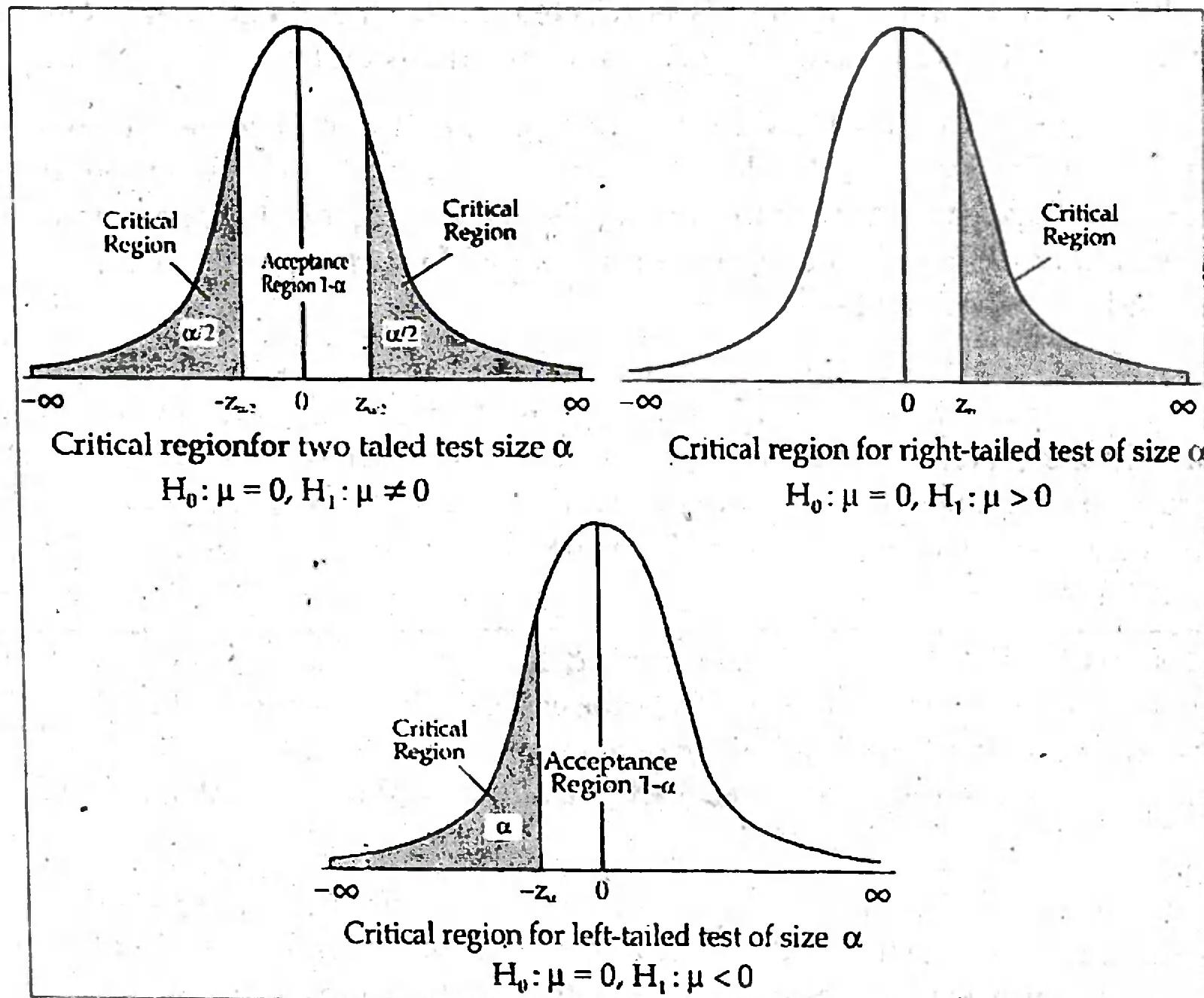


Fig. 16.1. Areas of acceptance and rejection regions for one tailed and two-tailed test.

16.2.11. Critical Value. A critical value is measured in the same units of measurement as the test statistics and identifies the value of the test statistic that would lead to the rejection or acceptance of null hypothesis at the specified level of significance. This is also called the theoretical value of statistic. The critical values are determined independently of the sample statistics from the sampling distribution under null hypothesis from the table. The critical value (s) for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is to be rejected. The critical value for any hypothesis test depends on i) the significance level at which the test is carried out, and ii) the type of test (one-sided or two-sided). There are two critical values for a two-tailed test, while one for a one-tailed test. The critical values of popular test statistics for different level of significance are available in Statistical tables.

Definition. Critical value. The value of the sample statistic that separates acceptance region and rejection region is called critical value. This is the value of the test statistic with which observed value is compared and decision regarding acceptance or rejection of null hypothesis is taken.

Some useful critical values of Z for both one-tailed and two-tailed tests at various levels of significance (negative value of Z stands for left-tailed test and positive value for right-tailed test) are presented in following table.

Table 16.2. Critical values of Z statistic for different level of significance.

Level of significance α	0.10	0.05	0.025	0.01	0.005	0.002
Critical values of Z for right-tailed test	1.28	1.645	1.96	2.33	2.58	2.88
Critical values of Z for left-tailed test	-1.28	-1.645	-1.96	-2.33	-2.58	-2.88
Critical values of Z for two-tailed test	± 1.645	± 1.96	± 2.33	± 2.58	± 2.81	± 3.08

16.2.12. p- value. The decision to reject or accept the null hypothesis is taken by comparing the observed value of test statistic with the critical value, which is obtained on the basis of level of significance. Since the critical value of test statistic is different at different levels of significance, so the decision about the acceptance or rejection may be different at different levels of significance. For example, suppose for a right-tailed test the value of Z-statistic observed from the sample is 2.03 and at the 0.05 level of significance the critical value of Z is 1.645, here 2.03 lies in the critical region, so the null hypothesis may be rejected at 5% level of significance. However, we cannot reject H_0 at the 0.01 level because the test statistic is less than the critical value $z=2.33$. That means the null hypothesis may be rejected or may not be rejected at varied level of significance. To avoid this type of confusion in decision making about rejection of a null hypothesis it is desirable to consider the smallest level of significance at which it is rejected. This smallest level of significance is termed as p-value. By doing this, the actual risk of committing Type I error can be established. We see from the table of standard normal distribution that the critical value of Z at 2.12% level of significance is 2.03, which is the smallest value at which the hypothesis may be rejected, so here the p value is 0.0212.

p-value. The p-value or observed significance level of a statistical test is the smallest value of the level of significance at which the null hypothesis can be rejected. It is the actual risk of committing a type I error, if H_0 is rejected based on the observed value of the test statistic. The p-value measures the strength of the evidence against H_0 .

A small p-value indicates that the observed value of the test statistic lies far away from the hypothesized value. This presents strong evidence that H_0 is false and should be rejected. A large p-value indicates that the observed test statistic is not far from the hypothesized value and does not support rejection of H_0 .

If the p-value is less than a pre-assigned significance level α , then the null hypothesis can be rejected, and we can report that the results are statistically significant at level α .

Many researchers classify p-values as follows:

- i) If the p-value is less than .01, H_0 is rejected. The results are highly significant.
- ii) If the p-value is between .01 and .05, H_0 is rejected. The results are statistically significant.
- iii) If the p-value is between .05 and .10, H_0 is usually not rejected. The results are only tending toward statistically significant.
- iv) If the p-value is greater than .10, H_0 is not rejected. The results are not statistically significant.

In particular, for testing the null hypothesis $H_0 : \mu = \mu_0$, against $H_1 : \mu > \mu_0$, p-value is given by

$$\text{p-value} = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq z_{\text{obs}} \mid H_0 : \mu = \mu_0\right)$$

where, z_{obs} is the observed value of the test statistic associated with the smallest significance level at which the null hypothesis can be rejected.

It is to be noted here that all hypothesis testing are done under the assumption that the null hypothesis is true. It is also important to understand that the rejection of null hypothesis is to conclude that it is false, while fail to reject it does not necessarily mean that it is true. We fail to reject null hypothesis since we have no sufficient evidence to believe otherwise.

16.2.13. Assumption. The suppositions regarding population and/or sample, which are needed to take decision about the distribution of test statistic, are known as assumption. For example, the test statistic Z follows $N(0, 1)$ under the assumption that the sample observations are drawn independently from normal population with known variance or the sample size is large. On the other hand, the test statistic t follows Student's t distribution under the assumption that the sample observations are independently drawn from a normal population with unknown variance and the sample size is small.

16.3. Survey of Important Test Statistics

The important test statistics are

- i) Z-test or normal test
- ii) t-test
- iii) χ^2 -test and
- iv) F-test.

A brief survey of the important parametric test statistics is provided below. Applications of these test statistics have been described in section 16.5.

16.3.1. Z-test or Normal test. In a normal tests we find U whose expected value $E(U)$ is specified by the null hypothesis. The standard error $\sigma(U)$ of U is either known or estimated from a large sample. Then a statistic

$$Z = \frac{U - E(U)}{\sigma(U)}$$

is taken as a normal variable with mean 0 and standard deviation 1. It is symbolically expressed as $Z \sim N(0,1)$. If the distribution of U is normal and $\sigma(U)$ is known, then Z is exactly normally distributed. Frequently, however, the distribution of U is approximately normal or $\sigma(U)$ is estimated from a sample or perhaps both. When samples are large this approximation is usually quite satisfactory. That is why; normal test is often regarded as a large sample test.

In particular, when U is sample mean \bar{X} then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

The normal curve is symmetrical about mean; hence, the critical value at a particular level of significance at right side is the same as that of left side with a negative sign. Normal tests can be used in case of one-tailed as well as two-tailed tests.

16.3.2. t- tests (Student's t-test). In case of normal tests it is assumed that population variance is either known or estimated from large sample (usually $n > 30$), but very often we have to deal with small samples where population variance is unknown. In this situation the test statistic t is to be used instead of Z-statistic. The distribution of t contains a parameter v (nu) known as degrees of freedom (d.f.). This is a positive integer and always less than n , the size of the sample. The relationship between v and n depends on how $\sigma(U)$ calculated. The normal test can be regarded as a special case of t-test when v is large. So, a t-test is also called a small sample test.

This statistic t is generally known as Student's t is defined in similar algebraic form of Z-statistic except that standard error is estimated from small sample. Thus the algebraic form of t is

$$t = \frac{U - E(U)}{\text{Estimated } \sigma(U)} \text{ with } v = n - 1$$

In particular, for sample mean \bar{X} , $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

$$\text{where } s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 \text{ with } v = n - 1.$$

Like the normal distribution, t distribution is also symmetrical about mean; hence, the critical value at a particular level of significance with certain degrees of freedom at right side is the same as that of left side with a negative sign. Like the normal test, the t -tests can also be one-tailed or two-tailed.

Another form of t -test is also used for testing the difference between two means in case of dependent or correlated or repeated samples (x, y). This is known as paired t -test, defined as

$$t = \frac{d - \bar{d}}{s(d)} \text{ with } v = n - 1, \text{ where, } d = x - y$$

16.3.3. χ^2 -tests. χ^2 -tests are used mainly for testing hypothesis that specify the nature of one or more distributions as a whole. Thus a hypothesis may define the mathematical form of a distribution or assert the two or more distributions are identical or two attributes are independent, etc. The elements common to the test statistics used for testing the above hypothesis is that each involves the comparison of an observed set of frequencies with a corresponding set of expected set of frequencies under null hypothesis. If O_i ($i = 1, 2, \dots, k$) denotes the observed frequency, and E_i denotes the corresponding expected frequency, then the test statistic χ^2 is defined as

$$\chi^2_{(v)} = \sum \frac{(O_i - E_i)^2}{E_i} \text{ where } v \text{ denotes the degrees of freedom, the}$$

only parameter of the theoretical χ^2 distribution.

χ^2 is used for testing varying types of hypotheses and the value or definition of v varies with type of hypothesis under consideration. For example,

- i) χ^2 - test for specific variance here, $v = n - 1$
- ii) χ^2 - test for goodness of fit, here, $v = k - r$
- iii) χ^2 - test for independence of attributes, here $v = ((r - 1)(c - 1))$.

Here n is the sample size, k is the number of cells or values of a variable, r is the number of restrictions imposed on the set of frequencies while calculating the expected frequencies, r is the number of rows and c is the number of columns. It is to be mentioned here that the first one is the parametric test while the next two are non-parametric. Tests of goodness of fit are beyond the scope of this book.

χ^2 is used for testing the hypothesis that normal population from which the sample is available has a specified variance. This is an exact and one can consider a one tailed as well as a two-tailed test. In this case χ^2 is defined as

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \text{ with } n-1 \text{ degrees of freedom}$$

However, for last two cases, χ^2 tests are approximate in the sense that the test statistic has an approximate χ^2 distribution under null hypothesis and these are one-tailed tests.

16.3.4. F-tests. R.A. Fisher originally devised this test and Snedecor called it F-test in honour of Fisher. Suppose s_1^2 and s_2^2 denote the sample variances of the same variance σ^2 of a normal population computed from two independent samples of sizes n_1 and n_2 respectively with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Then the statistic F is defined as

$$F = \frac{s_1^2}{s_2^2} \text{ with } v_1 \text{ and } v_2 \text{ degrees of freedom, where, } s_1^2 > s_2^2$$

The statistic F is so defined that s_1^2 in the numerator is expected to be larger than denominator s_2^2 when the null hypothesis is not true. Hence, only the upper tail of F-distribution is usually used as the critical region.

16.4. Steps in Hypothesis Testing

As it is clear from above discussion that in order to test the validity of claim or assumption about the population parameter, at first it is necessary to draw a sample from the respective population and then analyzed. Some assumptions regarding the population distribution may also be necessary for suitability of tests. The results of the analysis are used to decide whether the claim is true or not. Thus, the general procedure for hypothesis testing consists of following basic steps:

- i) **State the null hypothesis and alternative hypothesis:** Clear and precise formulation of the null and alternative hypothesis is the first and foremost step in the test of significance. Without formulation of this hypothesis, one cannot proceed to perform any test of significance. That

means, at first it is required to state the assumed value of population parameter which is to be tested as null hypothesis. A justified alternative hypothesis along with null hypothesis is also to be established depending on the statement of problem. Care should be taken regarding the alternative whether it would be one-tailed or two-tailed. For example, suppose we want to test the hypothesis that the monthly average price of a commodity is Taka 100 per kg. In this case, the following null and alternative hypothesis are to be considered $H_0: \mu = 100$, against the alternative, $H_1: \mu \neq 100$.

- ii) **Specify the level of significance (α) prior to sampling:** At the second step of test of hypothesis, it is required to specify the level of significance. Because, the risk of taking wrong decision depends on the nature of study, so maximum risk should be determined before drawing sample from the population. It is at the discretion of investigator to select its value. Although usually $\alpha = 0.05$ is considered, but value of α may vary depending on the sensitivity of the study. For example, 5% risk might be more for taking decision about the effectiveness of a drug, in that case, 1% or 0.1% level of significance may be considered.
- iii) **Select the suitable test statistic:** Depending on the formulated hypothesis, assumption made about the population distribution, sample size, at this stage the appropriate test statistic is to be selected.
- iv) **Establish the critical region:** At this stage the critical region is to be established on the basis of above steps. That means, the critical region is selected depending on alternative hypothesis whether it is one-tailed or two-tailed, the level of significance and the selected test statistic. For example, for the two tailed alternative $H_1: \mu \neq 100$ at 5% level of significance, suppose normal test statistic Z is selected for testing the hypothesis, then the critical region is the lower and upper 2.5% area of a standard normal distribution which are $Z < -1.96$ and $Z > 1.96$, hence the critical values are ± 1.96 (the values of z at certain level of significance can be obtained from the table 'Area under the normal curve' available in statistical tables or in the appendix of almost all books on statistics).
- v) **Collect sample and Compute the value of the test statistic:** After selecting the critical region, a sample of predetermined size n is collected. Then the value of selected test statistic is calculated from sample. In this case, it is assumed that the null hypothesis is true.
- vi) **Compare observed and critical values:** The value of the test statistic computed in earlier step is compared with the critical value or values. The computed value is checked whether it falls within or beyond the critical region.
- vii) **Make the decision:** The decision about the acceptance or rejection of hypothesis is taken on the basis of critical value. It is either "reject the

null hypothesis" or "fail to reject the null hypothesis or not reject the null hypothesis". If the observed value of test statistic lies within the critical region, then "reject the null hypothesis", on the other hand if the observed value of test statistic falls beyond the critical region, then decision is taken as 'fail to reject null hypothesis'.

For the convenience of the users, a set of decision rules for normal test for $\alpha = 1\%$, 5% and 10% are provided below:

Table 16.3. Decision rule for one-tailed and two-tailed test using Z-statistic

Alternative Hypothesis	Decision Rule		
	$\alpha = 0.01$ Reject H_0 if	$\alpha = 0.05$ Reject H_0 if	$\alpha = 0.10$ Reject H_0 if
$\mu \neq \mu_0$	$Z > 2.58$ or $Z < -2.58$	$Z > 1.96$ or $Z < -1.96$	$Z > 1.645$ or $Z < -1.645$
$\mu > \mu_0$	$Z > 2.53$	$Z > 1.645$	$Z > 1.28$
$\mu < \mu_0$	$Z < -2.53$	$Z < -1.645$	$Z < -1.28$

viii) **Draw conclusion:** This is a statement, which indicates the level of evidence (sufficient or insufficient) at given level of significance, and/or, whether the original claim is rejected or accepted. If decision is taken in favour of alternative hypothesis, then it is concluded that there is sufficient evidence to reject null hypothesis or accept the original claim.

16.5. Applications of Test Statistics

In this and the following sections we will discuss some specific applications of the test statistics viz. Z , t , χ^2 in testing various types of hypotheses related to business and management.

The applications are classified as follows:

Applications of Z-statistic

- i) Test of a single population mean
- ii) Test of equality of two population means
- iii) Test of a single population proportion
- iv) Test for difference between two population proportions
- v) Test of a specified correlation co-efficient
- vi) Test of equality of two-population correlation co-efficient

Applications of t-statistic (small sample test)

- i) Test of a single population mean
- ii) Test of difference between two population means
- iii) Test of significance of a correlation co-efficient with zero value
- iv) Test of a population of regression co-efficient with zero or specified value.
- v) Test of difference between two population regression co-efficient

Applications of χ^2 -statistic

- i) Test of a population variance with specific value.
- ii) Test of equality of several variances
- iii) Test of equality of several correlation co-efficient
- iv) Test of equality of several population proportions
- v) Tests of independence of attributes
- vi) Test of goodness of fit

Applications of F -statistic

- i) Test of significance of difference between two population variances
- ii) Test of significance of several population means
- iii) Test of significance of two or more regression co-efficient

The above-mentioned applications of the test statistics are discussed below.

16.6. Hypothesis Testing for Single Population Mean

Although it is difficult to draw a clear-cut line of demarcation between large and small samples, it is generally agreed that if the size of sample exceeds 29, then it may be regarded as a large sample. The test of significance used for large samples are different from that of small samples for the reasons that the assumptions we make in case of large samples do not hold for small samples. The following assumptions are to be made for Z-test.

For small sample: The sample is randomly selected from a normally distributed population with known variance, and

For large sample: The sample is randomly selected from a normally distributed population with unknown variance.

Some practical examples are cited below.

1. Population normal and variance known for any sample size (small and large)

Suppose X_1, X_2, \dots, X_n be a random sample of size n drawn independently from a normal population with mean μ and variance σ^2 . In this case,

$$X \sim N(\mu, \sigma^2), \text{ then } \bar{X} \sim N(\mu, \sigma^2/n).$$

The following null and alternative hypothesizes may be considered for testing the population mean:

- i) $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ (for a two tailed alternative)
- ii) $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ (for a right tailed alternative)
- iii) $H_0: \mu = \mu_0$ against $H_1: \mu < \mu_0$ (for a left tailed alternative)

The test statistic for testing the null hypothesis H_0 for all the alternatives is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Decision rule. (i) Suppose the level of significance is α . We find $z_{\alpha/2}$ from the standard normal integral table by using $P(|Z| > z_{\alpha/2}) = \alpha/2$. We reject the null hypothesis if the absolute observed value of z is greater than $z_{\alpha/2}$.

(ii) For the second case, we find the z_α from the standard normal integral table by using $P(Z > z_\alpha) = \alpha$. We reject the null hypothesis if the observed z value is greater than z_α .

(iii) For the third case, we find the $-z_\alpha$ from the standard normal integral table by using $P(Z < -z_\alpha) = \alpha$. We reject the null hypothesis if the observed z value is less than $-z_\alpha$.

The above three cases can be shown in the following table.

Table 16.4. Decision rule for a single mean test using Z statistic

Case No.	Type of test	Decision rule
		Reject H_0 , if
1	Two-tailed test $H_1: \mu \neq \mu_0$	$ Z > z_{\alpha/2}$
2	Right-tailed test $H_1: \mu > \mu_0$	$Z > z_\alpha$
3	Left-tailed test $H_1: \mu < \mu_0$	$Z < -z_\alpha$

Example 16.6.1. The managing director of a firm claims that his firm produces 110 items on average daily. A random sample of 15 days gives the following data set:

110, 118, 130, 140, 142, 146, 112, 100, 95, 98, 96, 122, 123, 124, 130.

It is known that the number of items produced by the firm follows normal distribution with variance 300.

Can we conclude at 5% level of significance that the average daily production of items of that firm is

- (a) 110 items
- (b) More than 110 items
- (c) Less than 110 items?
- (d) Compute p-value for each case.

Solution. (a) Steps involved in testing the hypothesis will be followed in this case:

- (i) First we have to formulate null and alternative hypothesis. It is a two-tailed test. Since if the average number of items produced by the firm is

more or less than 110 to some extent, then the claim of the managing director will be proved as false. In that case the claim would be rejected.

So, the null hypothesis and alternative hypothesis can be formulated as follows:

$$\text{Null hypothesis } H_0: \mu = 110$$

$$\text{Alternative hypothesis } H_a: \mu \neq 110$$

(ii) Level of significance is $\alpha = 0.05$

(iii) Here, the sample is taken from a normal population with known variance. The appropriate test statistic is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

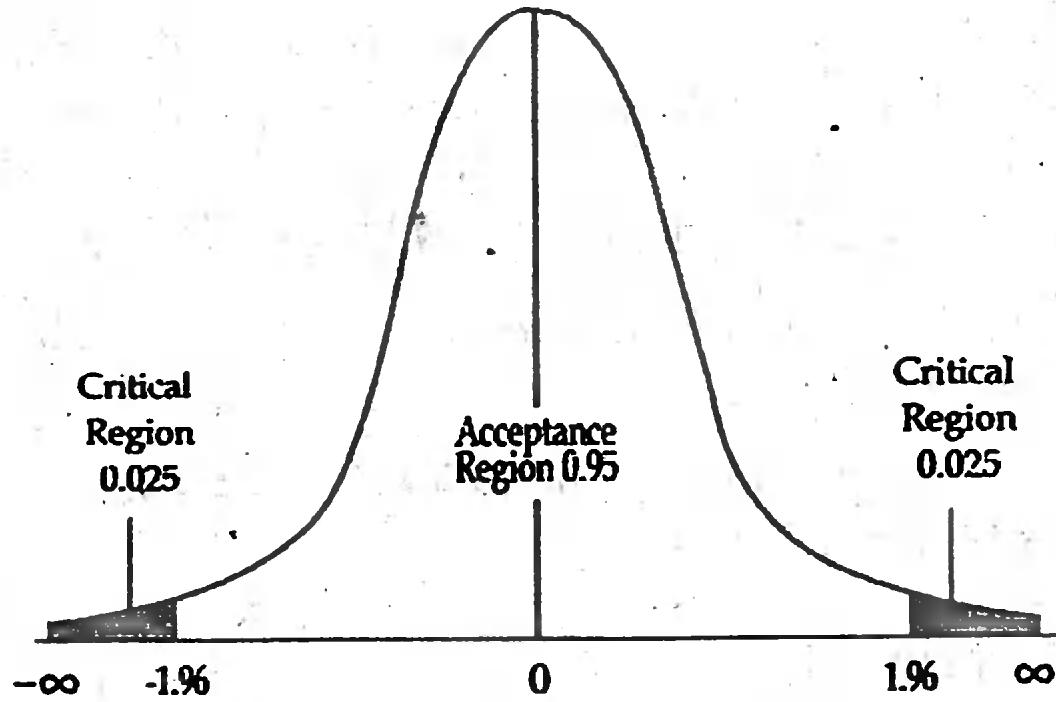
where \bar{X} = The sample mean

μ = Population mean (to be tested)

σ = Standard deviation of the population

σ/\sqrt{n} = Standard error of the sample mean \bar{X} .

(iv) Here $\alpha = 0.05$ and it is a two tailed-test, the critical region will be on both sides of curve of Z in such a way that the critical region will comprise 2.5% or 0.025 area at the right end and 2.5% at the left end. From the table of area of standard normal distribution, we see that these values of Z are ± 1.96 , that means the critical regions are $Z < -1.96$ (at the left end) and $Z > 1.96$ (at the right end), i.e. critical region is given by $|z_{0.025}| > 1.96$.



(v) Under the null hypothesis, the value of Z is

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$(vi) \text{ Here, } \bar{x} = \frac{\sum X}{n} = \frac{1786}{15} = 119.07, \mu_0 = 110$$

$\sigma^2 = 300$ and $\sigma = 17.32$ and $n = 15$. Substituting these values in the formula of Z , we have

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{119.07 - 110}{17.32 / \sqrt{15}} = 2.03.$$

- (vii) It is found that the observed value of Z is 2.03, which is greater than the right tail critical value 1.96, hence it falls in the upper critical region.
- (viii) Since the observed value of the test statistic falls in the critical region, so we fail to accept the null hypothesis at 5% level of significance.
- (ix) Conclusion. Hence we cannot accept the claim of the managing director at 5% level of significance.

p-value. From the table of the standard normal distribution we find that $P(Z > 2.03) = 0.0212$, since it is a two tailed test, the p-value is $0.0212 \times 2 = 0.0424$, that means the smallest level of significance at which the hypothesis may be rejected is approximately 4.34%

(b) Null hypothesis is the same as (a)

Null hypothesis $H_0: \mu = 110$

Alternative hypothesis $H_a: \mu > 110$

Level of significance is $\alpha = 0.05$

The appropriate test statistic is the same as (a). That is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Here $\alpha = 0.05$ and it is one-sided right tailed-test, the critical value $z_{0.05}$ will be found in such a way that $P[Z > z_{0.05}] = 0.05$. It is found from the standard normal distribution that $z_{0.05} = 1.645$.

The calculated value of Z under the null hypothesis is 2.03 which is greater than 1.645. That is observed value of Z lies in the rejection region. Hence we have no reason to accept the null hypothesis.

p-value. From the table of the standard normal distribution we find that

$$p = P(Z > 2.03) = 0.0212.$$

That means the smallest level of significance at which the hypothesis may be rejected is approximately 2.12%.

(c) Null hypothesis is the same as (a)

Null hypothesis $H_0: \mu = 110$

Alternative hypothesis $H_a: \mu < 110$

Level of significance is $\alpha = 0.05$

The appropriate test statistic is the same as (a). That is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Here $\alpha = 0.05$ and it is one-sided left tailed-test, the critical value $z_{0.05}$ will be found in such a way that $P[Z < -z_{0.05}] = 0.05$. It is found from the standard normal distribution that for left tail, $z_{0.05} = -1.645$.

The calculated value of Z under the null hypothesis is 2.03 lies outside the region $Z < -1.645$. That is observed value of Z lies in the acceptance region. Hence there is no evidence to reject the null hypothesis at 5% level of significance against the alternative hypothesis that the average production of the firm is less than 110.

p-value. From the table of the standard normal distribution we find that

$$p = P(Z > 2.03) = 0.0212.$$

This is a left tailed test, so the p-value is

$$p = P(Z > -2.03) = 1 - 0.0212 = 0.9788.$$

That means the smallest level of significance at which the hypothesis may be rejected is approximately 97.88%.

Note: In practice, if we are in a position to reject a null hypothesis, we compute p-value to find the exact level of significance. The p-value found in this case is quite unjustified.

2. Population normal, variance unknown and sample size is large ($n > 29$)

Example 16.6.2. Manager of a fertilizer factory claims that the average daily production of his factory follows normal distribution with mean production 880 kg. A random sample of 50 days shows that average production is 871 kg with standard deviation 21 kg. Test the significance of the claim of the manager at 5% level of significance. Also find p-value.

Solution. Here null and alternative hypotheses are

$$H_0: \mu = 880 \text{ and } H_1: \mu \neq 880$$

Here population is normal but variance is unknown and the sample size is large. The sample standard deviation can be taken as a good estimate of the population standard to estimate the standard error of the sample mean \bar{X} .

That is $\frac{\sigma}{\sqrt{n}}$ can be replaced by $\frac{s}{\sqrt{n}}$ for defining Z. The appropriate test

statistic is $Z = \frac{\bar{X} - \mu}{s / \sqrt{n}}$, which is approximately a standard normal variable.

The value of the test statistic Z under the null hypothesis is

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Here s^2 is the estimate of population variance σ^2 , defined as

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2.$$

It is a two-tailed test, so the critical region at 5% level of significance is

$$|Z| > 1.96$$

We have, $\bar{x} = 871$, $s = 21$, $n = 50$, so the computed value of Z under null hypothesis is

$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{871 - 880}{21 / 50} = -3.03.$$

Decision. Since the observed value of Z lies in the critical region, so we fail to accept null hypothesis. That means, the manager's claim is not justified.

p-value. From the standard normal integral table, we find that $P(Z > 3.03) = 0.0005$, and $P(Z < -3.03) = 0.0005$, so the value of p is $0.0005 \times 2 = 0.001$. Since the p-value is far less than 0.01, the value of Z is highly significant.

3. Population is not normal, variance known and sample size is large

Example 16.6.3. The producer of a company claims that the selling price of his product is very standard and it is Tk. 1500 per unit with standard deviation Tk.45. There is some doubt of CAB (Consumers' Association of Bangladesh) regarding this price. They want to verify this price using statistical testing procedure. A random sample of the sailing prices of 100 products of this company from different areas were collected. The average price per unit was found Tk.1510.

Can the CAB conclude at 5% level of significance that the average price of the product is standard? Also calculate p-value and 95% confidence interval for population mean.

Solution. We have to test the hypothesis that the sailing price of the product is TK.1500. So, the null hypothesis and alternative hypotheses are

Null hypothesis $H_0: \mu = 1500$

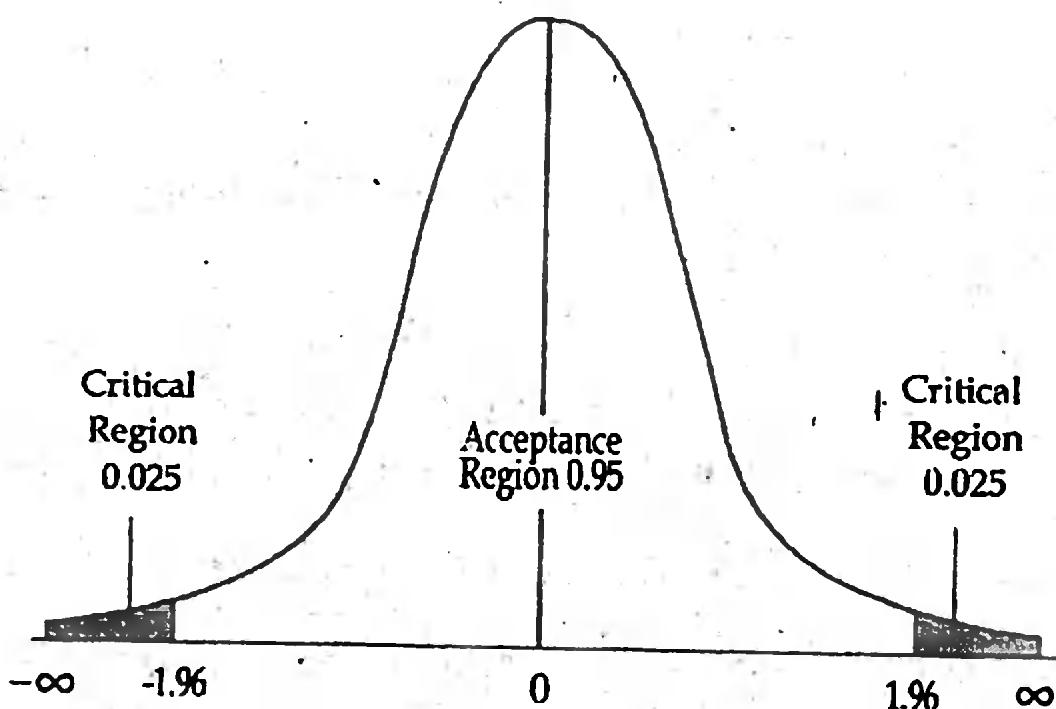
Alternative hypothesis $H_a: \mu \neq 1500$

Level of significance $\alpha = 0.05$

Here, we have to test the significance of a population mean with known population variance. But nothing is said about the form of distribution, but mean and variance exist. Since the sample size is large ($n = 100$), according to the central limit theorem, the sampling distribution of the mean is approximately normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$, so the appropriate test statistic is,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Here $\alpha = 0.05$ (given), and it is a two tailed-test, the critical region will be on both sides of curve of Z in such a way that the critical region will comprise 2.5% or 0.025 area at the right end and 2.5% at the left end. From the table of area of normal curve, we see that these values of Z are ± 1.96 , that means the critical regions are $Z < -1.96$ (at the left end) and $Z > 1.96$ (at the right end).



Under the null hypothesis Z is given by : $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.

Here, $\bar{x} = 1510$, $\mu_0 = 1500$, $\sigma = 45$ and $n = 100$, then

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1510 - 1500}{45/\sqrt{100}} = 2.22.$$

It is found that the observed value of Z is 2.22. It is greater than the critical value 1.96; hence it falls in the upper critical region. So we fail to accept the null hypothesis at 5% level of significance.

Conclusion. The claim of the producer is not right, that means, the price of the products as claimed by the producer is not standard.

p-value. From the table of area under the standard normal distribution we find that the value of $P(Z > 2.22) = 0.0132$, since it is a two tailed test, the p-value is $.0132 \times 2 = 0.0264$, that means the smallest level of significance at which the hypothesis is rejected is 2.64%.

Confidence interval: 95% confidence interval for μ is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \text{here } z_{\alpha/2} = z_{0.025} = 1.96,$$

So, the required CI is $1510 \pm 1.96 \frac{45}{\sqrt{100}} = (1501, 1519)$.

Thus, we may be 95% confident that the true mean price will be between TK. 1501 and TK. 1519.

4. Population is not normal, variance is unknown and the sample size is large

Example 16.6.4. A manufacturer of fluorescent tubes claims that his tubes have a lifetime of 1950 burning hours. A random sample of 100 tubes is taken from a day's output and tested for burning life. It is found to have a mean burning lifetime of 1900 hours with a standard deviation of 150 hours. Can the claim of the manufacturer be accepted at 5% level of significance? Also find p-value.

Solution. Here nothing is said about population and the population variance is not known.

$$H_0: \mu = 1900 \text{ and } H_1: \mu \neq 1900$$

The population from which the sample has been derived is not normal, but the sample size is large. So by the virtue of central limit theorem, we know

$$\text{that } Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1).$$

Thus the appropriate test statistic is : $Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1)$.

Here s^2 is the estimate of population variance σ^2 and is defined as

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2.$$

It is a two-tailed test, so the critical region at 5% level of significance is

$$|Z| > 1.96$$

Given, $\bar{x} = 1900$, $s = 150$, $n = 100$, so the computed value of Z under null hypothesis is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1900 - 1950}{150/100} = -3.33.$$

Decision. Since the observed value of Z lies in the critical region, so we fail to accept null hypothesis. That means, the manufacturer's claim is not accepted at 5% level of significance.

p-value. From the standard normal integral table, we find that $P(Z < -3.33) = 0.0004$, so the value of p is $0.0004 \times 2 = 0.0008$ (approx). Since the p-value is far less than 0.01, the value of z is highly significant.

5. Population normal, variance unknown and sample size is small ($n < 30$).

Suppose X_1, X_2, \dots, X_n be a random sample of size n drawn independently from a normal population with mean μ and unknown variance σ^2 .

The following null and alternative hypotheses may be considered:

- i) $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ (for a two tailed alternative)
- ii) $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ (for a right tailed alternative)
- iii) $H_0: \mu = \mu_0$ against $H_1: \mu < \mu_0$ (for a left tailed alternative)

The test statistic for testing the null hypothesis H_0 for all the alternatives is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \text{ where } s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2.$$

Here t follows Student's t-distribution with $n-1$ degrees of freedom.

The decision rules at 100α percent level of significance for hypothesis testing using t-statistic are as follows:

Table 16.5. Decision rule for t-test

Case No.	Type of test	Decision rule
		Reject H_0 , if
1	Two-tailed test $H_1: \mu \neq \mu_0$	$ t > t_{\alpha/2; (n-1)}$,
2	Right-tailed test $H_1: \mu > \mu_0$	$t > t_{\alpha; (n-1)}$,
3	Left-tailed test $H_1: \mu < \mu_0$	$t < -t_{\alpha; (n-1)}$,

Example 16.6.5. A wholesaler knows that on average sales in its store is 20% higher in December than in November. For the current year, a random sample of six stores was taken. Their percentage of sales increased in December was found to be 19.2, 18.4, 19.8, 20.2, 20.4, 19.0. Assuming that the sample has been drawn from a normal population with mean μ and unknown variance σ^2 ,

- (a) test the null hypothesis at 10% level of significance whether the true mean percentage sales increase is 20%, against the two sided alternative.
- (b) do you think that the true mean percentage sales increase is more than 20% at 10% level of significance.
- (c) do you think that the true mean percentage sales increase is less than 20% at 10% level of significance.

Solution. Here the population variance is unknown and the sample size is small. The estimated standard error of sample mean \bar{x} is given by

$$\hat{se}(\bar{x}) = s/\sqrt{n} \text{ where, } s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$$

- (a) We want to test the null hypothesis

$$H_0 : \mu = \mu_0 = 20 \text{ against the alternative } H_1 : \mu = \mu_1 \neq 20$$

Since the sample size is small and population variance is unknown, the value of the test statistic under null hypothesis is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

which is distributed as Student's t with $(n - 1)$ degrees of freedom

It is a two tailed test, so the decision rule is:

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ if } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha/2(n-1)} \text{ or } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{\alpha/2(n-1)}$$

The critical values of t at 10% level of significance with $n-1 = 5$ df are $\pm t_{n-1, \alpha/2} = \pm t_{5, 0.05} = \pm 2.015$ (From table of t-distribution).

Here, $n = 6$, $\bar{x} = 19.5$, $s^2 = 0.588$, and $s/\sqrt{n} = 0.24$.

$$\text{Then, we have, } t = \frac{19.5 - 20}{0.24} = -1.08.$$

Since the observed value $t = -1.08$ lies between -2.015 and 2.015 , hence we fail to reject the null hypothesis at 10% level of significance.

That means, the true mean sales is 20% higher in December than in November.

(b) In this case, we have to perform a one-tailed test, given by

$$H_0: \mu = \mu_0 = 20 \text{ against the alternative } H_1: \mu = \mu_1 > 20$$

The decision rule is to reject H_0 in favor of H_1 if $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha/2, n-1}$

The computed value of t is the same as before i.e., $t = -1.08$, the critical value of t at 10% level of significance is $t_{0.10, 5} = 1.476$

Since the observed value of $t = -1.08$ which is less than the critical value, we fail to reject the null hypothesis at 10% level of significance, which means the average sales increased by more than 20 percent is not evident from the given data.

(c) In this case, we have to perform a one-tailed test, given by

$$H_0: \mu = \mu_0 = 20 \text{ against the alternative } H_1: \mu = \mu_1 < 20$$

The decision rule is to reject H_0 in favor of H_1 if $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{\alpha/2, n-1}$

The computed value of t is the same as before i.e. $t = -1.08$, the critical value of t at 10% level of significance is $t_{0.10, 5} = -1.476$

Since the observed value of $t = -1.08$ lies beyond the critical region, we fail to reject the null hypothesis at 10% level of significance, which means the average sales increased by less than 20 percent is not evident from the given data.

Example 16.6.6. (Large sample with unknown variance when parent population is not normal)

A fertilizer factory manager claims that its average daily production is 912 kg. A random sample of 50 days shows that average production is 903 kg with standard deviation 21 kg. Test the significance of the claim of the manager at 5% level of significance.

Solution. Here, $H_0: \mu = 912$ and $H_1: \mu \neq 912$

The population from which the sample has been derived is not normal, but the sample size is large. So by the virtue of central limit theorem, we know

that $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1)$.

Thus under the null hypothesis, the appropriate test statistic is

$$z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

where s^2 is the estimate of population variance σ^2 , defined as

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

It is a two tailed test, so the critical region at 5% level of significance is

$$|Z| > 1.96$$

Given, $\bar{x} = 871$, $s = 21$, $n = 50$, so the computed value of Z under null hypothesis is .

$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{903 - 912}{21 / \sqrt{50}} = -3.03$$

Decision. Since the observed value of Z lies in the critical region, so we fail to accept null hypothesis. That means, the manager's claim is not justified.

Example 16.6.7. (Two-tailed test for small sample with known variance) The yields of wheat from a random sample of six test plots are as 1.40, 1.80, 1.30, 1.90, 1.60 and 2.20 tons per acre, test whether the information supports the claim that the average yield for this kind of wheat is 1.5 tons/acre with standard deviation 0.43 tons/acre. Also find the p-value.

Solution. The null and alternative hypotheses for this test are

Null hypothesis: $H_0: \mu = 1.5$

Alternative hypothesis: $H_A: \mu \neq 1.5$

Here, the sample size is small, the population variance is known and the sample is taken from normal population, so the appropriate test statistic is

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Let the level of significance $\alpha = 0.05$ (if α is not given, it is considered as 0.05).

Thus the critical values of Z are ± 1.96 , that means the critical regions are

$$Z < -1.96 \text{ and } Z > 1.96.$$

The average yield as obtained for the given observations is 1.7 tons/acre, so the value of Z under null hypothesis is

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{1.7 - 1.5}{0.43 / \sqrt{6}} = 1.14$$

Since the observed value of Z is less than 1.96, that means the observed value lies in the acceptance region, so we fail to reject null hypothesis. Hence, the information supports that the average yield is 1.5 tons/acre.

p-value. From the table of standard normal distribution, we find that $P(Z > 1.14) = 0.1271$ and $P(Z < -1.14) = 0.1271$, so, the p-value is $0.1271 \times 2 = 0.2542$, that means the test will be significance at 25.42% level of significance.

Example 16.6.8. (Right tailed test for small sample size with known variance) A stenographer claims that she can take dictation at the rate of more than 100 words per minute with a standard deviation of 15 words. Can we reject the claim on the basis of 10 trials in which she demonstrates a mean of 102 words per minute? Use 5% level of significance.

Solution. If the stenographer can take dictation of even at the rate of 100 words per minute, her claim can not be accepted. So the null hypothesis and alternative hypothesis to be considered are:

Null hypothesis: $H_0: \mu = 100$

Alternative hypothesis: $H_1: \mu > 100$

Since the population variance is known, the appropriate test statistic is

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

The critical region at 5% level of significance is $Z > 1.645$.

Here, $\bar{x} = 102$, $\sigma = 15$, $n = 10$, so the value of Z is

$$Z = \frac{102 - 100}{15 / \sqrt{10}} = 0.42$$

The computed value of Z does not exceed the critical value 1.645, so we fail to reject null hypothesis. So stenographer's claim is not correct.

Example 16.6.9. (Left tailed test for small sample size with known variance) An automobile manufacturer company claims that a new model car

achieves on an average 31.5 miles per gallon in highway driving. The distribution is known to be normal with standard deviation 2.4 miles per gallon. A random sample of sixteen automobiles provided an average of 30.6 miles per gallon in highway trials. Test the claim of company at the 5% level of significance against the population mean is less than 31.5 miles per gallon.

Solution. It is a left-tailed test, because, if the average coverage of distance is even equal to 31.5 miles, the company's claim will not be correct. Thus, the null and alternative hypotheses are:

$$\text{Null hypothesis: } H_0: \mu = 31.5$$

$$\text{Alternative hypothesis: } H_A: \mu < 31.5$$

Since, the population variance is known, the appropriate test statistic is

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

The critical region at 5% level of significance is $Z < -1.645$

Here, $\bar{x} = 30.6$, $\sigma = 2.4$, $n = 16$, so the value of Z is

$$z = \frac{30.6 - 31.5}{2.4 / \sqrt{16}} = -1.50$$

The computed value of Z does not fall in the critical region, so we fail to reject the null hypothesis, that means, the average achievement of car is not less than 31.5 miles per gallon.

Example 16.6.10. (Two-tailed test for large sample with known variance) A large manufacturer of stereo components is concerned about the efficiency of many new employees hired during the last six months. The efficiency rating of all employees has been reasonably stable with mean rating 200 and a standard deviation of 20. A random sample of 75 new employees has been selected and their average efficiency rating is found as 197.5. Test the null hypothesis at 1% level of significance that the mean efficiency rating is still 200. Find 99% confidence interval for population mean.

Solution. It is a two-tailed test because if the mean efficiency rating obtained from sample is too high or too low in comparison with population mean rating, the null hypothesis would be rejected, so the null and alternative hypotheses to be considered are

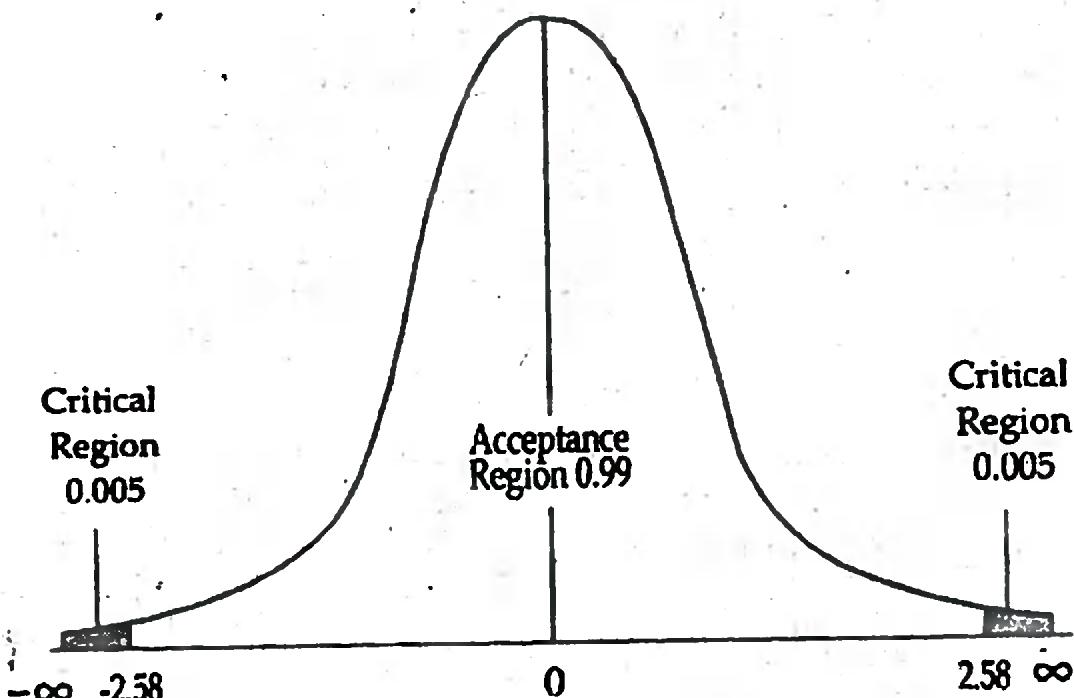
$$\text{Null hypothesis: } H_0: \mu = 200$$

$$\text{Alternative hypothesis: } H_1: \mu \neq 200$$

We have to test the significance of a population mean with known population variance $\sigma^2 = 20^2$, and the sample size is also large, so the sampling distribution of the mean is normally distributed with standard error $\frac{\sigma}{\sqrt{n}}$, the appropriate test statistic for the selected hypothesis is

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Here the level of significance is $\alpha = 0.01$ (as given in the problem).



It is a two-tailed test, the critical region will be on both ends of curve of Z that will comprise 0.5% or 0.005 area at the right end and 0.005 area at the left end. From the table of normal curve, we see that this critical values of Z are ± 2.58 , that means the critical regions are $Z < -2.58$ and $Z > 2.58$.

Now, we have to calculate $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$.

Here, $\bar{x} = 197.5$, $\mu_0 = 200$, $\sigma = 20$ and $n = 75$, we have

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{197.5 - 200}{20 / \sqrt{75}} = -1.08$$

The observed value of Z is greater than the lower critical value -2.58 . Since the observed value of test statistic does not fall within the critical region, so we fail to reject the null hypothesis at 1% level of significance. That means the mean efficiency rating is still 200.

Confidence interval: 99% confidence limits are given by $\bar{x} \pm z_{0.005} \frac{\sigma}{\sqrt{n}} = 197.5 \pm 2.58 \frac{20}{\sqrt{75}} = (191, 203)$, that means, we are 99% confident that true average efficiency rating of all employees will lie between 191 and 203.

Example 16.6.11. Suppose in Example 16.6.3, instead of checking if the price is different from the standard price, CAB decided to verify whether the specified price of the product is more than the standard price.

Solution. It is obviously a one-tailed test, because, if average price obtained from the sample is less than or equal to the standard price, the claim of the producer will be established. Hence we have to consider a composite hypothesis defined as

Null hypothesis: $H_0: \mu \leq 1500$

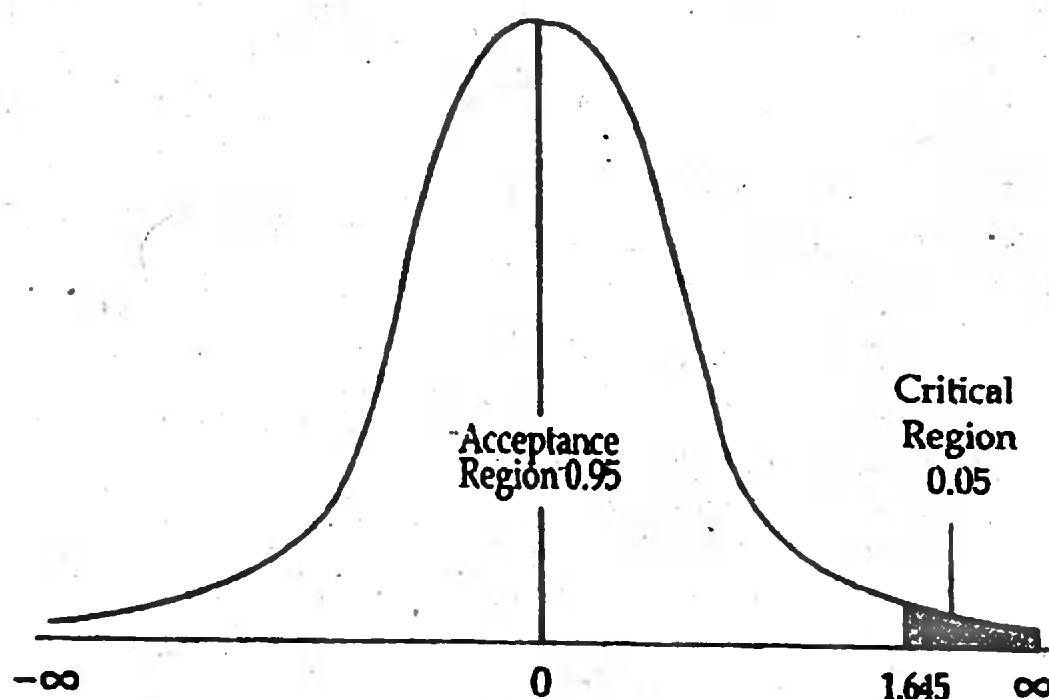
Alternative hypothesis $H_1: \mu > 1500$

The test statistic for testing the null hypothesis is

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Here $\alpha = 0.05$ (given).

It is a right-tailed test, the critical region will be on right side of curve of Z that will comprise 5% or 0.05 area at the right end. From the table of area of normal curve, we see that this critical value of Z is 1.645, that means the critical region is $Z > 1.645$ (at the right end).



We have found $z = 2.22$ (from Example 16.6.3)

Hence, the observed value of $Z = 2.22$ which is greater than the critical value 1.645 , hence, the observed value of test statistic falls in the critical region, we fail to accept the null hypothesis at 5% level of significance.

The claim of the producer is not right, that means, the average price of the products of the producer is more than the standard price.

p-value. The p-value is given by $P(Z > 2.22) = .0132$, that means the smallest level of significance at which the hypothesis is rejected is 0.0132 or 1.3% .

Example 16.6.12. The average petrol consumption of existing auto engines is 10.5 km per liter. 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. In order to verify company's claim, a sample of 50 new cars was randomly selected and it was found that the mean petrol consumption was 10 km per liter with a standard deviation of 3.5 km per liter. Test whether the claim of the company is acceptable at 5% level of significance.

Solution. Here we have to consider the following hypotheses

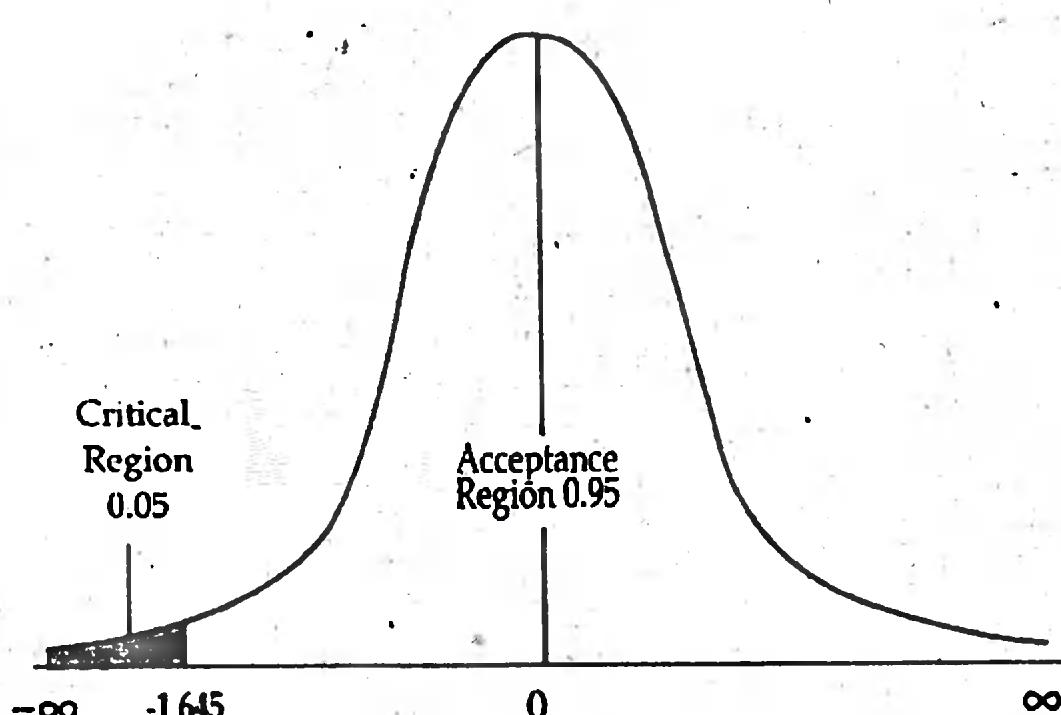
$$H_0: \mu = 10.5 \text{ and}$$

$$H_1: \mu < 10.5$$

Although the population variance is not known, since sample size is large,

the appropriate test statistic is $Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ where s^2 is sample variance and n is the sample size.

Since it is a left tailed test, the critical region lies in the left end of the curve given by $Z < -1.645$.



Given, $\bar{x} = 10$ km, $s = 3.5$, $n = 50$, so the value of test statistic Z under H_0 is given by

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{10 - 10.5}{\frac{3.5}{\sqrt{50}}} = -1.01$$

This observed value of Z does not lie in the critical region, so we fail to reject the null hypothesis. That means the claim of the company is not acceptable.

Example 16.6.13. Suppose the daily number of items produced by a firm for randomly selected 15 days are as follows:

110, 118, 130, 140, 142, 146, 112, 100, 95, 98, 96, 122, 123, 124, 130

Can we conclude at 5% level of significance that the average daily production of items of that firm is 110?

Solution: It is a two-tailed test because if the average hourly number of items produced by the company is more or less than 110, then the statement that the average daily production of items will be proved as false, then the null hypothesis that the average daily production of items is 110 would be rejected.

So, we the null hypothesis and alternative hypothesis are as follows:

Null hypothesis $H_0: \mu = 110$

Alternative hypothesis $H_a: \mu \neq 110$

Level of significance as given in the problem is $\alpha = 0.05$, since the variance is unknown and sample size is small, the appropriate test statistic is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \text{ which is distributed as Student's t with } n-1 \text{ df.}$$

Here, $n = 15$, so the df = $15 - 1 = 14$

Here $\alpha = 0.05$, and it a two tailed-test, the critical region will be on both sides of t-distribution, in such a way that the critical region will comprise 2.5% or 0.025 area at the right end and 2.5% at the left end. From the table of t-distribution, we find for df = 14, $\alpha = 0.025$, the values of t are ± 2.145 , that means the critical regions are $t < -2.145$ (at the left end) and $t > 2.145$ (at the right end).

Here, $\bar{x} = 119.07$, $\mu = \mu_0 = 110$

$$\begin{aligned} s^2 &= \frac{\sum (X - \bar{X})^2}{n-1} = \frac{1}{n-1} \left[\sum X^2 - \frac{(\sum X)^2}{n} \right] \\ &= \frac{1}{14} \left[216682 - \frac{(1786)^2}{15} \right] = 287.78, \text{ so, } s = 16.96 \text{ and } n = 15. \end{aligned}$$

$$\text{So, we have, } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{119.07 - 110}{16.96/\sqrt{15}} = 2.07$$

It is found that the observed value 2.07 is less than the critical value 2.145, hence it does not fall in the critical region, so we fail to reject the null hypothesis at 5% level of significance.

Conclusion. It can be concluded that the average daily production of the items of the given firm may be accepted as 110.

Example 16.6.14. (Small sample with unknown variance) A gas station repair shop claims that the average time it takes to do a lubrication job and oil change is maximum 30 minutes. The consumer protection department wants to test the claim. A sample of six cars were sent to the station for oil change and lubrication. The job took an average of 34 minutes with a standard deviation of 4 minutes. Assuming that the population is normal, do you think that the job took an average of time more than 30 minutes? (use $\alpha = 0.05$)

Solution. It is obviously a one-tailed test, because, the claim will not be rejected if the average time taken on a car for the oil change and lubrication job is considerably less than 30 minutes. So, in this case we have to consider a composite hypothesis given by (although in practice we use simple hypothesis):

$$H_0 : \mu = \mu_0 \leq 30 \text{ against the alternative } H_1 : \mu = \mu_1 > 30$$

Since the sample size is small and population variance is unknown, t statistic will be used.

Under the null hypothesis the value of t is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ which is distributed as Student's t with } n - 1 \text{ df}$$

$$\text{Here, } n = 6, \bar{x} = 34, \mu_0 = 30, s = 4, \text{ and } \hat{\sigma}(\bar{x}) = s/\sqrt{n} = 1.63,$$

$$\text{therefore, we get, } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{34 - 30}{1.63} = 2.45$$

The critical value of t at $\alpha = 0.05$ and $df = n - 1 = 5$, for a right tailed test is given by $t_{5,05} = 2.02$, since the computed value of $t = 2.45$ is higher than the critical value of $t = 2.02$, we reject the null hypothesis, that means the claim of the shop can not be considered to be correct.

p-value: In this case p-value is given by $P(t_5 > 2.45) = 0.03$ (from the table of critical value of t distribution), hence p-value is 0.03 or 3%.

Example 16.6.15. A process that produces bottles of shampoo, when operating correctly, produces bottles whose contents weigh, on average, 20 ounces. A random sample of nine bottles from a single production run yielded the following weights

21.4, 19.7, 19.7, 20.6, 20.8, 20.1, 19.7, 20.3, 20.9

Assuming that the population distribution is normal, test the hypothesis that the process is operating correctly at 5% level of significance. Also calculate 95% confidence interval for population mean.

Solution. It is a two-tailed test, because, the claim will be rejected if the average weight deviates from 20 ounces in any direction. So, the hypothesis are given by

$$H_0 : \mu = \mu_0 = 20 \text{ against the alternative } H_1 : \mu \neq 20$$

Since the sample size is small and population variance is unknown, t statistic will be used.

Under the null hypothesis the value of t is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ which is distributed as Student's } t \text{ with } n - 1 \text{ df}$$

Here, $n = 9$, $\bar{x} = 20.36$, $\mu_0 = 20$, $s = 0.61$, and $\sigma(\bar{x}) = s/\sqrt{n} = 0.203$,

$$\text{Hence, } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{20.36 - 20}{0.203} = 1.77$$

The critical values of t at $\alpha = 0.05$ with $n - 1 = 8$ df for a two-tailed test are given by $\pm t_{8;025} = \pm 2.316$.

Since the computed value of t is 1.77 which does not fall in the critical region, so we fail to reject the null hypothesis, that means the process is operating correctly.

95% confidence interval for population mean μ is given by

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} = 20.36 \pm 2.316 \times \frac{0.61}{\sqrt{9}}$$

So, the lower and upper confidence limits are 19.89 and 20.83 respectively.

16.7. Test of Hypothesis Concerning Two Population Means

Suppose $X_{11}, X_{12}, \dots, X_{1n_1}$ be a random sample of size n_1 drawn from normal population with mean μ_1 and variance σ_1^2 , and $X_{21}, X_{22}, \dots, X_{2n_2}$ be

another sample of size n_2 drawn from normal population with mean μ_2 and variance σ_2^2 . Suppose, the observed sample means are \bar{X}_1 and \bar{X}_2 . In the earlier chapter, it is mentioned that the difference between two sample means follows normal distribution if the sample sizes are large, that means

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

The possible null and alternative hypotheses considered for testing the significance of difference between two population means are:

- i) $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$ (for a two tailed alternative)
- ii) $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 > \mu_2$ (for a right tailed alternative)
- iii) $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 < \mu_2$ (for a left tailed alternative)

Again, the following alternative situations may arise in testing the above mentioned null and alternative hypotheses

- i) Independent samples with known population variances, sample sizes are large or small
- ii) Independent samples with unknown population variances, sample sizes are large
- iii) Independent populations for small sample sizes (≤ 29) with unknown but equal variances
- iv) Independent populations for small sample sizes (≤ 29) with unknown and unequal variances
- v) Correlated sample or matched sample (the sample obtained from a bi-variate normal population or paired observations)

The test statistic to be used for testing the simple hypothesis $H_0: \mu_1 - \mu_2 = 0$ against a one-tailed or two-tailed alternative is given by

Under situation (i): $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$

Under situation (ii): $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$

where $s_1^2 = \frac{\sum(x_1 - \bar{x}_1)^2}{n_1 - 1}$ and $s_2^2 = \frac{\sum(x_2 - \bar{x}_2)^2}{n_2 - 1}$

$$\text{Under situation (iii): } t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which is distributed as Student's t with $n_1 + n_2 - 2$ degrees of freedom, where, s^2 is pooled-estimate of variance, given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The value of $(\mu_1 - \mu_2) = 0$ for all Z and t defined above under null hypothesis $H_0 : \mu_1 - \mu_2 = 0$, but if any other value is specified (say, $\mu_1 - \mu_2 = \theta_0$) by null hypothesis for the difference between means, that means if

$$H_0 : \mu_1 - \mu_2 = \theta_0$$

then the value of $(\mu_1 - \mu_2)$ will be θ_0 instead of zero.

For testing the hypothesis regarding difference between two means under situation i) to iii) it is desirable to test the equality of two population variance to check if the assumptions of equal variances are valid or not. If the variances are found not to be equal, then, Student's t statistic can not be applied. Hence, under situation (iv), under null hypothesis, the test statistic is given by

$$t' = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Here, t' is not a Student's t statistic. The critical values of t' at $100\alpha\%$ level of significance are computed using the formula

$$t'_\alpha = \frac{\frac{s_1^2 t_1}{n_1} + \frac{s_2^2 t_2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t_1 and t_2 are Student's with $(n_1 - 1)$ and $(n_2 - 1)$ df respectively at $100\alpha\%$ level of significance.

Again, under situation (v), let us consider a random sample of n matched pairs of observations (x_i, y_i) from a bi-variate normal population, then the test of the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ requires to compute a statistic d defined as $d = x_i - y_i$, and the testing procedure is the same as testing the significance of a single mean of the observations obtained from the

difference of two variables assuming that small sample size and unknown population variance. The statistic used for testing this type of hypothesis is called paired-t test, defined as

$$t = \frac{\bar{d}}{se(\bar{d})} \sim t_{n-1} \text{ under situation (v)}$$

which is distributed as t with $n - 1$ df

$$\text{where, } \bar{d} = \frac{\sum d}{n}, s_d^2 = \frac{\sum (d - \bar{d})^2}{n-1} \text{ and } se(\bar{d}) = \frac{s_d}{\sqrt{n}}$$

Decision rule. The decision rules in all of the above cases are the same as that of testing the significance of a single mean using Z or t statistic.

Example 16.7.1. It is wanted to investigate if male and female typists earn comparable wages. The sample data for daily wages of male and female provide with the following information.

Table 16.6. Sample mean and variance of male and female typists.

	Male	Female
Sample size	60	60
Mean wage	Taka 158.50	Taka 141.60
SD (Population)	Taka 18.20	Taka 20.60

Test whether the mean wages of male typists is more than female typists at 5% and 1% level of significance.

Solution. Let the wages of male and female are normally and independently distributed with means μ_1 and μ_2 , and known variances σ_1^2 and σ_2^2 respectively. It is a one-tailed test, so we consider the following hypothesis

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 > \mu_2$$

Assuming that the sample sizes are large, under null hypothesis, the value

of test statistic is given by $z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

For $\alpha = 0.05$, the critical region is $Z > 1.645$, and for $\alpha = 0.01$, the critical region is $Z > 2.33$.

Here, $\bar{x}_1 = 158.50$, $\bar{x}_2 = 141.6$, $\sigma_1 = 18.20$, $\sigma_2 = 20.60$, $n_1 = n_2 = 60$

Thus, the computed value of Z is: $z = \frac{(158.50 - 141.6)}{\sqrt{\frac{(18.20)^2}{60} + \frac{(20.60)^2}{60}}} = 4.76$

Conclusion. The computed value of Z is greater than critical values at both the level of significance, hence in the given city, male typists have on the average higher earnings than their female counterpart at 1% and 5% levels of significance. Here, the value of z is highly significant.

Example 16.7.2. (Large sample sizes with unknown population variances) A potential buyer of electric bulbs bought 100 bulbs each of two famous brands A and B. Upon testing both these samples, he found that brand A had a mean life of 1500 hours with standard deviation of 50 hours whereas brand B had an average life of 1530 hours with standard deviation of 60 hours. Can it be concluded at 5% level of significance that the bulbs of two brands differ significantly in quality?

Solution. We assume that the parent population of these two lifetimes are independently distributed with means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2 . We also assume that there is no significant difference in the quality of both brands so that brand A is as good as brand B in terms of operating hours. Hence, we have to test

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 \neq \mu_2$$

Since sample sizes are large, so under the null hypothesis, the value of test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

Again, $\alpha = 0.05$ and since it is a two tailed test, the critical region is

$$|Z| > 1.96$$

Here, $\bar{x}_1 = 1500$, $\bar{x}_2 = 1530$, $s_1 = 50$ and $s_2 = 60$, $n_1 = n_2 = 100$

Now, $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(50)^2}{100} + \frac{(60)^2}{100}} = 7.81$

Thus, $z = \frac{(1500 - 1530)}{7.81} = -3.841 = -3.841$

Since the observed value of Z is more than the critical value, the null hypothesis may be rejected at 5% level of significance.

p-value. From the table of Z, we have $P(Z < -3.00) = P(Z > 3.00) = 0.00$, hence the null hypothesis may be rejected even at 0% level of significance, so, $p = 0.00$. It is said that the value is highly significant since the p-value is 0.00.

Confidence Interval. 95% confidence limits for the difference of two population means ($\mu_1 - \mu_2$) are given by

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (-30) \pm 7.81 = (22.19, 37.81)$$

Example 16.7.3. (Large sample sizes with unknown population variances) A professor taught two sections of an introductory marketing course using very different styles. In the first section approach was extremely formal and rigid, while in the second section an independent, more relaxed and informal attitude was adopted. At the end of the course, a common final examination was administered. In the first section the seventy two students obtained a mean score of 71.03 and the sample standard deviation was 22.91. In the second section there were sixty four students, with mean score 80.92 and standard deviation 23.11. Assume that these two groups of students can be regarded as independent random samples from the populations of all students who might be exposed to these teaching methods. Test at 5% level of significance (i) whether the performance of these two methods are the same, (ii) whether second method is better than first one.

Solution. We assume that the parent population of these two lifetimes are independently distributed with means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2 .

(i) We have to test that there is no significant difference in two method, so that first method is as good as second method as far as the teaching system is concerned. Hence, we have to test

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 \neq \mu_2$$

Since sample sizes are large, so under null hypothesis the value of test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

Again, $\alpha = 0.05$ and since it is a two tailed test, the critical region is

$$|Z| > 1.96$$

Here, $\bar{x}_1 = 71.03$, $\bar{x}_2 = 80.92$, $s_1 = 22.91$ and $s_2 = 23.11$, $n_1 = 72$, $n_2 = 64$

$$\text{Now, } \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(22.91)^2}{72} + \frac{(23.11)^2}{64}} = 3.95$$

$$\text{Thus, } z = \frac{(71.03 - 80.92)}{3.95} = -2.50$$

Since the computed absolute value of Z is greater than the critical value, the null hypothesis may be rejected at 5% level of significance. The sample supports that there is significant difference between average performance of the two methods.

p-value. From the table of Z, we have $P(Z < -2.50) = P(Z > 2.50) = 0.007$, hence the null hypothesis may be rejected even at 1.4% (since $0.007 \times 2 = 0.014$) level of significance, so $p = 0.014$.

Confidence Interval. 95% confidence limits for the difference of two population means ($\mu_1 - \mu_2$) are given by

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (-9.89) \pm 7.81 = (-17.70, 2.08)$$

(ii) Since the second method would be proved to be better than the first method, if the average score obtained by first method is significantly smaller than that of second method, so, in order to test whether second method is better than the first method, we have to consider a one-tailed test defined as

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 < \mu_2$$

Here, since the sample sizes are large, the test statistic is also Z as defined in (i), but the critical region will cover the 5% area only in the left tail, thus the critical region is $Z < -1.645$.

The observed value of Z is -2.50, as found in (i), falls in the critical region, so the null hypothesis may be rejected at 5% level of significance. That means, the sample supports that the second method is on an average better than the first method.

Example 16.7.4. (Large sample sizes with known variances) A firm believes that the tires produced by process I on an average last longer than tires produced by process II. To test this belief, random samples of tires produced by two processes were tested and the results are as

Table 16.7: Mean and standard deviation of lifetimes of tires.

	Process I	Process II
Sample size	50	50
Average lifetime (in km.)	22,400	21,800
Population Standard deviation (in km.)	1000	1000

Is there any evidence at 5% level of significance that the firm is correct in its belief?

Solution. We have to consider the following null and alternative hypothesis

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 > \mu_2$$

Since the sample sizes are large, the value of test statistic under null hypothesis is $z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Since it is a one tailed test, the critical region is $Z > 1.645$

Given, $\bar{x}_1 = 22400$, $\bar{x}_2 = 21800$, $\sigma_1 = 1000$, $\sigma_2 = 1000$, $n_1 = n_2 = 50$.

$$\text{So, the value of } Z \text{ is } z = \frac{(22400 - 21800)}{\sqrt{\frac{(1000)^2}{50} + \frac{(1000)^2}{50}}} = 3.00$$

Since the calculated value of Z is more than its critical value at 5% level, therefore null hypothesis may be rejected. Hence, we can conclude that the tires produced by process I has longer life than process II.

p-value. From the table of area under the standard normal probability distribution, we have $P(Z > 3.00) = 0.001$, so the smallest critical value for which null hypothesis may be rejected is 0.001, thus the p-value is 0.001.

Example 16.7.5. (Small sample sizes with unknown and equal population variance) Manager of a factory I claims that the average wage of its workers is higher than that of factory II. A firm conducted a survey on daily wages of workers of two factories to see if the claim of manager is justified or not. The summary of sample statistics obtained were as follows.

Table 16.8. Mean and standard deviation of lifetimes of tires.

	Factory I	Factory II
Sample size	16	11
Sample mean wage	Taka 290	Taka 250
Sample standard deviation	15	50

Solution. Let us assume that the wages of corresponding population of two factories are independent and normally distributed with common unknown variance σ^2 . Thus the null and alternative hypothesis are $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 > \mu_2$.

Since the sample sizes are small and variances are unknown but equal, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which is distributed as Student's t with $n_1 + n_2 - 2$ degrees of freedom,

where, $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is an estimate of common variance σ^2

Here, $n_1 = 16, n_2 = 11$, so the degrees of freedom is $n_1 + n_2 - 2 = 25$

It is a one tailed test, thus from the table of t-distribution we have the critical values of t at 5% level of significance with 25 df is 1.708, that means, $t_{25,0.05} = 1.708$, in other words, the critical region is $t > 1.708$.

Given, $\bar{x}_1 = 290, \bar{x}_2 = 250, s_1 = 15, s_2 = 50, s^2 = 1135$ and $s = 33.69$

So, under the null hypothesis, the value of t is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{290 - 250}{33.69 \sqrt{\frac{1}{16} + \frac{1}{11}}} = 3.03$$

Decision. The observed value of t is greater than the critical value, it lies in the critical region, so the null hypothesis may be rejected at 5% level of significance.

Conclusion. The claim of manager of factory-I is justified.

Example 16.7.6. (Small sample sizes with unknown and equal population variances) The residence of Dhaka city complains that traffic speeding fines given in their city are higher than the traffic speeding fines that are given in Chittagong city. The appropriate authority agreed to study the problem. To check if the complaints were reasonable, independent random samples of the amounts paid by the residents for speeding fines in each of two cities over the last three months were obtained and shown in following table.

Table 16.9. Amounts of traffic speeding fines.

Dhaka city	100	125	135	128	140	142	128	137	156	142
Chittagong city	95	87	100	75	110	105	85	95		

Assuming an equal population variance, test

- (i) whether there is any significant difference in the mean cost of speeding in these two cities and find the 95% confidence interval.
- (ii) whether the mean speeding cost in Dhaka city is higher than Chittagong city at 1% level of significance.

Solution. (i) Let X_1 be the speeding cost in Dhaka city and X_2 be the speeding cost in Chittagong city. Assuming the samples have been drawn independently from two normal populations with means μ_1 and μ_2 respectively with common variance σ^2 . We have to test the following hypothesis

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 \neq \mu_2$$

It is a two tailed test, the sample sizes are small and variances are also unknown, so, the appropriate test statistic is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which is distributed as Student's t with $n_1 + n_2 - 2$ degrees of freedom,

where, $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is an estimate of common variance σ^2 .

Given, $n_1 = 10, n_2 = 8$, so the degrees of freedom is $n_1 + n_2 - 2 = 16$

From the table of t-distribution we have the critical values of t at 5% level of significance with 16 df are ± 2.12 , that means, $t_{16,0.025} = 2.12$, in other words, the critical region is $|t| > 2.12$

From the given observations, we have,

$$\bar{x}_1 = 133.30, \bar{x}_2 = 94.00, s_1 = 18.20, s_2 = 20.60, s^2 = 371.98$$

So, under null hypothesis, the calculated value of t is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{133.30 - 94.00}{s \sqrt{\frac{1}{10} + \frac{1}{8}}} = \frac{133.30 - 94.00}{9.14} = 4.30$$

Decision. The observed absolute value of t is greater than the critical value, it lies in the critical region, so the null hypothesis may be rejected at 5% level of significance.

Conclusion. There is significant difference between the average traffic fines in two cities.

Again, the 95% confidence interval for $(\mu_1 - \mu_2)$ is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{16,0.025} \times s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + t_{16,0.025} \times s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{or, } 39.30 \pm 2.12 \times 4.30 = 39.30 \pm 9.12 = (30.18, 48.42)$$

P-value. From the table of critical values of t-distribution, we find that $t_{16,0.005} = 2.921$, that means $2 \times 0.005 = 0.010 = 1\%$ can be considered as the smallest level of significance at which the null is still rejected, so the required p-value is 0.01.

(ii) In this case we have to perform a one-tailed test given by

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 > \mu_2$$

Here the critical value of t at 1% level of significance with 16 df is $t_{16,0.01} = 2.583$

The observed value of t is same as previous one, so $t = 4.30$.

In this case too, the observed t falls in the critical region, so the null hypothesis may be rejected at 1% level of significance. We can conclude that on an average the traffic fines at Dhaka city is higher than that of Chittagong city.

p-value. In this case the p-value is 0.005 because, $P(t > 4.30) = 0.005$.

Example 16.7.7. (Matched observations or paired sample) A study was conducted by a pharmaceutical company to compare the difference in effectiveness of two particular drugs in cholesterol levels. The company used paired sample approach to control variation in reduction that might be due to factors other than the drug itself. Each member of a pair was matched by age, weight, lifestyle, and other pertinent factors. Drug X was given to one person randomly selected from each pair, and drug Y was given to the other individual in the pair. After a specific period of time each person's cholesterol level was measured again. Suppose a random sample of eight pairs of patients with known cholesterol problems is selected from the large populations of participants. The following table gives the number of points by which each person's cholesterol level was reduced.

Table 16.10. Reduction levels of cholesterol by drugs :

Pair	1	2	3	4	5	6	7	8
Drug X	29	32	31	32	32	29	31	30
Drug Y	26	27	28	27	30	26	33	36

- (i). Test whether there is any significant difference between the mean reduction of cholesterol levels by two drugs at 1% level of significance.

- (ii) Find 99% confidence interval for the difference between the population means

Solution. (i) Let the observations have been selected from a bi-variate normal population. So, it is necessary to use paired t-test for testing the difference between the mean reduction levels.

Let us formulate the null and alternative hypothesis as

$$H_0: \mu_x = \mu_y \text{ against } H_1: \mu_x \neq \mu_y$$

The appropriate test statistic is $t = \frac{\bar{d}}{se(\bar{d})} \sim t_{n-1}$, which is distributed as t

with $n - 1$ df,

$$\text{where, } \bar{d} = \frac{\sum d}{n}, s_d^2 = \frac{\sum (d - \bar{d})^2}{n-1} \text{ and } se(\bar{d}) = \frac{s_d}{\sqrt{n}}$$

Here, $n = 8$, degrees of freedom is $n - 1 = 7$

It is two-tailed test, so at 1% level of significance the critical region is

$$|t| > t_{n-1, \alpha/2} = t_{7, 0.005} = 3.499.$$

Now, let us construct the following table for calculation of mean and standard deviation of d.

Table 16.11. Calculation of mean and standard deviation of d

Pair	1	2	3	4	5	6	7	8
Drug X	29	32	31	32	32	29	31	30
Drug Y	26	27	28	27	30	26	33	36
$d = X - Y$	3	5	3	5	2	3	-2	-6
$(d - \bar{d})$	1.38	3.38	1.38	3.38	0.38	1.38	-3.63	-7.63
$(d - \bar{d})^2$	1.89	11.39	1.89	11.39	0.14	1.89	13.14	58.14

We have $\bar{d} = 1.625$, $s_d^2 = 14.27$, $s_d = 3.777$

$$\text{So, } t = \frac{\bar{d}}{se(\bar{d})} = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{1.625}{3.777/\sqrt{8}} = 1.22$$

Decision. The observed value of t is smaller than absolute value of critical value, that means it does not fall in the critical region, so we fail to reject null hypothesis.

Conclusion. There is no significant difference between the average reduction of cholesterol level by two drugs, that means, drug X and drug Y are equally effective.

(ii) The 99% confidence interval for the difference of population average reduction of cholesterol levels is given by $\bar{d} - t_{n-1, \alpha/2} \cdot se(\bar{d}) < \mu_x - \mu_y < \bar{d} + t_{n-1, \alpha/2} \cdot se(\bar{d}) = (-3.05, 6.30)$.

Example 16.7.8. (Matched observations or paired sample): Ten persons were appointed as probationary officers in an office. Their performance was noted by taking a test and the marks were recorded out of 100. After training for a 6-months period, another test was conducted. The marks obtained by the officers before and after training were as follows.

Table 16.12. Marks of employees before and after training.

Employees	A	B	C	D	E	F	G	H	I	J
Before training	80	76	92	60	70	56	74	56	70	56
After training	84	70	96	80	70	52	84	72	72	50

Were the employees benefited by the training?

Solution. It would be proved that the employees were not benefited by the training if there is no significant difference between the average score obtained by them before and after training. On the other hand, it would be proved to be benefited if the average score obtained by employees significantly improved after training.

So we have to conduct a one tailed test defined by the null and alternative hypothesis as

$$H_0: \mu_a = \mu_b \text{ against } H_1: \mu_a < \mu_b$$

The appropriate test statistic is $t = \frac{\bar{d}}{se(\bar{d})} \sim t_{n-1}$, which is distributed as t with $n-1$ df,

$$\text{where, } \bar{d} = \frac{\sum d}{n}, s_d^2 = \frac{\sum (d - \bar{d})^2}{n-1} \text{ and } se(\bar{d}) = \frac{s_d}{\sqrt{n}}$$

Here, $n = 10$, degrees of freedom is $n - 1 = 9$,

It is one-tailed test, so at 5% level of significance the critical region is

$$t < -t_{n-1, \alpha} = -t_{9, 0.05} = -2.62$$

Now, let us construct the following table for calculation of mean and standard deviation of d .

Table 16.13. Computation of mean and standard deviation of d

Employees	A	B	C	D	E	F	G	H	I	J
Before training (X)	80	76	92	60	70	56	74	56	70	56
After training (Y)	84	70	96	80	70	52	84	72	72	50
$d = X - Y$	-4	6	-4	-20	0	4	-10	-16	-2	6
$(d - \bar{d})$	0.00	10.00	0.00	-16.00	4.00	8.00	-6.00	-12.00	2.00	10.00
$(d - \bar{d})^2$	0.00	100.00	0.00	256.00	16.00	64.00	36.00	144.00	4.00	100.00

We have $\bar{d} = -40/10 = -4$, $s_d^2 = 80$, $s_d = 8.944$

$$\text{So, } t = \frac{\bar{d}}{s_e(d)} = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{-4}{8.944/\sqrt{10}} = -1.414$$

Decision. The observed value of t does not fall in the critical region, so we fail to reject null hypothesis, that means the null hypothesis holds true.

Conclusion. It can be concluded that the employees were not benefited by training.

16.8. Test of Hypothesis Concerning Attributes

In case of attributes we can only find out the presence or absence of a certain qualitative characteristics. For example, in the study of attribute 'employment', a survey may be conducted and the people may be classified as employed and unemployed, in the study of the attribute 'effectiveness' of a drug, the patients may be classified as cured and not cured and in the study of attribute 'size', the items may be classified as small, medium and large. The appearance of an attribute may be considered as success and its non-appearance as failure. Obviously, this type of two outcomes will follow binomial distribution. Thus, when the sample size is large, we can perform the following tests with the attributes having two categories:

- i) Test of a population proportion for a specified value.
- ii) Test of the difference between two population proportions.
- iii) Tests of independence of two attributes (having two or more categories of each attribute)

The test regarding independence of attributes is discussed in section 16.12.

16.8.1. Test of hypothesis about a population proportion. Suppose we have a sample of n observations from a population, a proportion P of population follow a particular attribute. Then, if the number of sample observations n is large, and the observed sample proportion is P . For testing the significance of this population proportion for a given value π_0 , we consider the following null and alternative hypothesis:

$H_0 : \pi = \pi_0$, against the alternative $H_1 : \pi \neq \pi_0$ (two tailed test)

We know, the sampling distribution of P is normal with mean π and variance $\sigma_p^2 = \frac{\pi(1-\pi)}{n}$ and under H_0 it is $\frac{\pi_0(1-\pi_0)}{n}$.

So, under the null hypothesis,

$$\text{the test statistic is } z = \frac{P - \pi_0}{\text{se}(P)} = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$$

The decision rule is: Reject H_0 in favor of H_1 at $100\alpha\%$ level of significance

$$\text{if } z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > z_{\alpha/2} \text{ or } \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < -z_{\alpha/2} \text{ for a two-tailed test,}$$

$$\text{or if } |Z| > z_{\alpha/2}$$

Similarly, for a right tailed test, the decision rule is :

$$\text{Reject } H_0 \text{ in favor of } H_1 \text{ at } 100\alpha\% \text{ level of significance if } \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > z_\alpha$$

and for a left-tailed test, the decision rule is :

Reject H_0 in favor of H_A at $100\alpha\%$ level of significance if

$$\frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < -z_\alpha$$

Example 16.8.1. For the following questions carry out the test of significance of population proportions at 5% level of significance (where x represents the number of things of particular category):

- i) $H_0 : \pi = 0.25, H_1 : \pi \neq 0.25; n = 100, x = 40$
- ii) $H_0 : \pi = 0.40, H_1 : \pi > 0.40; n = 200, x = 100$
- iii) $H_0 : \pi = 0.30, H_1 : \pi < 0.30; n = 400, x = 100$

Solution. The statistic to be used for testing the given hypothesis is given by

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \text{ where, } P \text{ is the estimate of proportion.}$$

- (i) This is a two tailed test, so the decision rule is: Reject H_0 in favour of alternative at 5% level if $|Z| > z_{0.025}$ or $|Z| > 1.96$

Here, $n = 100$, so, $P = \frac{40}{100} = 0.40$ and $\pi_0 = 0.25$, so the computed value of

test statistic is $z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.40 - 0.25}{\sqrt{\frac{0.25 \times (1 - 0.25)}{100}}} = 3.46$

Decision. $|Z| > 1.96$, the computed value of Z falls in the critical region, so we fail to accept null hypothesis.

(ii) This is a right tailed test, so the decision rule is: Reject H_0 in favour of alternative at 5% level if $Z > z_{0.05} = 1.645$

Here, $n = 200$, so, $P = \frac{100}{200} = 0.50$ and $\pi_0 = 0.40$, so the computed value of Z

is $z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.50 - 0.40}{\sqrt{\frac{0.40 \times (1 - 0.40)}{200}}} = 2.89$

Decision. The observed value of $Z > 1.645$, which falls in the critical region, so we fail to accept null hypothesis.

(iii) This is a left tailed test, so the decision rule is: Reject H_0 in favour of alternative at 5% level if $Z < -z_{0.05} = -1.645$

Here, $n = 400$, so, $P = \frac{100}{400} = 0.25$ and $\pi_0 = 0.30$,

so the computed value of Z is $z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.25 - 0.30}{\sqrt{\frac{0.30 \times (1 - 0.30)}{400}}} = -2.18$

Decision. The observed value of $Z < -1.645$, which falls in the critical region, so we fail to accept null hypothesis.

Example 16.8.2. Forecasts of corporate earnings per share are made on a regular basis by many financial analysts. In a random sample of 600 forecasts, it was found that 382 of these forecasts exceeded the actual outcome for earnings. Test against a two tailed alternative the null hypothesis that the population proportion of forecasts that are higher than actual outcomes is 0.50 at 5% level of significance.

Solution. Let P denotes the population proportion and p denotes the sample proportion of forecasts that are above the actual outcomes. We are interested to test the hypothesis

$$H_0 : \pi = \pi_0 = 0.50, \text{ against the alternative } H_1 : \pi \neq 0.50$$

The appropriate test statistic is $Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$

The decision rule is: Reject H_0 in favour of alternative if

$$|Z| > z_{\alpha/2} \text{ or } |Z| > 1.96$$

Here, $n = 600$, $P = \frac{382}{600} = 0.637$ and $\pi_0 = 0.50$

So, the computed value of test statistic is

$$z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.637 - 0.50}{\sqrt{\frac{0.50(1 - 0.50)}{600}}} = 6.71$$

Since 6.71 is much bigger than 1.96, the null hypothesis is clearly rejected. That means, forecasts of corporate earnings that exceed the actual values are significantly different from 0.50.

Example 16.8.3. A manufacturer claims that at least 95% of the equipments which he supplied to a factory conformed to the specification. An examination of the sample of 200 pieces of equipment revealed that 18 were faulty. Test the claim of manufacturer at 5% level of significance.

Solution. According to the statement of the problem, it is better if we consider a composite hypothesis such that at least 95% if the equipments supplied by the company conformed to the specification, that means

$H_0: \pi \geq \pi_0 = 0.95$, against the alternative $H_1: \pi < 0.95$ (one tailed test)

The appropriate test statistic is $Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$

$\alpha = 0.05$, the decision rule is : Reject H_0 in favour of alternative if $Z < -z_{\alpha/2}$ or $Z < -1.645$ (for this left-tailed test).

Here, $n = 200$, out of 200, 18 were found faulty, that means $(200-18) = 182$ equipments conform to the specification, so $P = \frac{182}{200} = 0.91$ and $\pi_0 = 0.95$

So, the computed value of test statistic is

$$z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.91 - 0.95}{\sqrt{\frac{0.95(1 - 0.95)}{200}}} = -2.67$$

Since the observed value of $z = -2.67$ is less than the critical value -1.645 , it lies in the critical region, so the null hypothesis is clearly rejected. That means, the proportion of equipments conforming to the specification is greater than 95%.

Example 16.8.4. An auditor claims that 10 percent of the customers' ledger accounts of a bank are carrying mistakes of posting and balancing. A random sample of 600 accounts was taken to test the accuracy of posting and balancing, and 45 accounts were found to have mistakes. Are these sample results consistent with the claim of auditor? (use 5% level of significance).

Solution. Let us take the null that the claim of the auditor is valid, that means, $H_0: \pi = \pi_0 = 0.10$, against the alternative $H_1: \pi_0 \neq 0.10$

$$\text{The appropriate test statistic is } Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}.$$

$\alpha = 0.05$, the decision rule is: Reject H_0 in favour of alternative if

$$|Z| > z_{\alpha/2} \text{ or } |Z| > 1.96$$

$$\text{Here, } n = 600, \text{ so, } P = \frac{45}{600} = 0.075 \text{ and } \pi_0 = 0.10$$

So, the computed value of test statistic is

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.075 - 0.10}{\sqrt{\frac{0.10(1 - 0.10)}{600}}} = -2.049$$

Since the observed value of $|z| = 2.049$ is greater than the critical value 1.96, it lies in the critical region, so the null hypothesis is rejected at 5% level of significance. That means, the claim of the auditor is not valid.

Example 16.8.5. Suppose, a machine produces 12% faulty items. A manufacturer of the same type of machine claims that there machine is better than this machine. In order to test the manufacturer's claim, a random sample of 300 items were checked and 30 items were found to be faulty. On the basis of the information, comment on the claim of manufacturer.

Solution. The manufacturer's claim will be proved to be justified if it produces less proportion of defective items than the existing one. So, let us think that the existing machine is as better as the new machine, and consider the null hypothesis as

$H_0 : \pi = \pi_0 = 0.12$, against the alternative $H_1 : \pi_0 < 0.12$ (one tailed test)

The appropriate test statistic is $Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$

Let $\alpha = 0.05$, the decision rule is: Reject H_0 in favour of alternative if

$$Z < -z_{\alpha/2} \text{ or } Z < -1.96.$$

Here, $n = 300$, so, $P = \frac{30}{300} = 0.10$ and $\pi_0 = 0.12$

So, the computed value of test statistic is

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.10 - 0.12}{\sqrt{\frac{0.12(1 - 0.12)}{300}}} = -1.066$$

Since the observed value of $z = -1.066$ does not fall in the critical region, so we fail to reject the null hypothesis at 5% level of significance. That means, the claim of the manufacturer is not justified and the new machine is as better as the existing one.

16.8.2. Test of hypothesis about difference between two population proportions. Let P_1 and P_2 be the sample proportions obtained from large samples of sizes n_1 and n_2 from respective population having proportions π_1 and π_2 . We are interested to test the hypothesis that there is no difference between the population proportions, that means,

$H_0 : \pi_1 = \pi_2$, against the alternative $H_1 : \pi_1 \neq \pi_2$

The sampling distribution of difference between sample proportions $(P_1 - P_2)$ is normal with mean $E(P_1 - P_2) = \pi_1 - \pi_2$ and variance $\sigma_p^2 = \frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}$.

Since the sample sizes are large, the test statistic is

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}} \sim N(0,1)$$

However, since all tests are undertaken under null hypothesis, so if the null hypothesis is true, then P_1 and P_2 are two independent unbiased estimators of the same population parameter $\pi_1 = \pi_2 = \pi$. Thus the best estimate of the common proportion π is the pooled proportions P for two samples. The

pooled estimate of π is the weighted mean of two sample proportions, given by,

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

The test statistic Z then becomes,

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{(P_1 - P_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \text{ under } H_0$$

The decision rule is: For a two tailed test, reject H_0 in favor of H_1 at $100\alpha\%$ level of significance:

$$\text{if } \frac{(P_1 - P_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} > z_{\alpha/2} \text{ or, } \frac{(P_1 - P_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} < -z_{\alpha/2} \text{ or, if } |Z| > z_{\alpha/2}$$

Similarly, for a right tailed test, the decision rule is : Reject H_0 in favor of H_1 at $100\alpha\%$ level of significance if

$$\frac{(P_1 - P_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} > z_\alpha$$

and for a left-tailed test, the decision rule is : Reject H_0 in favor of H_1 at $100\alpha\%$ level of significance if

$$\frac{(P_1 - P_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} < z_\alpha$$

Example 16.8.6. A company is considering two different television advertisements for promotion of a new products. Manager 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 is used in other area. In a random sample of 60 customers who saw the advertisement A, 18 had tried to buy the product, on the other hand, in a random sample of 100 consumers who saw advertisement B, 22 had tried to buy the product. Does this indicate that the advertisement A is more efficient than advertisement B, if level of significance is 5%?

Solution. Let π_1 and π_2 be the population proportions of customers who had tried to buy the products after seeing the advertisement A and advertisement B respectively, then we consider the null hypothesis as both advertisements are equally effective, that means,

$$H_0 : \pi_1 = \pi_2, \text{ against the alternative } H_1 : \pi_1 > \pi_2$$

(one tailed test, because, advertisement A will be considered more effective if proportion of customers who had tried to buy in this case is more than that of advertisement B)

Under null hypothesis, the appropriate test statistic is

$$Z = \frac{(P_1 - P_2)}{\sqrt{P(1-P) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

where, P is the pooled estimate of proportion.

Given, $\alpha = 0.05$, and it is a right-tailed test, so the decision rule is: Reject H_0 in favour of H_1 if $Z > z_{0.05}$ or $Z > 1.645$

Here, $n_1 = 60$, proportions $P_1 = \frac{18}{60} = 0.30$ and $n_2 = 100$, $P_2 = \frac{22}{100} = 0.22$

and the pooled estimate of proportion is

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{60 \times 0.30 + 100 \times 0.22}{60 + 100} = 0.25$$

$$\text{Thus, } z = \frac{0.30 - 0.22}{\sqrt{0.25(1-0.25) \left\{ \frac{1}{60} + \frac{1}{100} \right\}}} = 1.131$$

Since, observed values of z is less than the critical value 1.645, we fail to reject the null hypothesis at 5% level of significance.

Hence, we can conclude that there is no significant difference in the effectiveness of the two advertisements.

Example 16.8.7. 800 units from factory A are inspected and 12 are found to be defective, 500 units from factory B are inspected and 12 are found to be defective. Can it be concluded at 5% level of significance that production at factory A is better than factory B?

Solution. Let π_1 and π_2 be the population proportions of defectives of factory A and factory B respectively, then we consider that null hypothesis that performance of both factories are the same, that means,

$$H_0: \pi_1 = \pi_2, \text{ against the alternative } H_A: \pi_1 < \pi_2$$

(it is a one tailed test, because, factory A can be considered as better if the proportion of defectives found in factory A is less than the proportion of defectives found in factory B)

The appropriate test statistic is $Z = \frac{(P_1 - P_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$

where, P_1 and P_2 are the sample proportions, and P is the pooled estimate of proportion given by $P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$

Given, $\alpha = 0.05$, and it is a left-tailed test, so the decision rule is: Reject H_0 in favour of H_1 if $Z < -z_{0.05}$ or, $Z < -1.645$

Here, $n_1 = 800$; $P_1 = \frac{12}{800} = 0.015$ and $n_2 = 500$, $P_2 = \frac{12}{500} = 0.024$ and the pooled estimate of proportion

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{800 \times 0.015 + 500 \times 0.024}{800 + 500} = 0.018$$

$$\text{Thus, } z = \frac{0.015 - 0.024}{\sqrt{0.018(1 - 0.018)\left\{\frac{1}{800} + \frac{1}{100}\right\}}} = -1.184$$

Since, observed values of Z is not less than the critical value -1.645 , we fail to reject the null hypothesis, that means the null hypothesis holds good at 5% level of significance.

Hence, we cannot conclude that the production at factory A is better than B.

Example 16.8.8. In a random sample of 700 workers from a particular factory of Bangladesh 200 are found to be smokers. In another district out of 1300 workers 400 were found to be smokers. Can you conclude that there is a significant difference between the two factories with regard to the smoking habit?

Solution: Let π_1 and π_2 be the population proportions of smokers in two cities respectively. Let us take the hypothesis that there is no difference in smoking habits in the two factories, then the null hypothesis is

$$H_0 : \pi_1 = \pi_2, \text{ against the alternative } H_1 : \pi_1 \neq \pi_2$$

The appropriate test statistic is $Z = \frac{(P_1 - P_2)}{\sqrt{P(1-P)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$

where, P_1 and P_2 are the sample proportions, and P is the pooled estimate of proportion given by $P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$

Let $\alpha = 0.05$, and it is a two-tailed test, so the decision rule is: Reject H_0 in favour of H_1 if $|Z| > 1.96$

Here, $n_1 = 700$, $P_1 = \frac{200}{700} = 0.2857$ and $n_2 = 500$, $P_2 = \frac{400}{1300} = 0.3077$ and

the pooled estimate of proportion $P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{200 \cdot 400}{700 + 1300} = 0.30$

$$\text{Thus, } z = \frac{0.2857 - 0.3077}{\sqrt{0.30(1 - 0.30) \left\{ \frac{1}{700} + \frac{1}{1300} \right\}}} = -1.023$$

Since the observed absolute value of Z is less than 1.96, there is no evidence to doubt the hypothesis, that means, workers of two factories do not differ significantly with respect to the smoking habit.

16.9. Test of Hypothesis about Correlation Co-efficient

Suppose $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be n pairs of observations of a random sample drawn from a bi-variate normal population with correlation co-efficient ρ , then the sample correlation co-efficient between n pairs of observations is given by

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}}$$

The significance of population correlation co-efficient from which the sample has been drawn may be tested under following two assumptions:

- (i) $H_0 : \rho = 0$, when the population correlation co-efficient is zero (with one-tailed or two-tailed alternative),
- (ii) $H_0 : \rho = \rho_0$, when the population correlation co-efficient is equal to some specified value ρ_0 (with one-tailed or two-tailed alternative)

16.9.1. Testing the hypothesis when the population correlation co-efficient equals zero.

Here the null hypothesis is considered as there is no correlation in the population, that means, the relationship between the variables is not linear.

The test is undertaken considering the following null hypothesis $H_0: \rho = 0$. (in this case usually a two-tailed test $H_1: \rho \neq 0$ is conducted, however, one-tailed test is also conducted when the situation deserves)

Let r be the sample correlation co-efficient (which is the best estimate of population correlation co-efficient ρ). It is to be noted here that, for a sample

of size n , the variance of r is given by $\text{var}(r) = \frac{1-r^2}{n-2}$, then the appropriate test statistic to be used for testing the above mentioned hypothesis is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Which follows t-distribution with $n-2$ degrees of freedom.

Thus, if the computed value of r is greater than the tabulated value of t at 100α % level of significance, then the null hypothesis is rejected that means the decision rule is

$$\text{Reject } H_0 \text{ if } |t| > t_{n-2; \alpha/2} \text{ or } \left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| > t_{n-2; \alpha/2}$$

which indicates that the sample data provide sufficient evidence that $\rho \neq 0$

However, an approximate 'rule of thumb' for considering population correlation co-efficient ρ to be significantly different from zero is given by

$$|r| > \frac{2}{\sqrt{n}}$$

16.9.2. Testing the hypothesis when the population correlation co-efficient equals some specified value ρ_0 .

Let us consider the following null and alternative hypothesis

$$H_0: \rho = \rho_0 \text{ against } H_1: \rho \neq \rho_0$$

The test statistic t used for testing $H_0: \rho = 0$ is appropriate when $\rho = 0$ (under null hypothesis), however, when $\rho \neq 0$, the test statistic is not appropriate. Thus, in testing the above mentioned hypothesis, at first we have to use Fisher's z transformation. Hence, r is to be transformed into :

$$\text{given by } z = \frac{1}{2} \log_e \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+r}{1-r} \text{ (since } \log_e x = 2.3026 \log_{10} x)$$

It has been found that Z is normally distributed with mean

$$z_0 = \frac{1}{2} \log_e \frac{1+p_0}{1-p_0} = 1.1513 \log_{10} \frac{1+p_0}{1-p_0}$$

(under null hypothesis) and standard deviation $\frac{1}{\sqrt{n-3}}$.

Therefore, the test statistic for testing the null hypothesis $H_0: \rho = \rho_0$ is given

$$\text{by } Z = \frac{z - z_0}{\sqrt{\frac{1}{n-3}}} = (z - z_0) \sqrt{n-3}$$

which approximately follows $N(0, 1)$. The approximation is reasonably good if sample size is large.

16.10. Test of Hypothesis about Regression Co-efficient

We know, for n pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ drawn from a bi-variate normal population, the least squares estimate of parameter β , the regression co-efficient of a population regression model, $y = \alpha + \beta x + e$, where $e \sim NID(\mu, \sigma^2)$, is given by

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{sp(x, y)}{ss(x)}$$

It has been found that b is an unbiased estimate of β , that means $E(b) = \beta$,

and variance of b is given by $\text{Var}(b) = \frac{s^2}{\sum (x_i - \bar{x})^2}$, s^2 is unbiased estimate of σ^2

$$s^2 = \frac{\text{SS due to Error}}{n-2} = \frac{\sum (y_i - \hat{y})^2}{n-2} = \frac{\sum (y_i - a - bx)^2}{n-2},$$

s^2 can also be estimated using the relationship

$$s^2 = \frac{\sum (y - \bar{y})^2 - b \sum (y - \bar{y})(x - \bar{x})}{n-2}$$

$$\text{Then, } se(b) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

In order to test the significance of parameter β , the following null and alternative hypothesis are formulated:

$$H_0: \beta = \beta_0 \text{ against } H_1: \beta \neq \beta_0,$$

The test statistic for testing the hypothesis is given by $t = \frac{b - \beta_0}{se(b)} \sim t_{n-2}$

The decision rule is: Reject H_0 at $100\alpha\%$ level of significance, if $|t| > t_{n-2, \alpha/2}$

$$\text{or, if } t = \frac{b - \beta_0}{se(b)} > t_{n-2, \alpha/2} \text{ or } t = \frac{b - \beta_0}{se(b)} < -t_{n-2, \alpha/2}$$

Similarly, one-tailed test of regression parameter can be conducted, if necessary.

Note that very often testing the significance of regression co-efficient implies testing the null hypothesis of zero population regression co-efficient, in that case, it is required to formulate null hypothesis as $H_0 : \beta = 0$, that means, substitute zero in the place of β_0 , all other steps are the same as testing the null hypothesis $H_0 : \beta = \beta_0$.

The decision rule for testing the significance of correlation co-efficient and regression co-efficient using t-test statistic are as follows:

Table 16.14. Decision rule for correlation and regression co-efficient test

Case No.	Types of test		Decision rule Reject H_0 , if
	For correlation co-efficient	For regression co-efficient	
1	$H_1 : \rho \neq 0$	$H_1 : \beta \neq \beta_0$	$ t > t_{\alpha/2; (n-2)}$
2	$H_1 : \rho > 0$	$H_1 : \beta > \beta_0$	$t > t_{\alpha; (n-2)}$
3	$H_1 : \rho < 0$	$H_1 : \beta < \beta_0$	$t < -t_{\alpha; (n-2)}$

Confidence interval (CI) for β . If the regression errors ε 's are normally distributed and all the assumptions of linear regression model hold, then the $100(1-\alpha)\%$ CI for the population regression co-efficient β is given by

$$b - t_{n-2, \alpha/2} se(b) < \beta < b + t_{n-2, \alpha/2} se(b)$$

Example 16.10.1. In a study of relationship between expenditure (x) and sales volume (y), a sample of 10 firms yields the co-efficient of correlation $r = 0.93$. Can it be concluded on the basis of information that x and y are linearly related? (use $\alpha = 0.05$)

Solution. We have to test the hypothesis $\rho = 0$ against $H_0 : \rho \neq 0$.

The test statistic for testing the hypothesis is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Here, $n = 10$, $r = 0.93$, $df = n-2 = 8$

It is a two-tailed test and $\alpha = 0.05$, so from the table of t-distribution, we find the critical value of t at 2.5% level of significance with 8 df is 2.306.

$$\text{Now the value of test statistic is } t = \frac{0.93\sqrt{10-2}}{\sqrt{1-(0.93)^2}} = 7.03$$

Since the computed value of t is much greater than the tabulated value of t, the null hypothesis may be rejected at 5% level of significance. Hence, it may be concluded that x and y are linearly related.

Example 16.10.2. Suppose, in a study of demand and supply of a commodity for 20 months, it has been found that $r = 0.884$. Test whether this correlation is significantly different from a hypothesized value 0.92 at 5% level of significance.

Solution. Here we have to test the following null and alternative hypothesis

$$H_0: \rho = 0.92 \text{ against } H_1: \rho \neq 0.92$$

The test statistic for testing the hypothesis is given by

$$Z = (z - z_0)\sqrt{n-3}$$

$$\text{where, } z = \frac{1}{2} \log_e \frac{1+r}{1-r} = 1.1513 \times \log_{10} \frac{1+r}{1-r} \text{ and } z_0 = 1.1513 \times \log_{10} \frac{1+\rho_0}{1-\rho_0}$$

This is a two-tailed test, so the critical region of Z at 5% level of significance is $|Z| > 1.96$

Here, $r = 0.884$, $\rho_0 = 0.92$ and $n = 20$,

$$z = 1.1513 \times \log_{10} \frac{1+r}{1-r} = 1.1513 \times \log_{10} \frac{1+0.884}{1-0.884} = 1.3938$$

$$\text{and } z_0 = 1.1513 \times \log_{10} \frac{1+\rho_0}{1-\rho_0} = 1.1513 \times \log_{10} \frac{1+0.92}{1-0.92} = 1.5890$$

$$\text{Thus, } z = (z - z_0)\sqrt{n-3} = -0.80$$

Since this value of Z does not fall in the critical region, we fail to reject null hypothesis at 5% level of significance, that means, the population correlation co-efficient is not significantly different from the hypothesized value.

Example 16.10.3. A research team was attempting to determine if political risk in countries is related to inflation for these countries. For this purpose a survey of political risk analysis was conducted with the mean political risk score for each of 49 countries. The political risk score is scaled such that the higher the score, the greater the political risk. The sample correlation co-

efficient between political risk and inflation for these countries was found as 0.43. Test whether there is a positive linear relationship between the political risk and inflation at 0.5% level of significance.

Solution. We want to test $H_0: \rho = 0$ against $H_1: \rho > 0$

We have, $n = 45$, $r = 0.43$, so $df = 49 - 2 = 47$

The test is based on the statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$

This is a right tailed test, so the critical region is given by $t > t_{0.005, 47}$, or, $t > 2.704$

The computed value of test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.43\sqrt{49-2}}{\sqrt{1-(0.43)^2}} = 3.265$

Since the observed value of t lies in critical region, we can reject null hypothesis at 0.5% level of significance.

We have strong evidence of a positive linear relationship between inflation and political risk of countries.

Example 16.10.4. A regression of retail sales on disposable income provides with the following results. $n = 22$, $b = 0.3815$, $se(b) = 0.0253$. Test the significance of regression co-efficient at 1% level of significance. Also, compute 99% confidence interval for regression parameter.

Solution. We have to test the hypothesis $H_0 : \beta = \beta_0 = 0$ against $H_1 : \beta \neq 0$. Assume that the regression errors ϵ 's are normally distributed and all the assumptions of linear regression model hold, then the test statistic for testing the above mentioned hypothesis is given by

$$t = \frac{b}{se(b)} \text{ which distributed as Student's } t \text{ with } n - 2 \text{ df.}$$

From the given information, we have, the computed value of t as

$$t = \frac{b}{se(b)} = (0.3815 - 0)/0.0253 = 15.08$$

Now, from table of t -distribution, we have for $(n - 2) = 20$ df, $t_{20, 0.005} = 2.845$, hence, the null hypothesis is very clearly rejected, that means, the regression coefficient is highly significant.

Again, the 99% CI is given by $b \pm t_{n-2, \alpha/2} \times se(b) = 0.3815 \pm 2.845 \times 0.0253 = (0.31, 0.45)$.

Example 16.10.5. A Courier Express authority recorded the following data regarding distance (x) and the time usage (y) for hand delivery of 15 packets and obtained the following results:

$$\Sigma x = 243, \Sigma y = 503, \Sigma x^2 = 3999, \Sigma y^2 = 17181, \Sigma xy = 8229.$$

- (i) Compute the correlation between distance and time usage and test whether it is significantly different from zero against an alternative that there is a positive linear relationship between these two variables at 1% level of significance.
- (ii) Compute the regression co-efficient and test its significance.
- (iii) Also obtain 99% confidence interval for regression parameter.

Solution. (i) We know the correlation co-efficient is given by

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}}$$

Substituting the values from given information, we have,

$$\sum xy - \frac{\sum x \sum y}{n} = 8229 - (243 \times 503)/15 = 80.4$$

$$\sum x^2 - \frac{(\sum x)^2}{n} = 3999 - (243)^2/15 = 62.4$$

$$\sum y^2 - \frac{(\sum y)^2}{n} = 17181 - (503)^2/15 = 313.73$$

$$\text{Thus, } r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}} = \frac{80.4}{\sqrt{62.4 \times 313.73}} = 0.57$$

In order to perform the required test, we have to formulate the following null and alternative hypotheses

$$H_0: \rho = 0 \text{ against } H_1: \rho > 0$$

The appropriate test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Here the degrees of freedom is $n - 2 = 15 - 2 = 13$.

This is a right tailed test, so the critical region is given by $t > t_{0.01, 13}$ or $t > 2.650$ (From table of t-distribution).

The computed value of test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.57\sqrt{15-2}}{\sqrt{1-(0.57)^2}} = 2.50$

Thus, the observed value of t does not fall in the critical region, so we fail to reject null hypothesis at 1% level of significance.

Hence, it can be concluded that the population correlation co-efficient is not significantly different from zero, that means, there is no linear relationship between the variables.

(ii) We know, for the regression line $y = a + \beta x + e$, the least squares estimate of β and a are given by

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

Using the values of sum of the products and sum of squares of x from the calculation of correlation co-efficient, we have, $b = \frac{80.4}{62.4} = 1.29$ and

$$a = \bar{y} - b\bar{x} = \frac{503}{15} - 1.29 \times \frac{243}{15} = 35.53 - 1.29 \times 16.20 = 13.64$$

The fitted regression line is $y = 13.64 + 1.29x$

In order to test the significance of regression co-efficient, we have to consider the following null and alternative hypothesis:

$$H_0: \beta = 0 \text{ against a two-tailed alternative } H_1: \beta \neq 0.$$

The appropriate test statistic is $t = \frac{b - \beta_0}{se(b)} \sim t_{n-2}$

Here, degrees of freedom is also 13 and since it is a two-tailed test, at 1% level of significance the critical region is given by $|t| > t_{0.005;13}$ or $|t| > 3.012$

$$\text{Here, } s^2 = \frac{\sum (y - \bar{y})^2 - b \sum (y - \bar{y})(x - \bar{x})}{n-2} = \frac{313.73 - 1.29 \times 80.4}{13} = 16.15$$

and $s = 4.02$

$$se(b) = \frac{s}{\sqrt{\sum (x - \bar{x})^2}} = \frac{4.02}{\sqrt{62.4}} = 0.5089$$

Thus, under null hypothesis, the value of the test statistic is

$$t = \frac{b}{se(b)} = \frac{1.29}{0.5089} = 2.54$$

The computed value of t does not fall in the critical region, so we fail to reject null hypothesis at 1% level of significance.

Thus, we may conclude that population regression co-efficient is not significantly different from zero.

(iii) 99% confidence interval for β is given by

$$\begin{aligned} b - t_{n-2, \alpha/2} se(b) < \beta < b + t_{n-2, \alpha/2} se(b) \\ = 1.29 - 3.012 \times 0.6103 < \beta < 1.29 + 3.012 \times 0.6103 = (-0.548, 3.128) \end{aligned}$$

Example 16.10.6. The fitted regression model for profit (y) on costs (x) for 11 items produced by a company has been found as (the variables are measured in hundred Taka):

$$\begin{array}{lll} y = & 1922.40 & + 1.2865 x \\ se = & (274.9) & (0.0253) \end{array}$$

(where the values in the parenthesis represent the standard error of respective statistic)

The company claims that for every thousand of increase of the cost of their products, they incur an average profit of Taka 1.2 thousand. In the light of the claim of company, test the significance of population regression co-efficient, against a two-tailed alternative at 5% level of significance, find p value of the test and 95% confidence interval for population regression co-efficient.

Solution. For testing the significance of regression co-efficient, we have to formulate the null and alternative hypothesis as follows:

$$H_0: \beta = 1.2 \text{ against a two-tailed alternative } H_1: \beta \neq 1.2,$$

The appropriate test statistic is $t = \frac{b - \beta_0}{se(b)} \sim t_{n-2}$

Here, $n = 11$, so degrees of freedom is $11-2 = 9$ and since it is a two-tailed test, at 5% level of significance the critical region is $|t| > t_{0.025, 9}$ or $|t| > 2.262$

From the output of results we have, $b = 1.2865$, $se(b) = 0.0253$, $\beta_0 = 1.2$

$$\text{So, under } H_0 \text{ the value of } t = \frac{b - \beta_0}{se(b)} = \frac{1.2865 - 1.2}{0.0253} = 3.42$$

The computed value of t falls in the critical region, so we fail to accept null hypothesis at 5% level of significance. That means, the regression co-efficient is significant.

Thus, we may conclude that population regression co-efficient is significantly different from 1.2

p-value. From the table of area under t-distribution, we have, for degrees of freedom 9, $P(t < -3.250) = 0.005$, so the $P(t > 3.250) = 0.005$, so the value of $p = 2 \times 0.005 = 0.01$

Again, 95% confidence interval for β is given by

$$\begin{aligned} b - t_{n-2, \alpha/2} \times se(b) &< \beta < b + t_{n-2, \alpha/2} \times se(b) \\ = 1.2865 - 2.262 \times 0.0253 &< \beta < 1.2865 + 2.262 \times 0.0253 = (1.2292, 1.3437) \end{aligned}$$

Example 16.10.7. Suppose the fitted linear regression model for production (y) of certain commodity on electricity consumption (x) for 20 years has been found as :

$$\begin{array}{lll} y & = & 42.34 + 0.55x \\ se & = & (2.88) \quad (0.0197) \end{array}$$

Test the significance of regression co-efficient at 5% level of significance, find 95% CI for β .

Solution. For testing the significance of regression co-efficient, let us formulate the null and alternative hypothesis as follows:

$$H_0: \beta = 0 \text{ against a two-tailed alternative } H_1: \beta \neq 0,$$

$$\text{The appropriate test is by } t = \frac{b - \beta_0}{se(b)} \sim t_{n-2}$$

Here, $n = 20$, so degrees of freedom is $20-2 = 18$ and since it is a two-tailed test, at 5% level of significance the critical region is given by $|t| > t_{0.025; 18}$ or $|t| > 2.101$

From the output of results we have, $b = 0.55$, $se(b) = 0.0197$, $\beta_0 = 0$

$$\text{So under null hypothesis, the value of } t = \frac{b}{se(b)} = \frac{0.55}{0.0197} = 27.92$$

The computed value of t falls in the critical region and far away from the critical value, so we fail to accept null hypothesis at 5% level of significance. That means, the regression co-efficient is very significant.

95% confidence interval for β is given by

$$\begin{aligned} b - t_{n-2, \alpha/2} \times se(b) &< \beta < b + t_{n-2, \alpha/2} \times se(b) \\ = 0.55 - 2.101 \times 0.0197 &< \beta < 0.55 + 2.101 \times 0.0197 = (0.5086, 0.5914) \end{aligned}$$

Example 16.10.8. Twelve secretaries, already working for different periods, at an office were asked to take a special three-day intensive course to improve their keyboard skill. At the beginning and again at the end of the course, they were given a particular two-page letter to type and the improved flawless number of words typed are recorded. The recorded data are shown in the following table.

Table 16.15. Experience and improvement in typing speed.

Secretary	Number of Years of Experience	Improvement (words per minute)
A	2	9
B	6	11
C	3	8
D	8	12
E	10	14
F	5	9
G	10	14
H	11	13
I	12	14
J	9	10
K	8	9
L	10	10

- i. Compute the product moment correlation co-efficient and test its significant against a positive alternative.
- ii. Fit a regression line of improvement on experience and test the significance of regression co-efficient
- iii. Find 95% confidence interval for population regression co-efficient

Solution. (i) We know, the product moment correlation co-efficient is given by

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}}$$

The statistics needed for the calculation of correlation co-efficient are shown in following table.

Table 15.16. Calculating statistics for correlation and regression co-efficient

Secretary	x	y	x^2	y^2	xy
A	2	9	4	81	18
B	6	11	36	121	66
C	3	8	9	64	24
D	8	12	64	144	96
E	10	14	100	196	140
F	5	9	25	81	45
G	10	14	100	196	140
H	11	13	121	169	143
I	12	14	144	196	168
J	9	10	81	100	90
K	8	9	64	81	72
L	10	10	100	100	100
Total	$\Sigma x = 94$	$\Sigma y = 133$	$\Sigma x^2 = 848$	$\Sigma y^2 = 1529$	$\Sigma xy = 1102$

We have

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}} = \frac{1102 - 94 \times 133 / 12}{\sqrt{848 - \frac{(94)^2}{12}} \sqrt{1529 - \frac{(133)^2}{12}}} = 0.89$$

For testing the significance of correlation co-efficient, we formulate the null and alternative hypothesis $H_0: \rho = 0$ against $H_1: \rho > 0$

The appropriate test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Here the degrees of freedom of t is $n-2 = 12-2 = 10$

This is a right-tailed test, so at 5% level of significance, the critical region is given by $t > t_{0.05;10}$ or $t > 1.812$ (From table of t-distribution).

Under null hypothesis, the value of test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.89\sqrt{12-2}}{\sqrt{1-(0.89)^2}} = 6.25$$

Since, the observed value of t falls in the critical region, so we fail to accept null hypothesis at 5% level of significance. That means, there is a significant positive correlation between experience and improvement of typing speed.

(ii) The simple linear regression line to be fitted is $y = a + bx$

$$\text{Where, } b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\}} \quad \text{and } a = \bar{y} - b \bar{x}$$

$$\text{Thus, } b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\}} = \frac{1102 - 94 \times 133 / 12}{848 - (94)^2 / 12} = 0.54 \text{ approx.}$$

$$\text{and } a = \bar{y} - b \bar{x} = 133 / 12 - 0.54 \times 94 / 12 = 6.85 \text{ approx.}$$

Substituting the computed values of a and b leads to the following fitted regression equation $\hat{y} = 6.85 + 0.54x$

For testing the significance of regression co-efficient, let us consider $H_0: \beta = 0$ against a two-tailed alternative $H_1: \beta \neq 0$,

The appropriate test statistic is $t = \frac{b - \beta_0}{se(b)} \sim t_{n-2}$

$$\text{Here, } s^2 = \frac{\sum (y - \bar{y})^2 - b \sum (y - \bar{y})(x - \bar{x})}{n - 2} = 2.2427 \text{ and } s = 1.4976$$

$$\text{And } se(b) = \frac{s}{\sqrt{\sum (x - \bar{x})^2}} = \frac{1.4976}{\sqrt{60.3}} = 0.1928$$

$$\text{Thus, } t = \frac{b - \beta_0}{se(b)} = \frac{0.54 - 0}{0.1928} = 2.280$$

Here, at 5% level of significance with 10 degrees of freedom, the critical values are $|t| > t_{0.025, 10}$ or $|t| > 2.228$ (from table). The computed value of t falls in the critical region, so we fail to accept null hypothesis at 5% level of significance.

Thus, we may conclude that population regression co-efficient is significantly different from zero.

(iii) 95% confidence interval for β is given by

$$\begin{aligned} b - t_{n-2, \alpha/2} \times se(b) &< \beta < b + t_{n-2, \alpha/2} se(b) \\ &= 0.54 - 2.228 \times 0.1928 < \beta < 0.54 + 2.228 \times 0.1928 = (0.1104, 0.9695) \end{aligned}$$

16.11. Test of Significance of Single Variance (χ^2 Parametric Test)

Suppose, X_1, X_2, \dots, X_n be a random sample of size n from normal population with mean μ and variance σ^2 , where μ is unknown. We want to test the hypothesis that the population variance is σ_0^2 , a specified value of σ^2 , against possible three types of alternatives, that means,

- i) $H_0 : \sigma^2 = \sigma_0^2$ against $H_A : \sigma^2 \neq \sigma_0^2$ (against a two-tailed alternative)
- ii) $H_0 : \sigma^2 = \sigma_0^2$ against $H_A : \sigma^2 > \sigma_0^2$ (against a right-tailed alternative)
- iii) $H_0 : \sigma^2 = \sigma_0^2$ against $H_A : \sigma^2 < \sigma_0^2$ (against a left-tailed alternative)

Since μ is unknown, it is estimated by its unbiased estimate $\bar{X} = \frac{\sum X_i}{n}$ and variance σ^2 estimated by its unbiased estimate

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{1}{n-1} \left[\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right]$$

Then, under null hypothesis, the test statistic is $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$

Decision. In case of two tailed alternative, reject H_0 at α level of significance

$$\text{If } \chi^2_{\text{cal}} \leq \chi^2_{1-\alpha/2, n-1} \text{ or, } \chi^2_{\text{cal}} \geq \chi^2_{\alpha/2, n-1}$$

In case of a right-tailed alternative, reject H_0 at α level of significance

$$\text{if } \chi^2_{\text{cal}} \geq \chi^2_{\alpha, n-1}$$

and in case of a left-tailed alternative, reject H_0 at α level of significance

$$\text{if } \chi^2_{\text{cal}} \leq \chi^2_{1-\alpha, n-1}$$

Example 16.11.1. The following are the weights (in gram) of a randomly selected sample of 11 apples in a shop.

70, 85, 92, 90, 95, 79, 80, 85, 90, 85, 95

The weight of apples follows normal distribution with mean μ and variance σ^2 . Can we conclude that the population variance of apples of the shop is more than 50 gm^2 ?

Solution. The null and alternative hypothesis for this test are

$$H_0 : \sigma^2 = 50 \text{ against } H_A : \sigma^2 > 50$$

Under null hypothesis, the test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \text{ which is distributed as } \chi^2 \text{ with } n-1 \text{ df}$$

It is a right tailed test, so, the critical region is given by $\chi^2 \geq \chi^2_{\alpha, n-1}$

Here, $\bar{X} = \frac{\sum X_i}{n} = \frac{946}{11} = 86$ and

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{1}{n-1} \left[\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right]$$

$$= \frac{1}{10} \left[81930 - \frac{(946)^2}{11} \right] = 57.4$$

So, the computed value of χ^2 is $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{10 \times 57.4}{50} = 11.48$

Here, at 5% level of significance, the critical value is given by $\chi^2_{0.05, 10} = 18.307$ which is more than the computed value, so we fail to reject H_0 , that means, the variance of the weights of the apples is not more than 50 gm^2 .

Example 16.11.2. The daily duration of telephone calls received by the enquiry department of a small industry for a randomly selected 11 days over a quarter are as follows:

160, 172, 121, 144, 100, 108, 175, 200, 105, 95, 102

The manager of industry says that the population variance of the daily duration of calls over the quarter is 1500. The authority thinks that it is overestimated. How would you comment on the variance?

Solution. The null and alternative hypothesis for this test are

$$H_0 : \sigma^2 = 50 \text{ against } H_A : \sigma^2 < 1500$$

Under null hypothesis, the test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \text{ which is distributed as } \chi^2 \text{ with } n-1 \text{ df} = 10 \text{ df}$$

It is right tailed test, so, the critical region is given by $\chi^2 \leq \chi^2_{1-\alpha, n-1}$ where at 5% level of significance, $\chi^2_{0.95, 10} = 3.94$.

Here, $\sum X_i = 1482$ and

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{1}{n-1} \left[\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right]$$

$$= \frac{1}{10} \left[213340 - \frac{(1482)^2}{11} \right] = 1363.82$$

So, the computed value of χ^2 is $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{10 \times 1363.82}{1500} = 9.09$

Here, at 5% level of significance, the critical value is given by $\chi^2_{0.05, 10} = 3.94$ which is less than the computed value, so we reject H_0 , that means, the variance of daily duration of telephone calls is less than 1500. Hence, the authority's thought is beyond doubt.

16.12. Test of Hypothesis about Independence of Two Attributes (χ^2 Non-parametric Test)

Let a sample of size n be drawn from the population of interest and the observed frequencies are cross-classified according to the categories of two attributes. Let us consider two attributes A and B where A is assumed to have r categories and B is assumed to have c categories. The cross-classification can be conveniently displayed by means of a table called $r \times c$ contingency table (read r by c contingency table). A general contingency table of order $r \times c$ is shown in table 16.17. The frequencies O_{ij} in the cells are termed as observed frequencies and the totals of the frequencies in each row (say, R_i , $i = 1, 2, \dots, r$) and each column (say, C_j , $j = 1, 2, \dots, c$) are termed as marginal frequencies.

Contingency Table. Data for attributes arranged in two-directional tabular form for the test of independence is called a contingency table. The order of the table is determined by the number of categories of two attributes.

Table 16.17. $r \times c$ Contingency table

		Attribute B									Total
Attribute A		B ₁	B ₂	B _j	B _c		
	A ₁	O ₁₁	O ₁₂	O _{1j}	O _{1c}	R ₁	
	A ₂	O ₂₁	O ₂₂	O _{2j}	O _{2c}	R ₂	
	
	
	A _i	O _{i1}	O _{i2}	O _{ij}	O _{ic}	R _i	
	
	
	A _r	O _{r1}	O _{r2}	O _{rj}	O _{rc}	R _r	
	Total	C ₁	C ₂	C _j	C _c	N	

For testing the independence of two attributes A and B, cross classified in different categories, the null hypothesis is considered as

H_0 : Two attributes A and B are independent
 against the alternative H_1 : Two attributes A and B are associated or dependent.

Under null hypothesis, the test statistic for testing the hypothesis is given by

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

where, O_{ij} and E_{ij} are the observed and expected frequencies corresponding to the (i, j)th cell of contingency table. Expected cell frequencies are computed according to the multiplicative rule of probability. If two events are independent, the probability of their joint occurrence is equal to the product of their individual probabilities. Applying this rule to a contingency table, it is equivalent to say that, if two criteria of classification are independent, a joint probability is equal to the product of the two corresponding marginal probabilities. Thus, the expected cell frequencies are computed by the formula:

$$E_{ij} = \frac{R_i}{N} \times \frac{C_j}{N} \times N = \frac{R_i \times C_j}{N}$$

Note that total expected frequencies should be equal to the total observed frequency, so while computing the expected frequencies it is required to adjust the expected frequencies for $rc - (r - 1) \times (c - 1)$ cells, that is why for this test the degree of freedom is $(r - 1) \times (c - 1)$ that means for $(r - 1) \times (c - 1)$ cells expected frequencies are calculated independently.

It is one-tailed test (right-tailed) test. So, if the calculated value of χ^2 is less than the table value of χ^2 at a specific level of significance with $(r - 1) \times (c - 1)$ degrees of freedom, the null hypothesis holds true, that means, the two attributes are independent. If calculated value of χ^2 is greater than the table value, the null hypothesis is rejected, that means the two attributes are associated or dependent.

Observed Frequency. The frequencies obtained by observation. These are the sample frequencies.

Expected Frequency. The frequency corresponding to a particular cell obtained by dividing the product of row total and column total passing through that cell by the total frequencies.

Example 16.12.1. A tobacco company claims that there is no relationship between smoking and lung ailments. To investigate the claims a random sample of 300 males in the age group 40-50 are given medical test. The observed sample results are shown below:

Table 16.18. Number of males according to smoking and lung ailment.

	Found lung ailment	No lung ailment	Total
Smokers	75	105	180
Non-smokers	25	95	120
Total	100	200	300

On the basis of the information, can it be concluded that smoking and lung ailments are independent?

Solution. Let us consider the hypothesis that smoking and lung ailments are independent, that means,

H_0 : Smoking and lung ailments are independent
or there is no effect of smoking on lung ailments

Against, H_1 : They are not independent, or smoking causes lung ailments.

The test statistic for testing the hypothesis is $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E} \sim \chi^2_{(r-1)(c-1)}$

It is a 2×2 table, so the degree of freedom is 1, we have to calculate expected frequency for a cell independently, and others are to be obtained by adjustment so that the total marginal frequencies remain the same.

Again, let $\alpha = 0.05$, then the critical value of χ^2 with 1 df is 3.84 (which is also sometimes written as $\chi^2_{0.05; 1} = 3.84$).

Expected frequency for the cell corresponding to first row and first column

$$(E_{11}) \text{ is computed as } E_{11} = \frac{R_1 \times C_1}{N} = \frac{180 \times 100}{300} = 60$$

So the expected frequencies for the remaining cells are

$$\begin{aligned} E_{12} &= R_1 - E_{11} = 180 - 60 = 120, \quad E_{21} = C_1 - E_{11} = 100 - 60 = 40, \\ \text{and } E_{22} &= C_2 - E_{12} = 200 - 120 = 80, \quad E_{22} \text{ can also be computed using } R_2 \text{ as} \\ E_{22} &= R_2 - E_{21} = 120 - 40 = 80 \end{aligned}$$

Arranging the observed and expected frequencies in the following table, we can easily compute the necessary columns for χ^2 .

Table 16.19. Computation of χ^2 .

O	E	$(O - E)^2$	$(O - E)^2/E$
75	60	225	3.75
105	120	225	1.875
25	40	225	5.625
95	80	225	2.8125
		Total	14.0625

Here the observed value of chi-squares is much more than the critical values, therefore, the null hypothesis is rejected at 5% level of significance. That means, it is evident that smoking has significant effect on lung ailments.

Example 16.12.2. A certain drug is claimed to be effective in curing cold. In an experiment on 500 persons suffering from cold, half of them were given the drug and half of them were given the sugar pills. The reaction to the treatment on patients are recorded as in the following table:

Table 16.20. Reaction to the treatment

	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 information can it be concluded that there is a significance difference in the effect of the drug and sugar pills.

Solution. Let us take the null hypothesis that there is no difference in the drug and sugar pills as far as their effect on curing cold is concerned.

$$\text{The test statistic is } \chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Since it is a 2×3 table, the degrees of freedom would be $(2 - 1) \times (3 - 1) = 2$, that means, we will have to calculate only two expected frequencies independently and other four can be calculated or determined by adjustment.

Expected frequencies are computed as follows:

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{250 \times 280}{500} = 140, \quad E_{12} = \frac{R_1 \times C_2}{N} = \frac{250 \times 70}{500} = 35$$

Thus the table for the computation of expected frequencies is shown below:

Table 16.21. Computation of expected frequencies.

	Helped	Harmed	No effect	Total
Drug	140	35	75	250
Sugar pills	140	35	75	250
Total	280	70	150	500

Arranging the observed and expected frequencies in the following table, we can easily compute the necessary columns for χ^2 .

Table 16.22. Computation of χ^2 .

O	E	$(O - E)^2$	$(O - E)^2/E$
150	140	100	0.714
130	140	100	0.714
30	35	25	0.714
40	35	25	0.714
70	75	25	0.333
80	75	25	0.333
		Total	3.522

The calculated value of χ^2 is 3.522.

The critical value of χ^2 for 2 df at 5% level of significance is 5.99 which is larger than the calculated value. Therefore we fail to reject the null hypothesis, that means we can conclude that the drug and sugar pills do not make any significant difference in curing cold.

Example 16.12.3. In a survey on the daily production of a garments factory, the manager finds that 315 of the garments are of category A, 101 are of category B, 108 are of category C and 32 are of category D. According to the manager's long experience, he expects that the numbers of different categories of garments be in proportion 9 : 3 : 3 : 1. From the given information, can we say at 5% level of significance that the manager is correct?

Solution. Here, the null hypothesis is

H_0 : There is no significant difference between the observed and expected number of garments of different categories.

The appropriate test statistic for testing the hypothesis is

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

The expected frequencies in this case are to be calculated as proportional to the expected ratio. The total number of garments observed = 315 + 101 + 108 + 32 = 556 and the sum of ratios = 9 + 3 + 3 + 1 = 16.

So, the expected number of garments different categories are as follows:

$$E_A = \frac{556 \times 9}{16} = 312.75, E_B = E_C = \frac{556 \times 3}{16} = 104.25,$$

$$\text{and, } E_D = 556 - (E_A + E_B + E_C) = 34.75$$

Computation of the value of χ^2 is shown in following table.

Table 16.23. Calculation of χ^2 .

Category	Observed number O	Expected number E	$(O - E)^2$	$(O - E)^2/E$
A	315	312.75	5.062	0.016
B	101	104.25	10.562	0.101
C	108	104.25	14.062	0.135
D	32	34.75	7.562	0.218
			Total	0.470

The computed value of χ^2 is 0.470

Here, the degrees of freedom is $4 - 1 = 3$, and the critical value of χ^2 at 5% level of significance with 3 degrees freedom is 7.82 (from table).

Hence, we fail to reject null hypothesis at 5% level of significance, hence, it can be concluded that there is no doubt about the manager's expectation.

16.13. Power of a Test

We know, the power of a test is the probability of rejecting null hypothesis when it is false. Let us consider a sample of size n from normal population with mean μ and known variance σ^2 and suppose we are interested to test the hypothesis

$H_0: \mu = \mu_0$, against the alternative $H_1: \mu > \mu_0$ at α level of significance

We know, the test statistic to be used for testing the null hypothesis is given by

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

and the decision rule is that Reject

$$H_0: \text{if } Z > z_\alpha \text{ or } z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \text{ or, } \bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n}$$

Here, $\bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n}$ determine the values of sample mean that leads in rejecting the null hypothesis, which otherwise proves that the null hypothesis is false.

Steps in determination of power of a population mean test. The following steps are involved in computation of power of a test

- determine the value of $\bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n}$

(ii) find the probability that the sample mean will be in the acceptance region for given false null hypothesis, which will give the probability of Type II error, β . At this step, a value of $\mu = \mu^*$ is considered such that $\mu^* > \mu_0$ (for a right tailed test) and β is computed as

$$\beta = P(\bar{x} \leq \bar{x}_c \mid \mu = \mu^*) = P\left(z < \frac{\bar{x}_c - \mu}{\sigma/\sqrt{n}}\right)$$

(iii) compute power of the test as : Power = $1 - \beta$

Power for the tests of other parameters can also be derived using the steps same as population mean.

The value of β and power of the test will be different for different values of μ^* . If powers of a test are plotted against different values of μ^* and a smooth curve is drawn, the curve which will be obtained is known as power curve.

Example 16.13.1. (power for mean test). Suppose a ball bearing company claims that the average weight of its product is 5 ounces. The population distribution of weights is assumed to be normally distributed with standard deviation 0.1 ounce. An interested person drew a random sample of 16 observations to test whether the average weight is more than 5 ounces at 5% level of significance. Find power of this test.

Solution. Here, $H_0: \mu = 5$, against the alternative $H_1: \mu > 5$,

The value of the test statistic for testing the hypothesis is given by

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

At $\alpha = 0.05$, $z_\alpha = 1.645$, hence the decision rule is

Reject H_0 if $\frac{\bar{x} - 5}{0.1/\sqrt{16}} > 1.645$ or, $\bar{x} > 5 + 1.645 \times 0.1 \times 4 = 5.041 = \bar{x}_c$

Now if the sample mean is less than or equal to 5.041, then according to the rule, we will fail to reject null hypothesis.

Suppose, we want to determine the probability that the null hypothesis will not be rejected if the true mean is greater than 5.041 ounces, say 5.05 ounces. Thus the null hypothesis is wrong and the alternative is correct.

At this step, we want to determine the probability that we will fail to reject null hypothesis if $\mu = 5.05$, which is the probability of an incorrect decision and we will obtain β , as

$$\begin{aligned}\beta &= P(X \leq \bar{x}_c \mid \mu = 5.05) = P(Z \leq \frac{5.041 - 5.05}{0.1/\sqrt{16}}) \\ &= P(Z \leq -0.36) = 1 - 0.6406 = 0.3594\end{aligned}$$

This means, when the population mean is 5.05, the value of β is 0.3594.

Finally, power is computed using the relationship

$$\text{power} = 1 - \beta = 1 - 0.3594 = 0.6406$$

In the same way different values of β and power can be generated.

Example 16.13.2. (power for proportion test). Suppose the authority of a big firm claims that they follow the gender equality convention of UN in the employment of workers in their firm. A random sample of 600 workers was obtained from the firm, and found that number of female workers were 382. Using a significance level of $\alpha = 0.05$, test the claim of authority and find the power of test.

Solution. The null hypothesis and alternative hypothesis are respectively $H_0 : \pi = \pi_0 = 0.50$, against the alternative $H_1 : \pi \neq 0.50 = \pi_1$ (where π_1 is any value of π other than π_0)

The appropriate test statistic for testing the hypothesis is given by

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$$

Where P is the estimated value of proportion from sample.

The decision rule is

$$\text{Reject } H_0 \text{ if } \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} > z_{\alpha/2} \text{ or, } \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} < -z_{\alpha/2}$$

Here, $\alpha = 0.05$, $n = 600$, substituting the values, we have

$$\frac{P - 0.50}{\sqrt{\frac{0.50(1 - 0.50)}{600}}} > 1.96 \text{ or, } \frac{P - 0.50}{\sqrt{\frac{0.50(1 - 0.50)}{600}}} < -1.96$$

Hence, the decision rules becomes

$$\text{reject } H_0, \text{ if } P > 0.50 + 1.96 \times \sqrt{\frac{0.50(1 - 0.50)}{600}} = 0.50 + 0.04 = 0.54, \text{ or}$$

$$P < 0.50 - 1.96 \times \sqrt{\frac{0.50(1 - 0.50)}{600}} = 0.50 - 0.04 = 0.46$$

The observed value of π is $P = 382/600 = 0.637$, which is greater than 0.54, the upper critical value, so we fail to accept null hypothesis. That means, the authority's claim is not justified.

Since the test proves that the null hypothesis is not true, let us suppose that the true value is $\pi_1 = 0.55$, thus for the probability of Type II error, β , we have to find the probability that the sample proportion is between 0.46 and 0.54, if the population proportion is 0.55. This probability is given by

$$\beta = P(0.46 \leq P \leq 0.54 | P = 0.55)$$

$$= P\left[\frac{0.46 - \pi_1}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}} \leq \frac{P - \pi_1}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}} \leq \frac{0.54 - \pi_1}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}}\right]$$

$$= P\left[\frac{0.46 - 0.55}{\sqrt{\frac{0.55(1-0.55)}{600}}} \leq Z \leq \frac{0.54 - 0.55}{\sqrt{\frac{0.55(1-0.55)}{600}}}\right]$$

$$= P(-1.43 \leq Z \leq -0.49) = 0.3121$$

$$\text{Hence, Power} = 1 - \beta = 1 - 0.3121 = 0.6879$$

This probability β or power of the test can be calculated for any proportion P not equal to 0.50.

Questions

1. What do you mean by a statistical hypothesis? What is meant by the test of a statistical hypothesis?
2. Differentiate the following pairs of concepts:
 - i) null hypothesis and alternative hypothesis
 - ii) simple and composite hypothesis
 - iii) type I error and type II error
 - iv) level of significance (or size of a test) and power of a test
 - v) one-tailed test and two-tailed test
 - vi) critical region and acceptance region
 - vii) level of significance and p-value
 - viii) parameter and statistic
 - ix) Bi-variate table and contingency table.

3. Define student's t-test. Also define degrees of freedom. How does student's t-test differ from z-test? Cite two important applications of each of z-test and student's t-test.
4. What is paired t-test? State two situations when this test is applicable.
5. State the assumptions made in testing a hypothesis. Elucidate various steps involved in testing a hypothesis.
6. Define test statistic. State the names of all test statistics known to you and explain any one of them with applications.
7. Define χ^2 statistic. Explain the procedure of testing the independence of two attributes.
8. Under what conditions is it appropriate to use a one-tailed test? A two-tailed test?
9. If you have decided that a one-tailed test is the appropriate test to use, how do you decide whether it should a lower-tailed or an upper-tailed test?
10. What do you mean by confidence interval? Construct $100(1-\alpha)\%$ confidence interval for a population mean with necessary assumptions and interpret.
11. Define a contingency table. Mention its difference from a bivariate table. What do you mean by the degrees of freedom of a contingency table? How can you test the independence of two attributes with number of categories 3 and 4 respectively.
12. Define Type II error and power of a test. Illustrate how would you find the power of the following tests (i) $H_0 : \mu = \mu_0$, against $H_1 : \mu < \mu_0$, (ii) $H_0 : P = P_0$, against $H_1 : P > P_0$
13. Stating the underlying assumptions, explain how can you carry out the following tests:
 - i) significance of a population mean against a) a one-tailed and b) two-tailed alternative
 - ii) significance of the difference between two population means for two independent populations under different situations for a) a one-tailed and b) two-tailed alternative
 - iii) significance of a population proportion for a) a one-tailed and b) two-tailed alternative
 - iv) significance of the difference between two population proportions for a) a one-tailed and b) two-tailed alternative
 - v) significance of a population correlation co-efficient when it is a) specified b) not specified
 - vi) significance of a population regression co-efficient when it is a) specified b) not specified
 - vii) using a) a one-tailed and b) a two-tailed alternative
 - viii) independence of two attributes.

Exercises

Stating necessary assumption, if any, perform the tests of following hypotheses

1. $H_0: \mu = 21$ against $H_1: \mu \neq 21$, given, $n = 20$, $\bar{x} = 21.2$, $\sigma = 1.5$, $\alpha = 0.05$
(Ans. $z = 0.596$)
2. $H_0: \mu = 100$ against $H_1: \mu < 100$, given, $n = 36$, $\bar{x} = 98.5$, $s = 5.0$, $\alpha = 0.01$
(Ans. $z = -1.8$)
3. $H_0: \mu = 5$ against $H_1: \mu \neq 5$, given, $n = 25$, $\bar{x} = 6.1$, $\sigma = 3.0$, $\alpha = 0.05$
(Ans. $z = 1.83$)
4. $H_0: \mu = 15$ against $H_1: \mu > 15$, given, $n = 40$, $\bar{x} = 16.5$, $\sigma = 3.5$, $\alpha = 0.01$
(Ans. $z = 2.71$)
5. $H_0: \mu = 50$ against $H_1: \mu \neq 50$, given, $n = 60$, $\bar{x} = 48.9$, $\sigma = 4.0$, $\alpha = 0.01$
(Ans. $z = -2.13$)
6. $H_0: \mu = 5$ against $H_1: \mu \neq 5$, given, $n = 10$, $\bar{x} = 7.2$, $s = 4.89$, $\alpha = 0.05$
(Ans. $t = 1.42$)
7. $H_0: \mu = 10000$ against $H_1: \mu < 10000$, given, $n = 100$, $\bar{x} = 12480.0$, $s = 2047.98$, $\alpha = 0.05$
(Ans. $z = 12.11$)
8. $H_0: \mu = 25$ against $H_1: \mu > 25$, given, $n = 12$, $\bar{x} = 27.2$, $s = 3.89$, $\alpha = 0.05$
(Ans. $t = 2.23$)
9. $H_0: \mu = 60$ against $H_1: \mu \neq 60$, given, $n = 10$, $\bar{x} = 58.9$, $\sigma = 3.5$, $\alpha = 0.01$
(Ans. $z = -0.99$)
10. $H_0: \mu = 147$ against $H_1: \mu < 147$, given, $n = 15$, $\bar{x} = 138.9$, $s = 13.0$, $\alpha = 0.05$
(Ans. $t = -2.41$)
11. $H_0: \mu = 47$ against $H_1: \mu < 47$, given nine values of a sample as 45, 47, 50, 52, 48, 47, 49, 53, 50
(Ans. $t = 2.36$)
12. $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ given, $n_1 = 7$, $n_2 = 11$, $\bar{x}_1 = 16.29$, $\bar{x}_2 = 14.82$, $s_1^2 = 9.609$, $s_2^2 = 15.364$ assuming equal variance.
(Ans. $t = 0.87$)
13. $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 < \mu_2$ given, $n_1 = 31$, $n_2 = 30$, $\bar{x}_1 = 33.29$, $\bar{x}_2 = 34.5$, $s_1^2 = 1.41$, $s_2^2 = 1.09$
(Ans. $z = -4.23$)
14. $H_0: \mu_A = \mu_B$ against $H_1: \mu_A < \mu_B$ given that

	Set A	Set B
Mean	67.42	67.25
SD	2.58	2.50
Sample size	1000	1200

(Ans. $Z = 1.56$)

15. For each of the following information, test the given hypothesis at 5% level of significance (where, x is the number of objects of a particular category)
 - $H_0: P = 0.35$, $H_1: P \neq 0.35$; $n = 100$, $x = 40$
 - $H_0: P = 0.70$, $H_1: P \neq 0.70$; $n = 300$, $x = 220$
 - $H_0: P = 0.45$, $H_1: P > 0.45$; $n = 250$, $x = 150$

- d. $H_0: P = 0.65, H_1: P < 0.65; n = 500, x = 300$
e. $H_0: P = 0.55, H_1: P > 0.55; n = 300, x = 180$
f. $H_0: P = 0.85, H_1: P < 0.85; n = 350, x = 250$
g. $H_0: P_1 = P_2, H_1: P_1 \neq P_2; n_1 = 250, n_2 = 450, x_1 = 150, x_2 = 280$
h. $H_0: P_1 = P_2, H_1: P_1 \neq P_2; n_1 = 250, n_2 = 450, x_1 = 150, x_2 = 280$
16. Test the significance of population correlation and regression co-efficient for the following sample statistics (for first four cases, you have to compute correlation co-efficient and regression co-efficient, this type of information)
- a. $\Sigma x = 3000, \Sigma y = 6000, \Sigma x^2 = 187500, \Sigma y^2 = 937500, \Sigma xy = 367500, n = 60$.
b. $\Sigma x = 1200, \Sigma y = 6000, \Sigma x^2 = 240000, \Sigma y^2 = 965040, \Sigma xy = 330600, n = 60$.
c. $\Sigma x = 600, \Sigma y = 3000, \Sigma x^2 = 606000, \Sigma y^2 = 487500, \Sigma xy = 210000, n = 60$.
d. $\Sigma x = 3000, \Sigma y = 6000, \Sigma x^2 = 187500, \Sigma y^2 = 937500, \Sigma xy = 367500, n = 60$.
e. $r_{xy} = 0.65, n = 25$, against a two tailed alternative
f. $r_{xy} = 0.30, n = 10$, against a two tailed alternative
g. $r_{xy} = 0.40, n = 17$, against a two tailed alternative
h. $r_{xy} = 0.25, n = 22$, against an alternative of positive correlation
i. $r_{xy} = 0.90, n = 16$, against an alternative of positive correlation
j. $r_{xy} = -0.80, n = 20$, against an alternative of negative correlation
k. $r_{xy} = -0.45, n = 9$, against an alternative of negative correlation

Applications

1. The fasting blood sugar (FBS) of 15 randomly selected patients are given below:

110, 118, 130, 140, 142, 146, 112, 100, 95, 98, 96, 122, 123, 124, 130

(i) Do you think that the mean FBS of patients in the population is 110? It is known that the population variance of FBS is 300. (ii) If population variance is unknown, how would you conclude?

(Ans. $z = 2.03; s^2 = 287.78, t = 2.07$)

2. The average monthly expenditure of 100 randomly selected students from different private universities is recorded a 12480.00 taka. The variance of monthly expenditure is 4194236.15. From the information can we conclude at 5% level of significance that average monthly expenditure of all students of private universities is 10000.00 taka?

(Ans. $z = 12.11$)

3. A packaging device is set to fill detergent power packets with a mean weight of 5 kg. These are known to drift upwards over a period of time due to machine fault, which is not tolerable. A random sample of 100 packets is taken and weighed. This sample has mean weight of 5.03 kg

and a standard deviation of 0.21 kg. Can it be concluded that the mean weight produced by the machine has increased? (conduct a right tailed test using 5% level of significance) (Ans. $z = 1.428$)

4. An automobile company usually produces three cylinder model car whose mean petrol consumption is 9.5 km/liter. But the company launches a new four cylinder car whose mean petrol consumption is claimed to be lower than that of existing auto engine. It is found that the mean petrol consumption for 50 new cars is 10 km/liter with standard deviation of 3.5 km/liter. Test for the hypothesis at 1% level of significance whether new model's petrol consumption is less than the existing model. (Ans. $Z = 0.495$)
5. A poultry feed firm claims that their new food is less costly and more effective than older one. The owner of the firm known from the experience that the weights gained by the chickens after feeding old food for a month after birth is normally distributed with mean 2 kg and standard deviation 0.25 kg. In order to test the claim of the firm, an owner randomly selected 10 chickens, and after feeding the new food for a month found that the average weight of the chickens is 2.8 kg. State the appropriate null and alternative hypothesis, and test the claim of the firm at 5% level of significance. Also find the power of the test. (Ans. $Z = 10.12$)
6. A company claims that average life time of their fluorescent bulbs is more than 1570 hours. A sample of 400 fluorescent light bulbs produced mean life time 1600 hours with standard deviation of 150 hours. Is the claim of company justified at 1% level of significance? Also find the power of the test. (Ans. $Z = 4$)
7. A school teacher found in a journal that the school students usually watch TV for 13 hours. He thinks that his student watch TV for less than 13 hours. In order to test his thought he selected 50 students randomly from the school, and came to know that his students watch TV on an average for 8 hours with standard deviation 5 hours. On the basis of the information, comment whether the teacher can think that his students watch TV for less time than 13 hours at 1% level of significance. Also compute the power function. (Ans. $Z = 7.071$)
8. The following table shows the number of MBAs passing from two private universities A and B employed by stakeholders of different organizations in different periods:

University	No. of MBAs employed as manager
A	18, 16, 15, 20, 18, 15, 12
B	20, 14, 12, 22, 16, 14, 15, 10, 12, 18, 10

Assuming the same population variance for both population, can it be concluded that the employment facility for both universities are the same? (Ans. $t = 0.87$)

9. A sample of 20 items in a shop has mean length 42 inch and standard deviation 5 inch. Test the assumption that the length follow a normal distribution with mean 45 inch. (Ans. $t = 2.683$)
10. A fruit seller claims that the average weight of individual apples of his shop is not less than 110 gms. A random sample of 15 apples from that shop provides the mean weight 122 gms and standard deviation 6 gms. Do the data support seller's claim at 5% level of significance?
11. A manager of a firm wants to purchase electric bulbs. In order to take decision about the brand to be bought, he tested randomly selected 200 bulbs from each of two popular brands A and B. He finds that brand A has a mean life of 2500 hours with standard deviation of 90 hours, and brand B has a mean life of 2650 hours with standard deviation 75 hours. Help the manager to take decision in this regard (use 5% level of significance) (Ans. $Z = 10.864$)
12. Two brands of electric bulbs are quoted at the same price. A buyer tested random sample of 100 bulbs of each brand and obtained the following results:

	Mean life time in hours	SD of lifetime in hours
Brand A	1000	82
Brand B	1248	93

Test whether there is any significant difference in the quality of two brands of bulbs at 1% level of significance and comment on which brand of bulbs the buyer should buy.

13. In a factory, all bulbs are produced by two machines A and B. The manager of the factory thinks that effectiveness of the two machines are not the same. That's why he decided to compare the effectiveness of two machine on the basis of mean life times of the bulbs produced by two machines. Accordingly, he made a survey and obtained the following summary of the results.

	Machine A	Machine B
No. of bulbs	150	250
Average lifetime	1200 hours	1000 hours
Standard deviation	280 hours	240 hours

Test the significance of the difference between life time of bulbs produced by two machines at 1% level of significance and find p-value.

(Ans. $Z = 2.06$)

14. IQ tests on two groups of boys and girls revealed the following results:

For Girls: Sample size = 50 Mean IQ = 75 SD = 10
 For Boys: Sample size = 100 Mean IQ = 70 SD = 12

Is there any significant difference in the intelligence status of boys and girls? (Ans. $Z = 2.7$)

15. To compare the price of a certain product in two cities, ten shops were selected at random in each city. The following observations were obtained:

City A	61	63	56	63	56	62	59	60	61	65
City B	55	54	47	59	51	61	57	54	64	58

Test whether the average price can be said to be the same in the two cities.

16. Two groups of workers each consisting of randomly selected 9 workers of a garments factory were given training using two methods, one is traditional theoretical method and another is a new practical method, in order to improve the working capacity related to time taken by them in packaging stage. After completion of the training The time (in minutes) taken by the workers were recorded as follows:

Traditional method	32	34	31	35	41	44	37	28	35
New method	40	35	27	31	29	32	25	31	34

Can we say that the new method is more effective than the traditional method?

(Ans. $t = 1.64$)

17. An insurance company claims that it takes 2 weeks, on an average to process an auto accident claim. The standard deviation is 6 days. To test the validity of this claim, an investigator randomly selected 36 people who recently filed claims. The sample revealed that the company took on an average 16 days to process these claims. At 1% level of significance, check whether the company actually takes more than 2 weeks to process the claims. State the null and alternative hypothesis clearly. Also find the power of the test. (Ans. $z = 2$)
18. Ten workers were given training program with a view to shorten their assembly time for a certain product. The time (in minute) taken by the workers to assemble the product were recorded before and after training and summarized as below:

Worker	A	B	C	D	E	F	G	H	I	J
Time Before training	15	18	20	17	16	14	21	19	13	22
Time After training	14	16	21	10	15	18	19	16	14	20

On the basis of the data, can it be concluded that the training program has shortened the average assembly time at 5% level of significance?

(Ans. paired $t = 1.309$, not shortened)

19. The measurements (in inch) of lengths of 8 bolts produced in a factory by vernier caliper and micrometer are as follows:

Vernier reading (X)	2.265	2.267	2.264	2.237	2.268	2.263	2.264	2.258
Micrometer reading (Y)	2.270	2.268	2.269	2.273	2.270	2.270	2.268	2.267

- Test at 5% level whether the difference between measurements of length by two instruments is significant or not. (Ans. Paired $t = 5.05$)

20. Eleven BBA students were given a test in Accounting out of 30. They were given coaching for two months and a second test was conducted at the end of it. Marks obtained by the students were recorded as follows :

Student	A	B	C	D	E	F	G	H	I	J	K
Marks (1st test)	23	20	19	21	18	20	18	17	23	16	19
Marks (2nd test)	24	19	22	18	20	22	20	20	23	20	17

Do the information support that the students were benefitted by the extra coaching? (consider 5% level of significance)

(Ans. paired $t = 1.48$, not benefitted)

21. The following data represent the blood sugar of a group of patients before (B) and after (A) a specific treatment. Is the treatment successful?

Patient	A	B	C	D	E	F	G	H	I	J
Blood sugar (B)	142	146	156	120	138	160	150	155	180	213
Blood sugar (A)	120	130	120	115	100	160	150	138	160	180

(Ans. paired $t = 4.22$, successful)

22. In a dairy farm two new feeds are introduced for milking cows so that the milk production is increased. The milk production of randomly selected 10 cows are recorded on two different occasions, where feed-I is given to the cows at first instance and after sometime feed-II is given. The milk production (in liter) are shown in following table:

Cow	1	2	3	4	5	6	7	8	9	10
Feed-I	12.8	30.6	26.4	18.7	20.5	30.6	30.5	28.6	25.4	25.0
Feed-II	15.0	30.0	32.4	20.6	18.2	28.4	35.6	32.4	25.0	22.0

Are the two feeds same ?

(Ans. paired $t = -1.03$, same)

23. A farmer grows crops on two fields A and B. On A he uses fertilizer of worth taka 32 per kg and on B he uses fertilizer of worth taka 38 per kg. The net returns, taka per acre, exclusive of the cost of fertilizers on the two fields in the recent five years are as follows:

Year	2005	2006	2007	2008	2009
Field A	24	28	42	37	44
Field B	36	33	48	38	50

Other things being equal, discuss whether the farmer should continue using more expensive fertilizer

(Ans. Paired $t = 3.814$)

24. A certain medicine given to each of 9 patients resulted in the following increase of blood pressure: 7, 3, -1, 4, -3, 5, 6, -4, -1
Can it be concluded that the medicine will, in general, be accompanied by an increase in blood pressure? (Ans. paired $t = 1.305$, not accompanied)
25. A mobile phone company expects that at least 20% of the consumers of a certain area should use their mobile set. Accordingly a survey is made on randomly selected 400 consumers of that area after intensive

campaign, advertisement and significant reduction on sale, and it has been found that 100 use that particular set. Do the data support that the company could achieve its goal at 5% level of significance. State the null and alternative hypothesis clearly. Also find the power of this test.

26. It has been found that 45% the students in all universities in Bangladesh are smokers. A survey showed that out of 300 students in Chittagong University, 120 are smokers. Is it evident from the information that proportion of smokers in Chittagong University is less in comparison with all Universities in Bangladesh? (Ans. $Z = -2.02$)
27. In a factory two items A and B are produced simultaneously. It is expected that both of the items would be produced equally. A sample of 210 is drawn from the production, 140 of the items are found as from A, and the rest from B. Can it be believed that there will be in general 50% of items A? (Ans. $Z = 1.82$)
28. A cancer Specialist claims that 3 out of 100 can survive for five years if attacked by a certain type of cancer. But this proportion may be improved if a special treatment is taken. After applying that special treatment on a randomly selected 10 patients, it has been found everybody except one of the patients survive for more than five years. Test the Specialist's claim at 1% level of significance.
29. A survey was conducted to compare the performance of two machines A and B, on the basis of proportion of defectives. The result of the survey is summarized below:

	Machine A	Machine B	Total
No. of Defective	28	96	124
No. of non-defectives	572	304	876
Total	600	400	1000

Test at 5% level of significance whether performance of Machine A is better than that of B. (Ans. $Z = 34.13$)

30. In a manufacturing process the proportion of faulty items has been found, from long experience, to be 0.1. The proportion of faulty items in a randomly selected batch produced by a new machine is measured and found as 0.09. The manufacturer wishes to test at 5% level of significance whether or not there has been a reduction in proportion of faulty items, help the manufacturer in this regard.
31. A random sample of 360 customers from a super market shows that 120 leave without buying anything. Can we say at 5% level of significance that most of the customers leave the market without buying anything?
32. A survey was conducted in two districts A and B to know whether the consumption of certain brand of tea in district B is more than district A. A sample of 600 households from district B showed that 28 households consume that brand, while a sample of 400 households from district A

- reveals that 24 households use that brand of tea. Do the data support the assertion?
33. Out of 900 consumers in Dhaka city, 450 use toothpaste of particular brand, while in Chittagong 450 out of 600 use that brand. Test the significance of difference between proportions of consumers of that particular brand of toothpaste in two cities. (Ans. $Z = -9.69$)
34. It has been found that out of 956 new born babies who take birth in a week in Rajshahi city, 52.5% are male, while in Khulna city out of 450, 43.4 are male. Do the information establish the fact that more male babies born in Dhaka city than Rajshahi?
35. To compare the correlation between height of father and height of son in a particular community, a sample of 400 is taken. The sample gives a correlation co-efficient of +0.8, test whether there is any significant correlation between heights of father and son.
36. The correlation between days of absenteeism of workers and distance of their house from factory is found to be 0.85. Can it be concluded at 5% level of significance that the correlation between the two factors is 0.90? (Ans. $Z = 0.72$)
37. The correlation co-efficient between typing speed and the age of 20 typist has been found as 0.63. Test whether the correlation is significantly different from i) zero ii) 0.5 against the alternative that it is positive. (Ans. i) $Z = 0.741$, ii) $Z = 0.549$)
38. The sample correlation 68 pairs of annual returns n common stocks in country A and country B was found to be 0.51. Test the null hypothesis that the population correlation co-efficient is zero against the alternative that it is positive.
39. The correlation between price and demand of a commodity over a month (12 observations) is found as -0.75, test whether the population correlation is significantly different from i) zero ii) 0.90 against the alternative that it is negative.
40. In an advertising study the researchers wanted to determine if there was a relationship between per capita cost and the per capita revenue. The data provide a correlation co-efficient as 0.38. Test against a two sided alternative whether the population correlation co-efficient is significantly different from i) zero ii) 0.5 at 10% level of significance.
41. The results obtained for the fitted regression model of sales (y) on income (y) of a retail shop for seven days of a week is given below:

$$Y = 1230.5 + 3.8150 x$$

$$Se \quad (184.91) \quad (2.456)$$

Test the significance of regression co-efficient at 1% level of significance and find 95% confidence interval for population regression co-efficient.

42. A Courier Express authority recorded the following data regarding distance (X) and the time usage (Y) for hand delivery of 16 packets:

Distance and time usage for delivery of 16 packets

Packet No.	Distance (in miles)	Time usage (in hour)	Packet No.	Distance (in miles)	Time usage (in hour)
1	3	9	9	15	33
2	6	20	10	16	27
3	8	20	11	16	32
4	9	16	12	17	34
5	11	23	13	19	41
6	11	19	14	19	34
7	13	20	15	21	47
8	14	30	16	22	47

- i. Find the product moment correlation co-efficient and test its significance against a positive alternative.
- ii. Fit a regression line of time usage on distance and test significance of regression co-efficient
- iii. Compute 95% confidence interval for population regression co-efficient.
43. In a manufacturing process the assembly line speed (meters per minute) (x) was thought to affect the number of defective parts found (y) during the inspection process. In order to verify this theory, the management devised a situation where the same batch of parts was inspected visually at a variety of line speeds. The summary of collected data is as follows:
- $$\Sigma x = 70, \Sigma x^2 = 936, \Sigma y = 102, \Sigma xy = 1136, \Sigma y^2 = 1768, n = 10$$
- (i) Compute correlation co-efficient between speed and number of defective parts and test the significance at 1% level of significance
- (ii) Developed the regression line that relates line speed to the number of defective parts found and comment how much assembly line is affected by the number of defective parts
- (iii) Test whether there is any significant effect of number of defective parts on the assembly line speed (test regression co-efficient) at 5% level of significance.
- (iv) Find 95% confidence interval for population regression co-efficient.
44. The following table shows the results of inoculation against Hepatitis B. in a certain village

	Attacked	Not attacked	Total
Inoculated	267	37	304
Not inoculated	757	155	912
Total	1024	192	1216

Test whether inoculation is effective for the disease or not at 5% level of significance. (Ans. $\chi^2 = 3.992$)

45. A fashion house is interested to see whether there is any association between the preference of colour and sex of customers. She collected the information regarding the preference of colour from a randomly selected sample of 200 customers. The result is summarized in following table:

Colour	Male	Female	Total
Red	10	40	50
Green	70	30	100
White	30	20	50
Total	110	90	200

Comment on the association between the preference of colours and sex of customers at 1% level of significance. (Ans. $\chi^2 = 34.34$)

46. 500 families were selected at random in a city to test the belief that high income families usually send their children to public school and the low income families often send their children to government schools. The following table is obtained from the sample:

	Public School	Govt School
Low income	185	215
High income	65	35

Test at 5% level of significance whether the income and choice of school are independent.

47. A survey is conducted by an ice cream company to test if the quality of ice cream consumed by the customers depends on their sex. For that the company interviewed 1000 people of different corners, and tabulated the result as follows:

Sex of customers	Quality of Ice Cream		
	Type I	Type II	Type III
Male	160	290	210
Female	90	110	140

Test at 5% level if there is any association between the quality of ice cream and sex of consumers.

48. A survey on a number of consumers regarding the quality of a product in rural and urban areas produces the following results:

Opinion	Rural	Urban
Highly Satisfactory	40	150
Satisfactory	55	130
Not Satisfactory	125	90

On the basis of information, can it be concluded at 5% level of significance that the opinion is independent of the area?

49. A market survey conducted in four cities pertaining to the preference of a brand of particular brand of soap yields the following results.

	Dhaka	Chittagong	Rajshahi
Preferred	40	50	55
Not Preferred	40	45	35
No opinion	15	10	15

On the basis of information, can it be concluded at 5% level of significance that the preference of the commodity is independent of the area?

50. A study is undertaken on 500 items of goods of different qualities in order to verify the dependence of customers' preference of goods on the quality. The results are summarized as below:

Quality of goods	Preferred	Not-preferred
High	95	55
Medium	80	100
Low	125	45

Test whether customers' preference of goods depends on the quality.

51. A fish seller claims that the average weight of his fishes is not less than 12 kg. A random sample of 25 large fishes gave an average weight of 10.95 kgs per tin with a standard deviation of 3.21 kg. Could we accept the claim of seller at 5 % level of significance? Also mention the assumptions involved.
52. A mango seller claims that his mangoes do not decay before 28 hours of displaying in the shop. He also knows that the SD of decaying periods is 3.56 hours. A customer checked the duration of 10 mangoes and found the following observations (in approximate hours):

22, 35, 29, 30, 40, 25, 33, 28, 31, 27

Perform a test to verify the seller's claim at 5% level of significance.

53. Two companies produce two types of choc bars, say, Type A and Type B. The company claims that the children prefers Type B to Type A. An interested person observes the sales of bars for a day, and found that out of 200 bars sold in that day, the children bought 90 of Type A and rest of Type B. On the basis of the information provided, comment on the company's claim.
54. Before an increase in the excise duty on tobacco, 200 people out of a sample of 500 people were found to be habituated with smoking. After the increase of duty 190 people out of 600 were found to smoke. Establish the necessary null and alternative hypothesis and comment whether smoking habit was significantly reduced due to imposing excise duty.
55. An e-commerce research company claims that 60% or more graduate students have bought merchandise on-line. A consumer group is suspicious of the claim and thinks that the proportion is lower than

60%. A random sample of 80 graduate students show that only 22 students have ever done so. Is there enough evidence to show that the true proportion is lower than 60%?

56. A composition teacher wishes to see whether a new grammar program will reduce the number of grammatical errors per student make, when writing a two-page essay. The data are shown here. Can it be concluded that the number of errors has been reduced by the new program at $\alpha = 0.05$?
57. In order to test whether there is any significant difference between the mean income of two groups of people, the following information have been obtained:

	Group A	Group B
Sample size	120	125
Mean income	Tk 12000	Tk 15000
SD	Tk 8050	Tk 8970

Establish the necessary null and alternative hypothesis and comment, use 10% level of significance.

58. A private University claims that the average score of their students after final exam is not less than 3.40. An interested person made an effort to verify this claim. Accordingly, he collected a sample of 8 scores from the third semester BBA students. The scores are 3.00, 3.75, 3.30, 2.90, 3.60, 3.40, 3.80, and 3.35. What should be your comment on the University's claim at 2.5% level of significance (write in details).
59. Two types of batteries are tested for their length of life considering a sample of size 9 from Type A and 8 from Type B. It has been found that mean life in hours and variance for Type A are 600 and 121 respectively, and that of Type B are 640 and 144 respectively. Is there any significant difference between the average length of life of two types of batteries at 5% level of significance?

CHAPTER - 17

STATISTICAL QUALITY CONTROL

17.1. Introduction

Now a day's high quality of the products has appeared as an important component of any sale. Buyers do not bother to pay higher price for a high quality product, particularly, because of its reliability of service over a long period of time. Anderson et al (1994) define the quality as the totality of features and characteristics of a product or service that bears on its ability to satisfy given needs. Quality is important both for the buyer as well as for the producer. Producers should be aware of maintaining the quality of their products. Thus, statistical quality control is a powerful productivity technique for effective diagnosis of lack of quality (or conformity to the desired standards) in any of the materials, process, machines or end products. It is essential that the end products possess the qualities that consumer expects of them, because the progress on industry depends on the successful marketing of the products. Quality control ensures this by insisting on quality specifications all along the period from the arrival of materials through each of their processing to the final delivery of goods.

Definition. Statistical quality control is an effective statistical method for determining the extent to which desired qualities of a product are being met without checking every item produced and for identifying whether or not the variations occurring in the product are exceeding the expectations.

Statistical quality control enables us to decide whether to accept or reject a particular product. It is considered as an effective system which coordinates between the quality management and quality improvement efforts of various groups in an organization so as to enable production at the most economical levels which allow for all customer satisfaction.

Quality control, therefore, covers all the factors and processes of production, which may be broadly classified as follows:

- i) **Quality of materials:** Material of good quality will result in smooth processing thereby reducing of waste and increasing the output. It will also give better finish to end products.
- ii) **Quality of manpower:** Trained and qualified personnel will give increased efficiency due to the better quality production through the application of skill and also reduce production cost and waste.
- iii) **Quality of machines:** Better quality equipment will result in efficient work due to lack or scarcity of breakdown and thus reduce the cost.

- iv) **Quality of management:** A good management is imperative for increase in efficiency, harmony in relations, growth in business and markets.

17.2. Objectives of Quality Control

The main objectives of quality control are to (i) establish standards of achievable quality and (ii) set up suitable controls to implement the quality standards by measuring characteristics of the raw material, products and services, and comparing these measurements with established standards.

17.3. Causes of Variations

It is true that variations in the quality of manufactured product in the repetitive process in any industry are inherent and inevitable. These variations are either natural or unnatural. Depending on the nature, these variations are broadly classified as being due to two causes, viz, (i) chance causes and (ii) assignable causes.

17.3.1. Chance causes of variation. Some stable pattern of variation or a constant cause system is inherent in any particular production and inspection process. This pattern results from many minor causes that behave in a random manner. The variation due to these causes is beyond the control of human hand and cannot be prevented or eliminated under any circumstances. We have to allow this type of variation within the stable pattern, usually termed as allowable variation or chance variation, also termed as common causes or random causes or uncontrollable causes. The range of variation due to this type of causes is known as natural tolerance of the process.

Definition. Chance causes. The variation in the production which causes due to unassignable problem inherent in the process or which is natural and can not be removed or prevented totally, is called chance causes of variation.

17.3.2. Assignable causes of variation. The second type of variation attributed to any production process is due to non-random or the so-called assignable causes or special causes, also termed as preventive variation. The assignable causes may creep in at any stage of the process, right from the arrival of the raw materials to the final delivery of goods. Some of the important factors of assignable causes of variation are substandard or defective raw material, new techniques or operations, negligence of the operators, wrong or improper handling of machines, faulty equipment, unskilled or inexperienced technical staff and so on. These causes can be discovered in a production process before it goes wrong, that means, before the production becomes defective.

Definition. Assignable causes. The variation in the production, which is due to identifiable reasons and can be removed from the process, is called assignable causes of variation.

The statistical quality control means planned collection and effective use of data for studying causes of variation in quality either as between processes, procedures, materials, machines, etc. or over periods of time. This cause-effect analysis is then fed back into the system with a view to continuous action on the processes of handling, manufacturing, packaging, transporting and delivery at end-use.

Thus, the main purpose of Statistical Quality Control is to devise statistical techniques which would help us in separating the assignable causes from the chance causes, thus enabling us to take immediate remedial measure/action whenever assignable causes are found to be present. A production process is said to be in a state of statistical control, if it is governed by chance causes alone, in the absence of any assignable causes of variation.

17.4. Uses of Statistical Quality Control (SQC)

The following are some of the uses of SQC that might result when a process is brought in good statistical control

- i) The act of getting a process in statistical control involves the identification and elimination of assignable causes of variation and possibly the inclusion of good ones viz. new material or method. This helps in the direction and correction of many production troubles and brings about a substantial improvement in the product quality and reduction of spoilage and rework.
- ii) It tells us when to leave a process alone and when to take action to correct troubles, thus preventing frequent and unwarranted adjustments.
- iii) It provides better quality assurance at lower inspection cost.
- iv) The very presence of a quality control scheme in a plant improves and alerts the personnel. Such a scheme is likely to breed quality consciousness throughout the organization, which is of immense long run value.
- v) Statistical quality control reduces waste of time and material to the absolute minimum by giving an early warning about the occurrence of defects. Savings in terms of the factors stated above means less cost of production and hence may ultimately lead to more profits.

17.5. Techniques of Statistical Quality Control

If the process is such that the items are of high quality and there is hardly any chance of defective item, the product may be sold without inspection. In that case there is no need of controlling or inspecting the quality of the product.

However, if it is necessary to inspect the quality, one can do it in two ways. There are two different ways of controlling the quality of a product. These are:

- i) Through 100% inspection (which is not statistical) and
- ii) Through statistical quality control of process or product.

17.5.1. Through 100% inspection. The system of 100 percent inspection of items is not a statistical way of controlling the production process or production. Hence, this is not a satisfactory way for the purpose due to following reasons:

- i) It is an expensive way and may induce poor performance from concerned person in the belief that poor work will be detected later.
- ii) Where inspection is destructive or detrimental to the product produced, the use of 100% inspection may remove the entire output of the process.
- iii) It is not always reliable because it becomes too much a routine for a person inspecting each and every product, hence defective products may also be noted as 'satisfactory'.
- iv) For the output of an ongoing process, 100% inspection is not possible.

Anyway, in spite of the above-mentioned reasons against the 100% inspection, it may be essential to inspect the quality in the following situations:

- i) Any single defective item may cause danger to life
- ii) Any single defect may stop the whole function of system
- iii) The lot size is small
- iv) The incoming quality of the product is very poor or quite unsatisfactory.

17.5.2. Through Statistical Quality Control. As it is mentioned, 100% inspection is not economical or sometimes not possible. So instead of this process, statistical tools are used to measure the incoming or outgoing quality of products or services. The application of random sampling provides each item with an equal chance of being selected and the value of sample statistic permits logical inference to be made about the quality of entire production of unit. The statistical quality control can be exercised over the process level, called process control and product level, called product control.

Advantages of quality control

Statistical quality control is one of the tools of scientific management. It has the following advantages over 100% inspections:

- i. *Reduction of cost*: Since only a fraction of output is inspected, costs of inspection are greatly reduced!
- ii. *Greater efficiency*: It is not that only costs of inspection is reduced, but the efficiency also goes up because much of the monotony of inspecting all items is avoided, the work of inspection being considerably reduced.
- iii. *Easy to apply*: An excellent feature of quality control is that it is easy to apply. Once the system is established, persons who have not had the extensive specialized training or a highly mathematical background can operate it. It may appear difficult, only because the statistical principle on which it is based are unrecognized or unknown. However, as these principles are based on commonsense, the quality control method finds wide applications.
- iv. *Early detection of faults*: Quality control ensures an early detection of faults and hence a minimum waste of rejects production. The moment a sample point falls outside the control limits, it is taken to be a danger signal and necessary corrective action is taken. On the other hand, with 100% production, unwanted variations in quality may be detected at a stage, when a large amount of faulty products have already been produced. Thus, there would be a big wastage. Control chart, on the other hand, provides a graphic picture of how the production is proceeding and to tell management where to look for trouble.
- v. *Advances to specifications*: Quality control enables a process to be brought into and held in a state of statistical control, that means, a state in which variability is only due to chance causes. So as long as statistical quality control continues, specifications can be accurately predicted for the future, which even 100% inspection can not guarantee. Consequently, it is possible to assess whether the production processes are capable of turning out products, which will comply, with the given set of specifications.
- vi. *The only course*: In certain cases, 100% inspection can not be carried out without destroying all the products inspected, for example, testing breaking strength of pencils, proofing of ammunition, testing longevity of marker pen, etc. In such cases, if 100% inspection method is followed, then all the items inspected will be spoiled. Hence, statistical quality control method is more economical.

Limitations of quality control

In spite of its great significance in controlling the quality of products, the statistical quality control is not free from limitations. Some of the limitations of quality control are mentioned below:

- i. **Not a unique solution:** It is not a perfect solution to all problems regarding quality.
- ii. **May be dangerous:** The application of standard process without adequate study of the process may be dangerous.
- iii. **Not the responsibility of statistician:** The responsibility for quality and process decisions rests with the manager in-charge of the process, not with statistician who identifies if the process is not within statistical control.

17.6. Process and Product Control

The main objective in any production process is to control and maintain the quality of the manufactured product so that it conforms to specified quality standards. In other words, we want to ensure that the proportion of defective items in the manufactured product is not too large. This is called **process control** and is achieved through the technique of control charts. A control chart provides a basis for deciding whether the variation in the output is due to random causes or due to assignable causes. Hence, it helps in taking decision whether to adjust the process or not.

On the other hand, once the production is over, the process through which the quality of the products ensured is known as **product control**, this is achieved through the technique of sampling inspection of final products.

Definition. Process control. When the function of production process is monitored in order to control the quality of product using control chart technique, then it is called statistical process control

Definition. Product control. When the quality of product is controlled by critical examination using sampling inspection of product, it is called statistical product control.

17.6.1. Control Charts (Process control). Control chart introduced by Dr. Walter Shewhart in 1924 provides a basis of deciding whether the variation in the output is due to random causes or due to assignable causes. It is possible to monitor the variation in the predetermined quality of a product or process through control chart. The control chart is used for following purposes.

- i) To focus on the time dimension in which a system produces products or services
- ii) To identify the nature of variation in the process during the operation
- iii) To ensure that only desired quality of product is maintained
- iv) Above all to decide whether the process of production is in control or not.

Definition. Control Chart. A control chart is a graphic device for presenting data so as to identify the frequency and extent of variations from a preset standard of product

A control chart is designed to display successive measurements of process through the three equidistant lines. Thus, a typical control chart consists of the three horizontal lines viz. i) A central line (CL) to indicate the desired standard or the level of the process, ii) Upper control limit (UCL) and iii) Lower control limit (LCL), together with a number of sample points as shown in the following diagram.

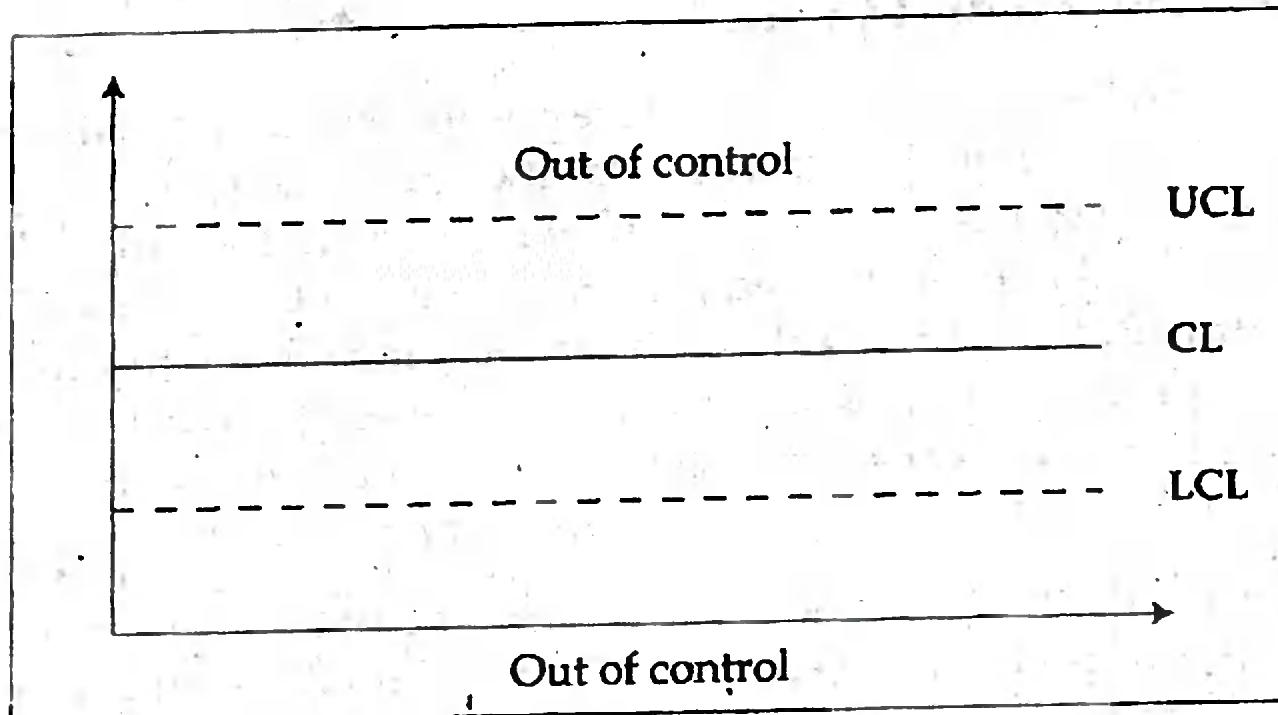


Fig. 17.1. Outline of a typical Control Chart.

In the control chart, upper control limit and lower control limit are usually plotted as dotted lines and central line is plotted as a bold line. If t is the underlying statistic then these values depend on the sampling distribution of t and are given by

$$UCL = E(t) + 3SE(t), \quad LCL = E(t) - 3SE(t) \quad \text{and} \quad CL = E(t)$$

This is called $\pm 3\sigma$ limits. There is statistical logic of considering this type of limits in the construction of control chart.

Again, if the statistic t is normally distributed with mean μ and variance σ^2 , then we have

$$\Pr[\mu - 3\sigma < t < \mu + 3\sigma] = 0.9973$$

$$\text{Or, } \Pr[|t - \mu| < 3\sigma] = 0.9973$$

In other words, the probability that a random value of t goes outside the 3σ limits viz. $\mu \pm 3\sigma$ is 0.0027, which is very small. That's why in the construction of control chart $\pm 3\sigma$ limits are used. Dr. Walter Shewhart, hence, gives these limits. This type of charts are also known as Shewhart's Control Charts. $\mu \pm 3\sigma$ are the limits within which the process will normally

perform. The width of this interval 6σ is sometimes called the natural tolerance of the process.

Definition. Natural tolerance. The width of the interval, which provides the measure of the variability that can be expected in specifications of a product.

Control charts are mainly constructed for variables and attributes related to products.

The steps needed to set up a control procedure

The manufacturer may follow the following steps in order to establish basic procedures for the operation of quality control program:

- i) Select the quality characteristics that are controlled.
- ii) Analyze the production process to determine the kind and location of possible causes of irregularities.
- iii) Determine how the inspection data are to be collected and recorded, and how they are to be subdivided.
- iv) Choose the appropriate statistical measures that to be used in the charts.

When to use what type of control charts.

Depending on the type of inspection data available, any one of the following control charts may be used:

- i) **Control chart for mean and standard deviation, and mean and range:** These types of control charts are used when measured values of the quality characteristics are at hand.
- ii) **Control chart for mean alone:** It is used where experience with control charts for mean and range, or mean and standard deviation has already demonstrated that instances of lack of control are almost always associated with causes that effect mean rather than standard deviation or range.
- iii) **Control chart for standard deviation or range alone:** Control chart for standard deviation or range is used alone when control for mean is known or be very much expensive.
- iv) **Control chart for number of defects:** This chart is used when the inspection consists of determining the number of defects, say, c in a sample, for example, in the examination of finished textiles, materials, wire, etc.
- v) **Control chart for fraction defectives or number of defectives:** These are used when the record of inspection of testing show merely the number of articles inspected and the number found defectives.

17.7. Control Chart for Variables

No production process is perfect enough to produce all the items exactly alike. Some amount of variation, in the produced items, is inherent in any production scheme. This variation is the totality of different steps in a production process viz. raw material, machine setting and handling, operators, etc. The control limits in the \bar{X} (mean) and R (Range) are so placed that they reveal the presence or absence of assignable causes of variation in the

- a) Average – mostly related to machine setting.
- b) Range – mostly related to the negligence on the part of the operator.
- c) Control chart for variability or Standard deviation.

17.7.1. \bar{X} and R charts. The stepwise procedure of the construction of \bar{X} chart and R chart are described below:

I. Measurement: Actually the work of a control chart starts first with the measurements. Any method of measurement has its own inherent variability. Errors in measurement can enter into the data by:

- i) The use of faulty instruments,
- ii) Lack of clear cut definitions of quality characteristics and the method of taking measurements, and
- iii) Lack of experience in the handling or use of the instruments, etc.

II. Selection of samples or subgroups: In order to make the control chart analysis effective, it is essential to pay due regard to the rational selection of the samples or sub-groups. The choice of the sample size n and the frequency of sampling i.e. time between the selection of two groups, depend on the process and no hard and fast rules can be laid down for this purpose. Usually, n is taken to be 4 or 5 while the frequency of sampling depends on the state of control exercised. Normally, 25 samples size 4 or 20 samples of size 5 each under control give good estimate of the process average and dispersion.

III. Calculation of \bar{X} and R for each sub-group: Like the usual definition of these two measures of numerical characteristics, the mean and range for i th sample are defined as

$$\bar{X}_i = \frac{1}{n} \sum x_{ij} \quad \text{for } j = 1, 2, \dots, n$$

$$R_i = \max x_{ij} - \min x_{ij}$$

and $s_i^2 = \frac{1}{n} \sum_j (x_{ij} - \bar{X}_i)^2$ for all $i = 1, 2, \dots, k$ if there are k samples.

Next it is required to compute \bar{X} , \bar{R} , \bar{s} , i.e. the averages of sample means, average of sample ranges and average of sample standard deviations, respectively as follows:

$$\bar{X} = \frac{1}{k} \sum \bar{X}_i, \quad \bar{R} = \frac{1}{k} \sum \bar{R}_i \text{ and } \bar{s} = \frac{1}{k} \sum s_i$$

IV. Setting of control limits: It is well known that if σ is process standard deviation (SD of the population from which samples are taken), then the standard error of sample mean is $\frac{\sigma}{\sqrt{n}}$, where n is the sample size, i.e.

$SE(\bar{X}_i) = \frac{\sigma}{\sqrt{n}}$ for $i = 1, 2, \dots, k$ and $SE(\bar{R}) = c \bar{R}$, where c is a constant depending on n , to be collected from the statistical tables.

Control limits for \bar{X} chart:

There are two cases under which we have to construct control chart for mean and ranges. For example, the case when the values of parameters μ and σ are known, and when they are unknown. The method of construction of control chart under these two situations are as follows.

Case I : When standards are given i.e., when both μ and σ are known. The 3σ control limits for \bar{X} chart are given by $E(\bar{X}) \pm 3 SE(\bar{X}) = \mu \pm 3 \frac{\sigma'}{\sqrt{n}}$, so

if μ' and σ' are the specified values of μ and σ respectively, then

$$LCL = \mu' - 3 \frac{\sigma'}{\sqrt{n}}, \quad CL = \mu' \text{ and } UCL = \mu' + 3 \frac{\sigma'}{\sqrt{n}}$$

Case II : On the other hand, if standards are not given, i.e., if the values of the parameters μ and σ are to be estimated from the sample, then, the limits are given as

$$LCL = \bar{X} - A_1 \bar{s}, \quad CL = \bar{X} \text{ and } UCL = \bar{X} + A_1 \bar{s}$$

However, these control limits can also be computed using sample range with the following formula,

$$LCL = \bar{X} - A_2 \bar{R}, \quad CL = \bar{X} \text{ and } UCL = \bar{X} + A_2 \bar{R},$$

where the values of A_1 and A_2 can be found in the statistical table for definite n .

Control limits for R chart:

The 3σ control limits for R chart are given by $E(R) \pm 3 SE(R)$, $E(R)$ is estimated by \bar{R} , and σ_R is estimated from the relation $\sigma_R = cE(R) =$

$c\bar{R}$ where c is a constant to be collected from the table. Thus the control limits for R chart are given by

$$LCL_R = \bar{R} - 3c\bar{R} = (1-3c)\bar{R} = D_3\bar{R}, CL_R = \bar{R}$$

$$\text{and } UCL_R = \bar{R} + 3c\bar{R} = (1+3c)\bar{R} = D_4\bar{R},$$

However, the control limits for R-chart can be computed directly from the assumed or known value of σ as follows:

$$LCL_R = D_2\sigma, CL = \bar{R} \text{ and } UCL = D_4\sigma$$

Where, D's are the values to be collected from table for different values of n .

V. Construction of control chart : After computing the values of the control limits, it is necessary to draw a graph with these limits and the sample variables (means or ranges).

VI. Decision making : Generally, if all the sample points lie within the control limits, the process is said to be within control and if any of the sample mean falls outside the limits, it can be concluded that the process is not under statistical control. However, control chart can indicate the lack of control of a process under one or more of the following situations.

17.7.1. Criteria for detecting lack of control in \bar{X} and R charts. Lack of control in the process may be recommended under one or more of the following situations:

- a. A sample point (mean or range) outside the control limits
- b. A run of seven or more points above or below the central line in the control chart
- c. The sample points on \bar{X} and R charts are too close to the central line
- d. Presence of trends
- e. Presence of cycles.

Measures of process capability

Again, whether the process is capable of meeting the specifications can also be judged by using two indices, namely, (i) Capability index (C_p) and (ii) C_{pk} index. These indices are briefly discussed below. These two indices are suitable in two situations viz. when data are centered between the tolerance limits and when are not centered respectively.

Suppose the management sets lower (L) and upper (U) tolerance limits for process performance. Process capability is judged by the extent to which control limits $\mu \pm 3\sigma$ (or, $\bar{x} \pm 3\hat{\sigma}$, when standards are not given) lies between these limits, $\hat{\sigma} = \frac{\bar{s}}{C_4}$ is an unbiased estimate of σ , where, \bar{s} is average of

sample standard deviations and the value of C_p depends on the sample size and can be found in statistical tables.

(i) **Capability index (C_p)** : This measure is appropriate when the sample data are centered between the tolerance limits $\bar{x} = (L + U)/2$. The index is given by

$$C_p = \frac{U - L}{6\hat{\sigma}}$$

A satisfactory value of this index is usually taken to be one that is at least 1.33 [which implies that natural tolerance of the process should be no more than 75% of $(U - L)$, the width of the range of accepted values].

(ii) **C_{pk} index** : When the sample data are not centered between the tolerance limits, it is necessary to allow for the fact that the process operating closer to one tolerance limits than the other. The resulting measure, called the C_{pk} -index, is defined as :

$$C_{pk} = \text{Min} \left[\frac{U - \bar{x}}{3\hat{\sigma}}, \frac{\bar{x} - L}{3\hat{\sigma}} \right]$$

The process is also supposed to be satisfactory if minimum value of this index is 1.33.

17.7.2. Interpretation of \bar{X} and R charts. In order to judge if a process is in control, \bar{X} and R charts should be examined together and the process should be deemed in statistical control if both the charts show a state of control. In addition to this, calculating the values of process capability indices C_p and C_{pk} and comparing them with 1.33, it is possible to take decision whether the process is capable of meeting the specifications.

Example 17.7.1. Using the information given below, construct lower control limit (LCL), central line (CL) and upper control limit (UCL) for mean and range chart (i) for 10 samples each of size $n = 6$, it has been found that $\bar{x} = 546$, $\bar{R} = 84$, (ii) for 12 samples each of size $n = 5$ it has been found that $\bar{x} = 1367.5$, $\bar{R} = 427.5$, (iii) also compute $3 - \sigma$ control limits for mean using the information $n = 4$, $\bar{x} = 0.5230$, $\sigma = 0.0032$.

Solution. (i) Control limits for \bar{X} chart:

$$\text{Central line} = \bar{x} = 546, \text{LCL} = \bar{x} - A_2 \bar{R} = 546 - 0.483 \times 84 = 505.43$$

$$\text{and } \text{UCL} = \bar{x} + A_2 \bar{R} = 546 + 0.483 \times 84 = 586.57$$

Similarly, Control limits for R chart:

$$\text{Central line} = \bar{R} = 84, \text{LCL} = D_2 \bar{R} = 0 \times 84 = 0$$

$$\text{and } \text{UCL} = D_4 \bar{R} = 2.004 \times 84 = 168.34$$

(ii) Control limits for \bar{X} chart:

$$\text{Central line} = \bar{x} = 1367.5,$$

$$LCL = \bar{x} - A_2 \bar{R} = 1367.5 - 0.0577 \times 427.5 = 1120.83$$

$$\text{and } UCL = \bar{x} + A_2 \bar{R} = 1367.5 + 0.0577 \times 427.5 = 1614.17$$

Similarly, Control limits for R chart:

$$\text{Central line} = \bar{R} = 427.5, LCL = D_2 \bar{R} = 0 \times 427.5 = 0$$

$$\text{and } UCL = D_4 \bar{R} = 2.115 \times 427.5 = 904.16$$

(in both cases, the values of A_2 , D_2 and D_4 have been collected from the table)

(iii) 3- σ limits for mean are :

$$\text{Central line} = \bar{x} = 0.523, LCL = \bar{x} - 3 \frac{\sigma}{\sqrt{n}} = 0.5182$$

$$\text{and } UCL = \bar{x} + 3 \frac{\sigma}{\sqrt{n}} = 0.5278.$$

Example 17.7.2. Construct a control chart for mean and the range for the following data on the basis of fuses. 12 samples each of size 5 being taken in every hour (each set of 5 has been arranged in ascending order of magnitude). Comment on whether the production seems to be under control.

Samples	1	2	3	4	5	6	7	8	9	10	11	12
	42	42	19	36	42	51	60	18	15	69	64	61
	65	45	24	54	50	74	60	20	30	109	90	78
	75	68	80	69	57	75	72	27	39	113	93	94
	78	72	81	77	59	78	95	42	62	118	109	109
	87	90	81	84	78	132	138	60	84	153	112	136

Solution. The following table is made for the construction of required control charts.

Table 17.1. Calculation of \bar{X} chart and R chart.

Samples	1	2	3	4	5	6	7	8	9	10	11	12
	42	42	19	36	42	51	60	18	15	69	64	61
	65	45	24	54	50	74	60	20	30	109	90	78
	75	68	80	69	57	75	72	27	39	113	93	94
	78	72	81	77	59	78	95	42	62	118	109	109
	87	90	81	84	78	132	138	60	84	153	112	136
Sample total	347	317	285	320	287	410	425	167	230	562	468	478
Sample means (\bar{X}_i)	69.4	63.4	57.0	64.0	57.4	82.0	85.0	33.4	46.0	112.4	93.6	96.6
Sample range (R_i)	45	48	62	48	36	81	78	42	69	84	48	75
Sample SD (s_i)	17	20	32	19	13	30	33	18	27	30	19	29

Here,

i) Combined mean $\bar{X} = \frac{1}{12} \sum \bar{X}_i = \frac{859.2}{12} = 71.60$

ii) Combined range $\bar{R} = \frac{1}{12} \sum R_i = \frac{716}{12} = 59.67$

iii) From the statistical table of control charts, we have for $n = 5$,

$$A_2 = 0.58, D_3 = 0, \text{ and } D_4 = 2.11$$

So, control limits for \bar{X} chart are as follows:

$$UCL = \bar{X} + A_2 \bar{R} = 71.6 + 0.58 \times 59.67 = 106.21,$$

$$LCL = \bar{X} - A_2 \bar{R} = 36.99, \text{ and } CL = \bar{X} = 71.60$$

The control chart for mean is shown in figure 17.2.

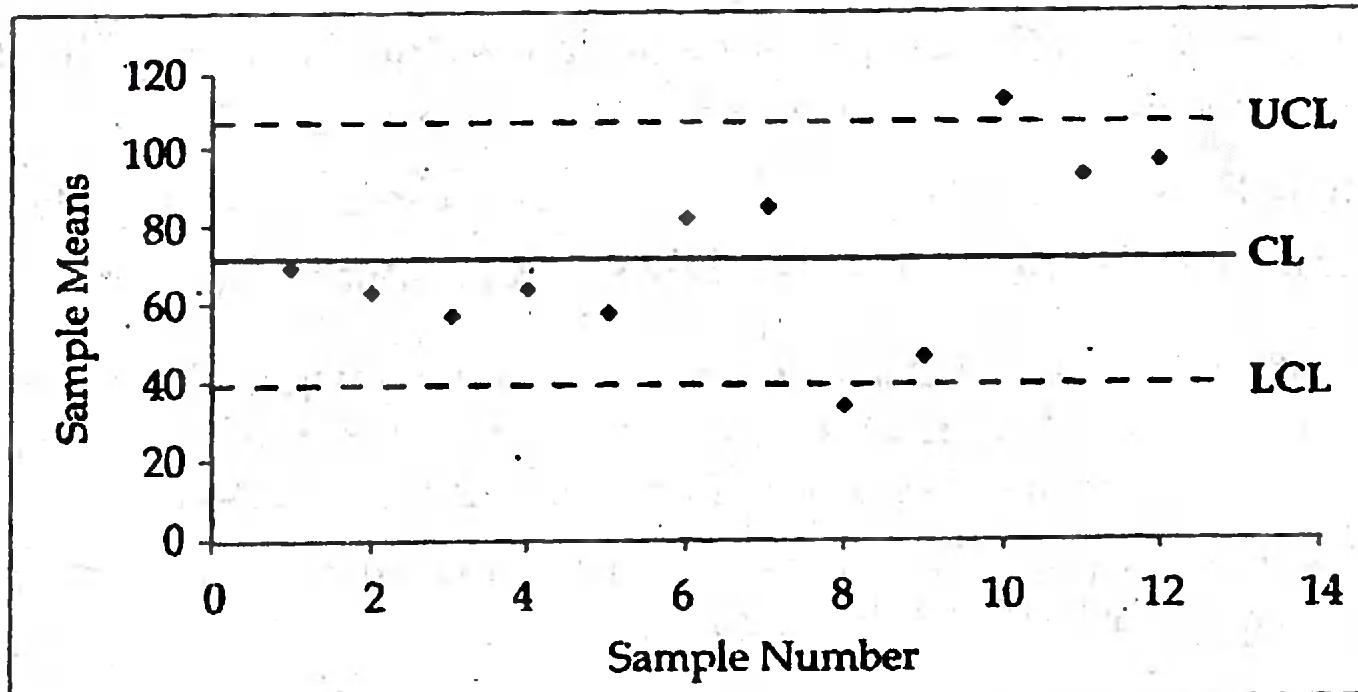


Fig. 17.2. Control Chart for Mean.

From the chart, it is observed that the process is out of control, since the points corresponding to 8th and 10th samples lie outside the control limits. Again, control limits for R charts are as follows:

$$UCL = D_4 \bar{R} = 2.11 \times 59.67 = 125.904, LCL = D_3 \bar{R} = 0 \text{ and } CL = \bar{R} = 59.67$$

The control charts for range are shown in figure 17.3.

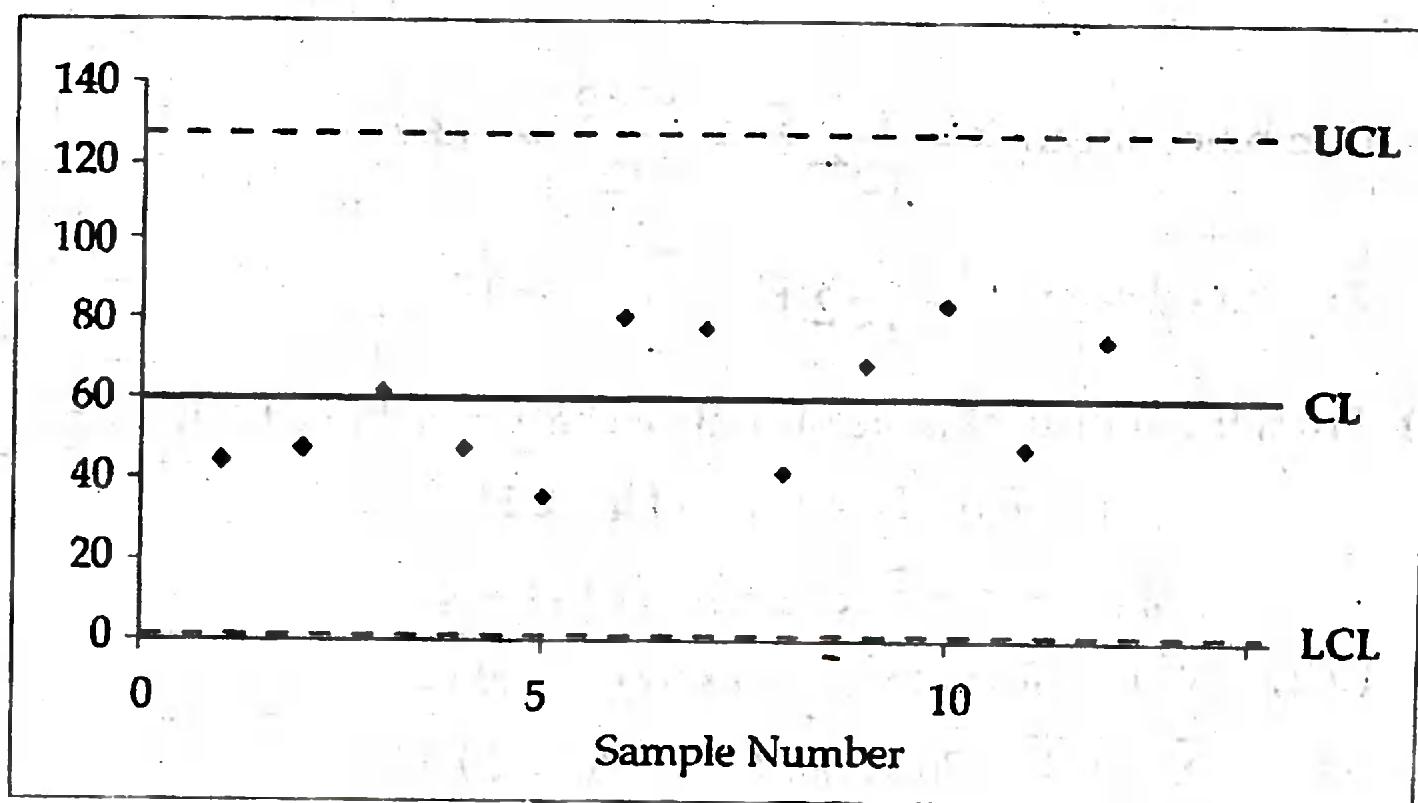


Fig. 17.3. Control Chart for Range.

Conclusion. Since all the sample range fall within the control limits, the chart shows that process is in control.

Although R-chart depicts control, the process can not be regarded as in statistical control because \bar{X} chart shows lack of control.

The choice between \bar{X} chart and R chart is a managerial problem. It is better to construct R chart at first, if this chart indicates that the dispersion of the quality by process, it is out of control and the process can be verified using \bar{X} chart. Generally, it is better not to construct other chart until the quality dispersion is found under control.

Let us now check whether the process is capable of meeting specification or if the process is satisfactory by applying process capability criteria C_p and C_{pk} indexes.

Here, for $n = 5$, $c_4 = 0.940$ and $\bar{s} = \frac{\sum s_i}{12} = 288/12 = 24$, so, $\hat{\sigma} = \frac{\bar{s}}{c_4} = 25.53$.

$$\text{Thus, } C_p = \frac{U - L}{6\hat{\sigma}} = \frac{153 - 19}{6 \times 25.53} = 0.87$$

$$\begin{aligned} \text{And } C_{pk} &= \min \left[\frac{U - \bar{x}}{3\hat{\sigma}}, \frac{\bar{x} - L}{3\hat{\sigma}} \right] = \min \left[\frac{153 - 71.60}{3 \times 25.53}, \frac{71.60 - 19}{3 \times 25.53} \right] \\ &= \min (1.06, 0.69) = 0.69 \end{aligned}$$

Hence, values of both indices are less than 1.33, which indicates that the production process is not capable of meeting the specification.

Example 17.7.3. A food company puts mango juice into cans advertised as containing 10 ounces of the juice. The weights of the cans immediately after filling for 20 samples are taken by a random method at an interval of every 1-hour. Each of the samples includes 4 cans. The sample values are tabulated in the following table. The weights in the table are given in units of 0.01 ounces in excess of 10 ounces. For example, the weight of the juice drained from the first can of the sample is 10.15 ounces, which is in excess of 10 ounces by 0.15 ounces. Since the unit in the table is 0.01 ounces, the excess is recorded as 15 units in the table. Construct a mean chart and range chart to take decision whether the process is under control.

Table. Excess weights of cans of juice.

Sample No.	Weight of cans (4 cans in each sample)				Sample No.	Weight of cans (4 cans in each sample)			
1	15	12	13	20	11	5	12	20	15
2	10	8	8	14	12	3	15	18	18
3	8	15	17	10	13	6	18	12	10
4	12	17	11	12	14	12	9	15	18
5	18	13	15	4	15	15	15	6	16
6	20	16	14	20	16	18	17	8	15
7	15	19	23	17	17	13	16	5	4
8	13	23	14	16	18	10	20	8	10
9	9	8	18	5	19	5	15	10	12
10	6	10	24	20	20	6	14	12	14

Solution. Calculation of \bar{X} chart and R chart are shown in table 17.2.

Table 17.2. Calculation of \bar{X} chart and R chart.

Sample No. (1)	Weight of cans (4 cans in each sample) (2)				Total weight of 4 cans (3)	Sample mean (4)	Sample range (5)
1	15	12	13	20	60	15	8
2	10	8	8	14	40	10	6
3	8	15	17	10	50	12.5	9
4	12	17	11	12	52	13	6
5	18	13	15	4	50	12.5	14
6	20	16	14	20	70	17.5	6
7	15	19	23	17	74	18.5	8
8	13	23	14	16	66	16.5	10
9	9	8	18	5	40	10	13
10	6	10	24	20	60	15	18
11	5	12	20	15	52	13	15
12	3	15	18	18	54	13.5	15
13	6	18	12	10	46	11.5	2
14	12	9	15	18	54	13.5	9

15	15	15	6	16	52	13	10
16	18	17	8	15	58	14.5	10
17	13	16	5	4	38	9.5	12
18	10	20	8	10	48	12	12
19	5	15	10	12	42	10.5	10
20	6	14	12	14	46	11.5	8
Total						263	211

Here, (i) Combined mean $\bar{X} = \frac{1}{20} \sum \bar{x}_i = \frac{263}{20} = 13.15$

(ii) Combined range $\bar{R} = \frac{1}{12} \sum R_i = \frac{211}{20} = 10.55$

(iii) From the statistical table of control charts, we have for $n = 4$, $A_2 = 0.729$

(iv) So, control limits for \bar{X} chart are as follows:

$$\begin{aligned} UCL &= \bar{X} + A_2 \bar{R} = 13.15 + 0.729 \times 10.25 = 20.84, LCL = \bar{X} - A_2 \bar{R} \\ &= 13.15 - 0.729 \times 10.25 = 5.46 \text{ and } CL = \bar{X} = 13.15 \end{aligned}$$

(It is to be noted that the values in the above computations are expressed in units of 0.01 ounces in excess of 10 ounces. Thus the actual value for the UCL is $20.84 \times 0.01 = 0.208$, and that for LCL and CL are 0.0546 and 0.131 respectively). The control chart for mean is shown in figure 17.4.

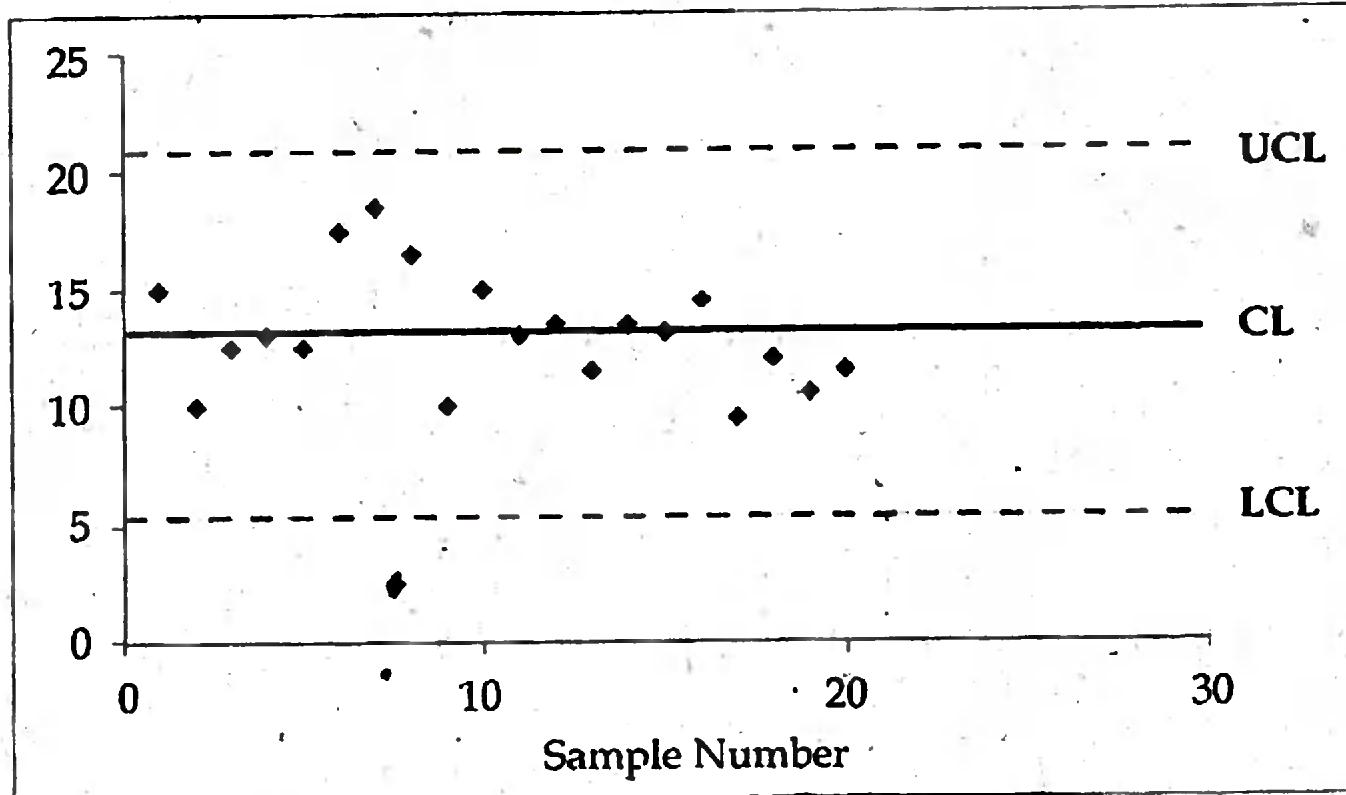


Fig. 17.4. Control Chart for Mean.

Again, the construction of control limits for R charts are as follows:

i) Combined range $\bar{R} = \frac{1}{12} \sum R_i = \frac{211}{20} = 10.55$

ii) From table for $n = 4$, $D_4 = 2.282$ and $D_3 = 0$

$$\text{iii) Hence, } UCL = D_4 \bar{R} = 2.282 \times 10.55 = 20.075,$$

$$LCL = D_3 \bar{R} = 0 \text{ and } CL = \bar{R} = 10.55$$

(However, the actual value for the UCL is 0.201, and that for LCL and CL are 0 and 0.1055 respectively)

The control chart for range is shown in following figure.

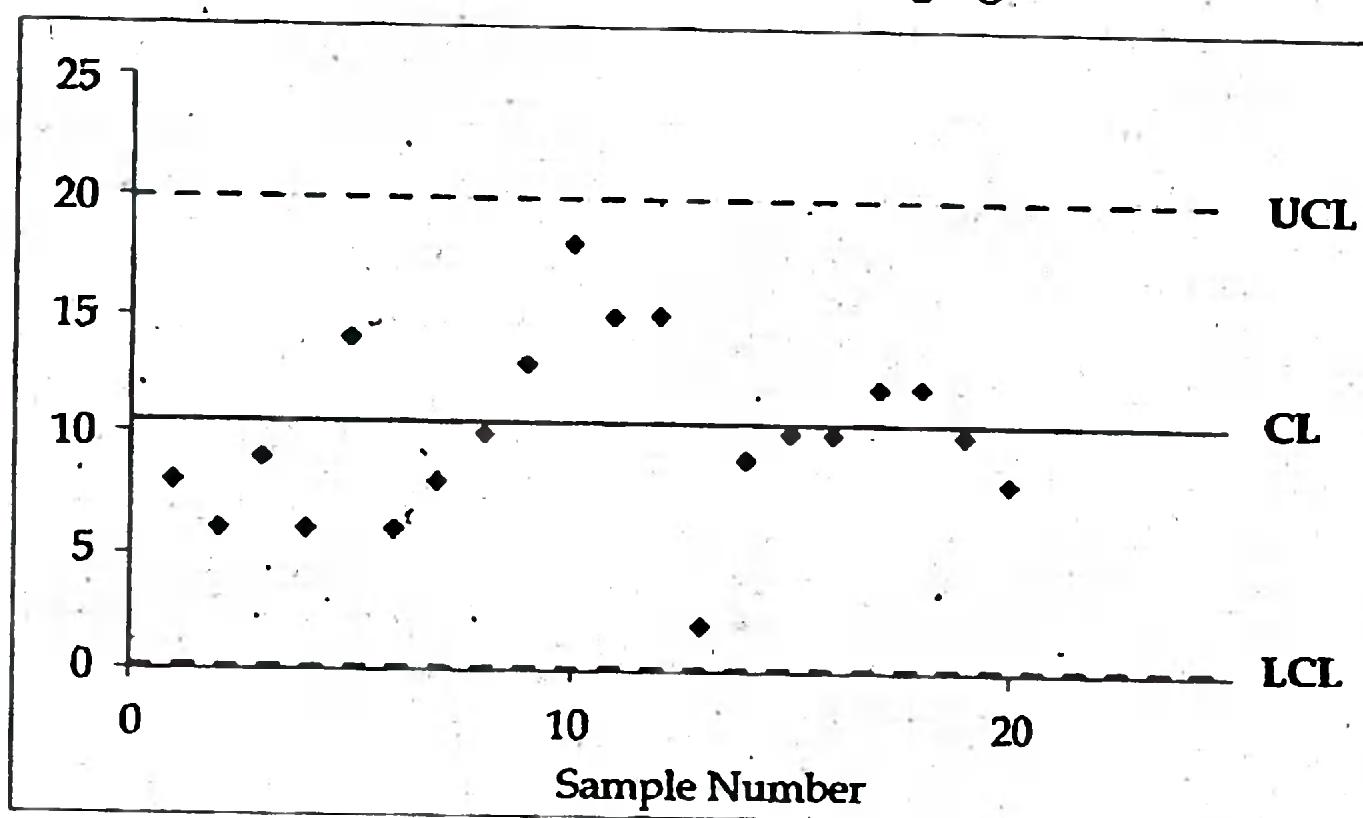


Fig. 17.5. Control Chart for Range.

Conclusion. From the chart, it is observed that the process is in statistical control, because all the sample points of mean and range are falling within the control limits.

Example 17.7.4. A machine is set to deliver an item of a given weight. 10 sample each of size 5 were recorded. The summary of information relevant to data are as follows:

Table 17.3. Mean and range of weights.

Samples	1	2	3	4	5	6	7	8	9	10
Mean (\bar{X}_i)	15	17	15	18	17	14	18	15	17	16
Range (R_i)	7	7	4	9	8	7	12	4	11	5

Calculate the values for the central line and control limits for mean chart and range chart and then comment on the state of control. (Given, Conversion factors for $n = 5$ are $A_2 = 0.58$, $D_3 = 0$, $D_4 = 2.115$)

Solution. From the given information, we have,

$$\text{Combined mean } \bar{X} = \frac{1}{n} \sum \bar{X}_i = \frac{162}{10} = 16.2$$

$$\text{Combined range } \bar{R} = \frac{1}{n} \sum R_i = \frac{74}{10} = 7.4$$

So, control limits for \bar{X} chart are as follows:

$$UCL = \bar{X} + A_2 \bar{R} = 16.2 + 0.58 \times 7.4 = 20.469$$

$$LCL = \bar{X} - A_2 \bar{R} = 16.2 - 0.58 \times 7.4 = 11.931$$

$$CL = \bar{X} = 16.2$$

Again, control limits for R chart are as follows:

$$UCL = D_4 \bar{R} = 2.115 \times 7.4 = 15.614$$

$$LCL = D_3 \bar{R} = 0 \times 7.4 = 0$$

$$CL = \bar{R} = 7.4$$

The control charts for mean and range are shown in figure 17.6 and 17.7 respectively.

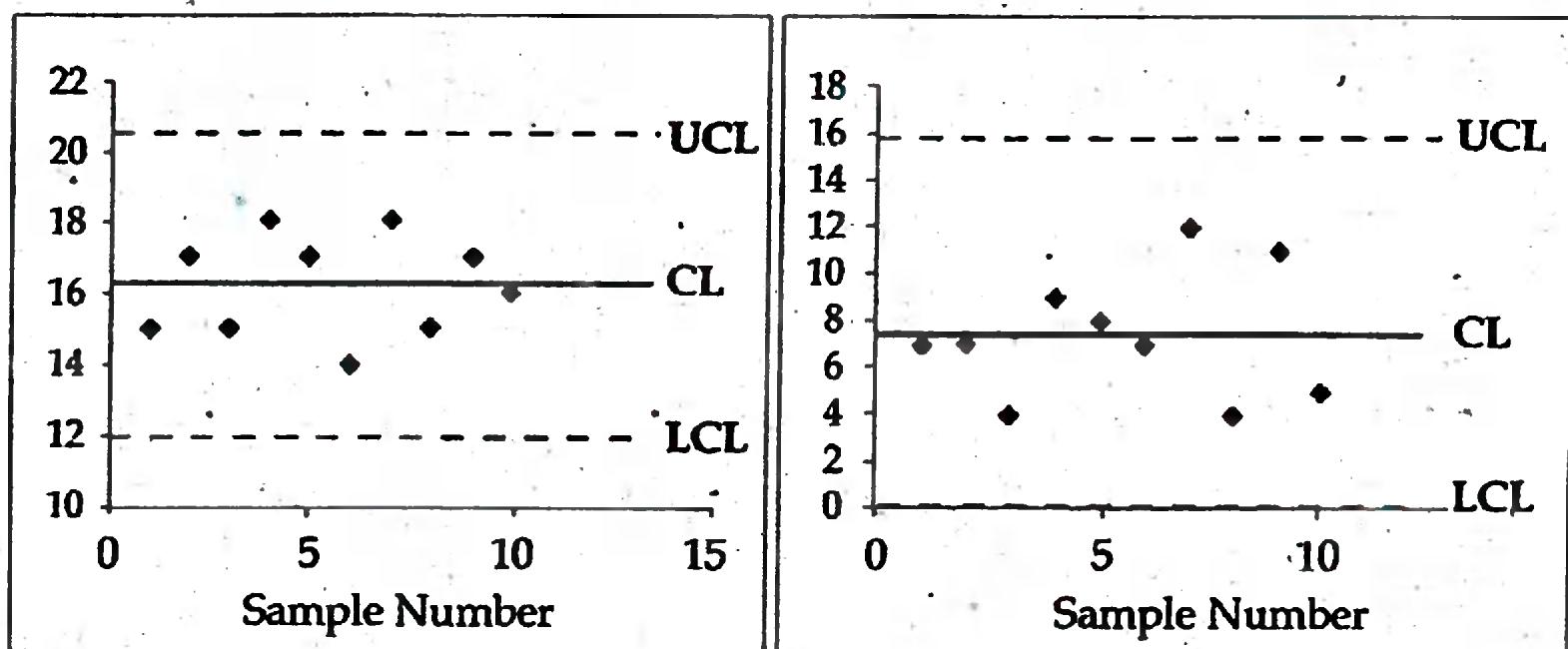


Fig. 17.6. Control Chart for Mean.

Fig. 17.7. Control Chart for Range.

Conclusion. From the charts, it is observed that all sample points of mean and range fall within the control limits, so the process is under statistical control.

Example 17.7.5. A company manufactures a product that is packed in one kg tins; it utilizes automatic filling equipment. It takes a sample of 5 cans every two hours and measures the filling in each of the 5 cans. The table gives the measurements of filling (grams) in the last 5 samples. Set up control charts for mean and range and state whether the process is under control or not. (Given $A_2 = 0.58$, $D_3 = 0$, $D_4 = 2.115$).

Sample No.	Individual measurements				
	1	2	3	4	5
1	1001	1002	1000	998	999
2	999	998	1001	998	999
3	995	1002	1003	1001	1002
4	1000	1001	999	998	1302
5	994	996	996	1000	999

Solution. Calculations for sample mean (\bar{X}) and sample range (R) are shown in following table.

Table 17.4. Calculation of sample mean and sample range.

Sample Number	Individual measurement (Sample values)					Total	Sample Mean(\bar{X})	Sample Range(R_i)
	1	2	3	4	5			
1	1001	1002	1000	998	999	5000	1000.0	4
2	999	998	1001	998	999	4999	999.0	3
3	995	1002	1003	1001	1002	5003	1000.6	7
4	1000	1001	999	998	1302	5000	1000.0	4
5	994	996	996	1000	999	4985	997.0	4
Total							4996.6	22

The mean of sample means \bar{X} is obtained as

$$\bar{X} = \frac{1}{n} \sum \bar{X}_i = \frac{4996.6}{5} = 999.32$$

$$\text{The value of } \bar{R} \text{ is obtained as } \bar{R} = \frac{1}{n} \sum R_i = \frac{22}{5} = 4.4$$

Thus, the control limits for \bar{X} chart are as follows:

$$UCL = \bar{X} + A_2 \bar{R} = 999.32 + 0.58 \times 4.4 = 1001.872$$

$$LCL = \bar{X} - A_2 \bar{R} = 999.32 - 0.58 \times 4.4 = 996.872$$

$$CL = \bar{X} = 999.32$$

Here, the entire sample means fall within the control limits; therefore the process is under statistical control.

Again, control limits for R chart are as follows:

$$UCL = D_4 \bar{R} = 2.115 \times 4.4 = 9.306$$

$$LCL = D_3 \bar{R} = 0 \times 4.4 = 0$$

$$CL = \bar{R} = 4.4$$

The control charts for mean and range are shown in following figures.

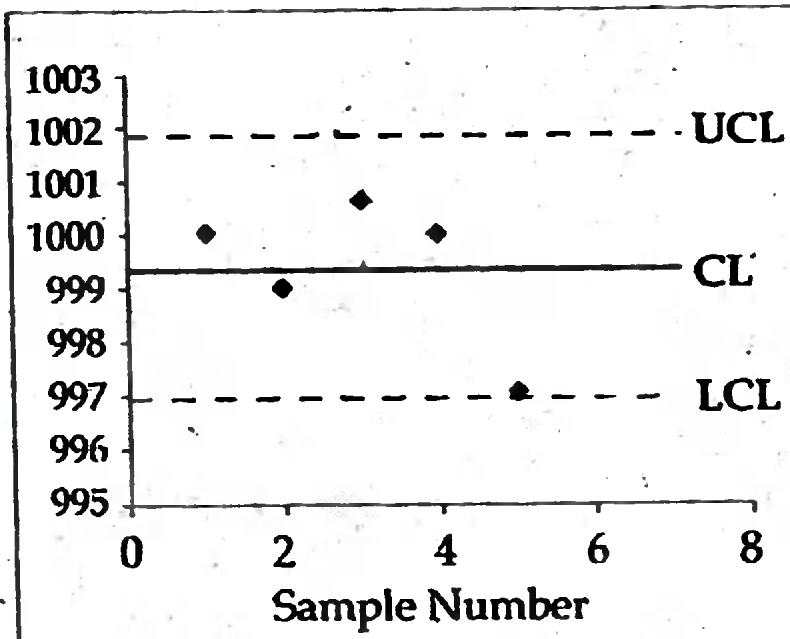


Fig. 17.8. Control chart for Mean.

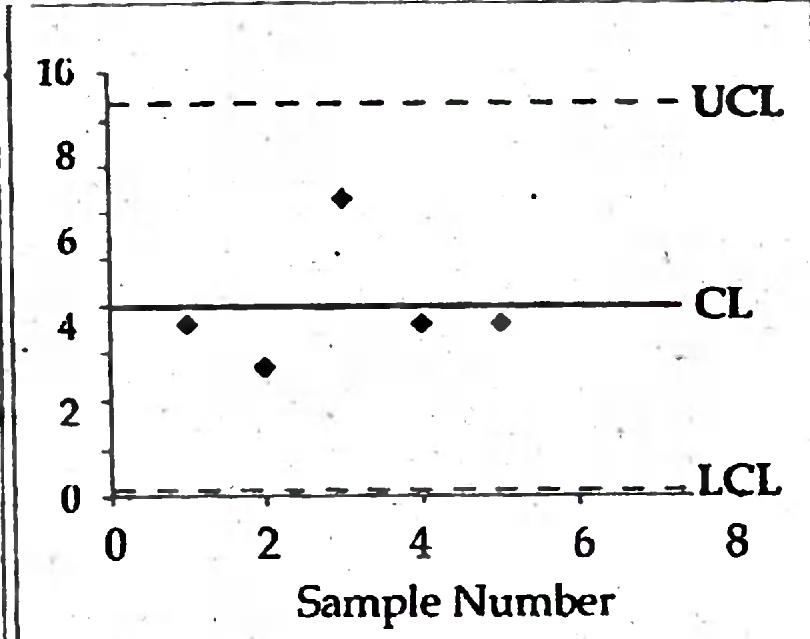


Fig. 17.9. Control chart for Range.

In this case too, all the sample range fall within the control limits, so the process is under statistical control.

17.7.2. Control chart for process variability or dispersion (standard deviation). In a sampling from a population with standard deviation σ , we know,

$$E(s^2) = (n-1)\sigma^2/n \text{ and } E(s) = C_4\sigma \text{ where } C_4 = \sqrt{\frac{2}{n} \left(\frac{n-2}{2} \right)}$$

Hence in sampling from normal population, we have

$$\text{Var}(s) = \left[\frac{n-1}{n} - C_4^2 \right] \sigma^2 \text{ and } SE(s) = C_3 \sigma.$$

$$SE(s) = 3C_3.$$

Thus UCL, CL and LCL for standard deviation are given by

$$UCL = E(s) + 3 SE(s) = (C_4 + 3C_3)\sigma = B_2\sigma$$

$$CL = E(s) = C_4\sigma$$

$$LCL = E(s) - 3 SE(s) = (C_4 - 3C_3)\sigma = B_1\sigma$$

If the value of σ is not specified or known, then we use its estimate, based on \bar{s} and estimate the σ as $\sigma = \bar{s}/C_4$. In this case,

$$UCL = \bar{s} + 3(C_3/C_4) \bar{s} = (1 + 3C_3/C_4) \bar{s} = B_4 \bar{s}$$

$$CL = \bar{s}$$

$$LCL = (1 - 3C_3/C_4) \bar{s} = B_3 \bar{s}$$

However, the control chart for dispersion or variability is not commonly used to check whether the process is under control or not.

17.8. Control Chart for Attributes

As an alternative to \bar{X} and R charts, we have the control chart for the attributes which can be used for quality characteristics of product. For example, characteristics which can be observed only as attributes by classifying an item as defective or non-defective that means conforming to specifications or not is justified in terms of this attribute.

There are two control charts for attributes, (i) control chart for fraction defective (p-chart) or the number of defectives (np-chart or d-chart) and (ii) control chart for the number of defects per unit (c-chart).

Definition. Defective and defect. A defective item is one, which does not conform the specified requirements. On the other hand, the quality or attribute lack of which makes an item defective, is called defect. Or a defect is a single nonconforming quality characteristic of an item. An item may have several defects. The term defective refers to the item having one or more defects.

For example, the purpose of producing marker pen is to write with it. Suppose after production somehow its ink has dried out, so now it is not usable for writing purpose, it does not conform the requirement. Hence it is a defective marker. The lacking of ink in the marker pen is its defect.

17.8.1. p-chart or control chart for fraction defective. The p-chart is a time plot of the sequence of sample proportion of non-conforming items. This p-chart is designed to control the percentage or proportion of defectives per sample. While dealing with attributes, a process will be adjusted in statistical control if all the samples or sub-groups are ascertained to have the same population proportion P.

If d is the number of defective items in a sample of size n, then the sample proportion defective is $p = d/n$. Hence, d is a binomial variate with parameters n and P, thus,

$$E(d) = nP \text{ and } \text{Var}(d) = nPQ, \text{ where } Q = 1 - P, \text{ therefore,}$$

$$E(p) = E(d/n) = P \text{ and } \text{Var}(p) = \text{var}(d/n) = PQ/n$$

Thus, $\pm 3\sigma$ control limits for p-chart are given by,

$$E(p) \pm 3 \text{SE}(p) = P \pm 3(PQ)^{1/2},$$

Again, if P' is the given or known value of P, then, the control limits are given by

$$P' \pm 3 \{ P' (1 - P') \}^{1/2} \text{ and that of central line is given by } CL = P'.$$

The sample proportion p is estimated by using \bar{p} given by $\bar{p} = \frac{\sum \hat{p}_i}{k}$ and in this case the control limits are given by

$$LCI = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; CL = \bar{p} \text{ and } UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

17.8.2. d-chart or control chart for number of defectives. If instead of p , the sample proportions defective, we use d , the average number of defectives in the sample, and then the $\pm 3\sigma$ control limits for the d-chart are given by

$$E(d) \pm 3 SE(d) = nP \pm 3(nPQ)^{1/2}$$

If P' is the given value of P , then,

$$UCL = nP' + 3[nP'(1-P')]^{1/2}, CL = nP', LCL = nP' - 3[nP'(1-P')]^{1/2}$$

Again, if \bar{p} is used as an estimate of P , then, average number of defectives is given by, $n\bar{p}$ or (Total number of defectives)/number of samples

$$UCL = n\bar{p} + 3[n\bar{p}(1-\bar{p})]^{1/2},$$

$$CL = n\bar{p} \text{ and }$$

$$LCL = n\bar{p} - 3[n\bar{p}(1-\bar{p})]^{1/2}$$

Where, $n\bar{p}$ is the average number of defectives per sample based on all possible samples of constant size from the process.

Since p or $n\bar{p}$ cannot be negative, if LCL as given by the above formula comes out to be negative, then it is taken to be zero.

17.8.3. c - chart or control chart for number of defects. There is some situation when it is more relevant to evaluate performance by keeping track of the number of undesirable occurrences (c), such as number of defects per unit or number of complaints received per 100 customers served or per Tk. 1000 sales.

c-chat is used when opportunities for defect in each production unit is complaint from a customer are very large while the probability of their occurrence per unit tends to be very small and constant. A Poisson distribution can describe the outcome of such a sampling process.

The steps for construction of control chart for a number of defects where the sample size is constant are as follows.

Step I. Count the defects individually in a sample.

Step II. If c stands for the number of defects counted per unit, then the mean rate with which such defects occur can be calculated as follows:

$$\bar{c} = \frac{\text{Number of defects in all samples}}{\text{Total number of samples}}$$

$$= \frac{c_1 + c_2 + \dots + c_n}{N}$$

Step III. Placing control limits using the standard deviation of the Poisson distribution, which is equal to \sqrt{c} as follows:

$$UCL = \bar{c} + 3\sqrt{c}, CL = \bar{c}, \text{ and } LCL = \bar{c} - 3\sqrt{c}.$$

Since number of defects cannot be negative, so in case the LCM becomes negative, it is considered as zero.

Step IV. The sample points c_1, c_2, \dots, c_n are plotted as points on a graph paper by taking the sample characteristics c along the y-axis and the sample number along the x-axis. The control lines are drawn accordingly.

Step V. With appropriate adjustment ensure whether a process is under control or not.

Example 17.8.1. Twenty samples each of TV tubes were taken from daily production of large output of pens and the numbers of defective tubes are recorded. On the basis of the information given below, prepare a p chart and state your conclusion.

Sample No	No. of defectives	Sample No.	No. of defectives	Sample No.	No. of defectives	Sample No.	No. of defectives
1	4	6	2	11	4	16	7
2	6	7	0	12	5	17	8
3	10	8	1	13	3	18	13
4	8	9	1	14	0	19	3
5	3	10	5	15	14	20	4

Solution. We know, the limits of control chart for p are given by

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; CL = \bar{p} \text{ and } UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

The calculations of fraction defectives for each sample are shown in following table.

Table 17.4. Fraction defectives.

Sample No.	No of defectives	Fraction defectives	Sample No.	No of defectives	Fraction defectives
1	4	0.04	11	4	0.04
2	6	0.06	12	5	0.05
3	10	0.10	13	3	0.03
4	8	0.08	14	0	0.00
5	3	0.03	15	14	0.14
6	2	0.02	16	7	0.05
7	0	0.00	17	8	0.07
8	1	0.01	18	13	0.13
9	1	0.03	19	3	0.03
10	5	0.05	20	4	0.04
Total					$\Sigma p = 1.00$

Here, central line $\bar{p} = \frac{1.00}{20} = 0.05$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.05 - 3\sqrt{\frac{0.05(1-0.05)}{100}} = -0.016 \text{ or } 0$$

$$\text{and } UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.05 + 3\sqrt{\frac{0.05(1-0.05)}{100}} = 0.116.$$

It is clear from the limits of the control chart that two points (sample 15 and 18) lie outside control limits; hence the process seems to be out of control.

Example 17.8.2. Twenty samples, each of 200 observations made over time of an electronic component are taken. For each sample the number and proportions of sampled components not conforming to standards are recorded as in following table. Construct p-chart and d-chart.

Table. Non-conforming items in Samples of 200 electronic components.

Sample No	No. of non-conforming	Sample No.	No. of non-conforming	Sample No.	No. of non-conforming	Sample No.	No. of non-conforming
1	18	6	29	11	19	16	14
2	15	7	11	12	26	17	25
3	23	8	21	13	11	18	17
4	9	9	25	14	28	19	23
5	17	10	14	15	22	20	18

Solution. Necessary calculations for the construction of control charts are shown in following table.

Table 17.5. Calculation of proportion defectives..

Sample	No. non-conforming	\hat{p}	Sample	No. non-conforming	\hat{p}
1	18	0.09	11	19	0.095
2	15	0.075	12	26	0.13
3	23	0.115	13	11	0.055
4	9	0.045	14	28	0.14
5	17	0.085	15	22	0.11
6	29	0.145	16	14	0.07
7	11	0.055	17	25	0.125
8	21	0.105	18	17	0.085
9	25	0.125	19	23	0.115
10	14	0.07	20	18	0.09
Total	182	0.91		203	1.015

p-chart

Here, population proportion is to be estimated by the average of sample proportion that is given by

$$\bar{p} = (0.09 + 0.075 + \dots + 0.09) / 20 = 1.925 / 20 = 0.0965$$

The control limits for the p-chart are as follows:

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.09625 - 0.06256 = 0.03369$$

(here, n is the number of samples)

$$CL = \bar{p} = 0.09625 \text{ and } UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.09625 + 0.06256 = 0.15881.$$

d-chart

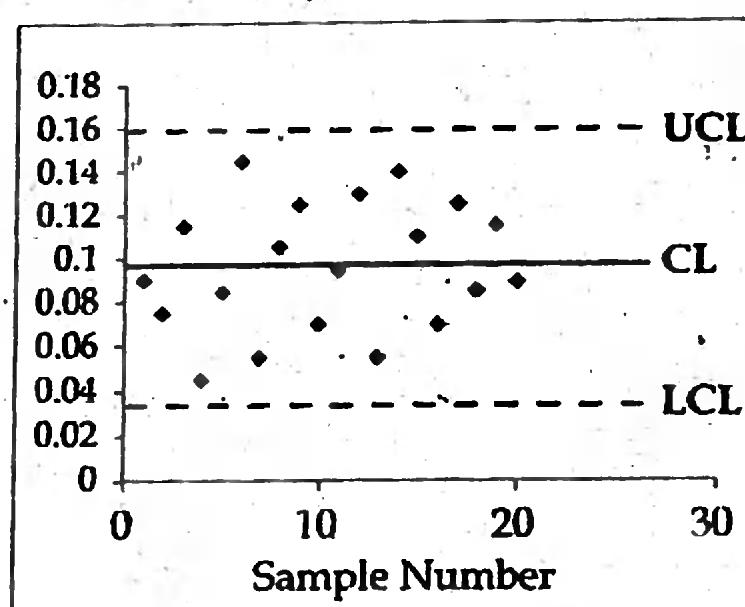
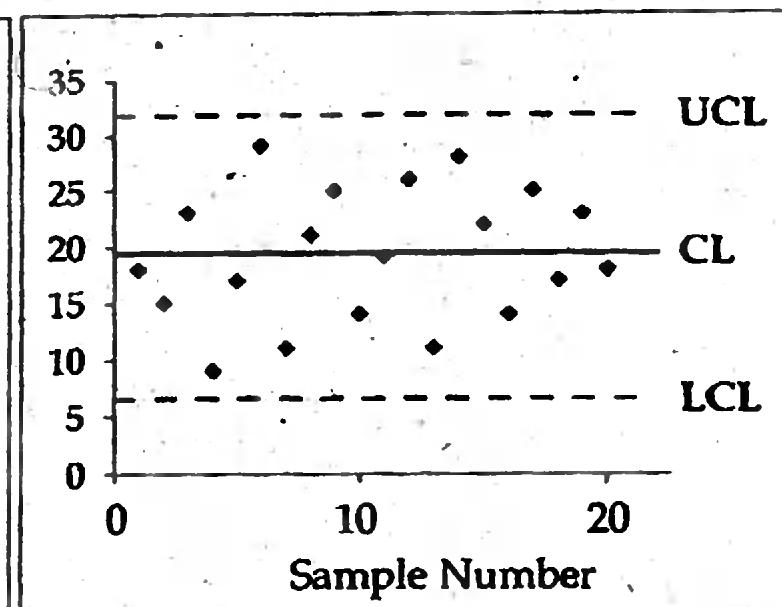
Control limits for d-chart are as follows:

Here, n = 200 (size of each sample). $CL = n\bar{p} = 200 \times 0.09625 = 19.25$ which can also be obtained directly from average number of defectives as (Total number of defectives)/number of samples = $(182 + 203) / 20 = 19.25$

$$LCL = n\bar{p} - 3[n\bar{p}(1-\bar{p})]^{1/2} = 200 \times 0.09625 - 3 \times [200 \times 0.09625(1-0.09625)]^{1/2} = 6.737$$

$$UCL = n\bar{p} + 3[n\bar{p}(1-\bar{p})]^{1/2} = 200 \times 0.09625 + 3 \times [200 \times 0.09625(1-0.09625)]^{1/2} = 31.76298.$$

Control charts for fraction defective (p) and number of defectives (d) are presented below.

Fig. 17.10. Control chart for p .Fig. 17.11. Control chart for d .

Conclusion. It is clear from the LCL and UCL of both control charts that the sample points of these factors (proportion and number of defectives) lie within the limits. So, we can say that the process is under control.

Example 17.8.3. The following table refers to the defects found at the inspection of the first 10 samples of an item each of size 100. Use the data to obtain the upper and lower control limits for proportion defectives in samples of 100.

Sample No	1	2	3	4	5	6	7	8	9	10
No. of Defectives	2	1	1	1	2	3	4	2	2	0

Solution. Here, population proportion is to be estimated by the average of sample proportion that is given by

Here, total number of defective items found = 20 and total number of items inspected is $10 \times 100 = 1000$

$$\text{Therefore, } \bar{p} = 20/1000 = 0.02$$

The control limits for the p -chart are as follows:

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.02 - 3\sqrt{\frac{0.02(1-0.02)}{10}} = -0.1128 = 0$$

$$\text{CL} = \bar{p} = 0.02 \text{ and } \text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.02 + 3\sqrt{\frac{0.02(1-0.02)}{10}} = 0.1528.$$

The fraction defectives for the samples are as follows.

Sample No.	1	2	3	4	5	6	7	8	9	10
No. of Defectives	2	1	1	3	2	3	4	2	2	0
Fraction Defectives	0.02	0.01	0.01	0.03	0.02	0.03	0.04	0.02	0.02	0

Plot of the fraction defectives on control chart is shown in following figure.

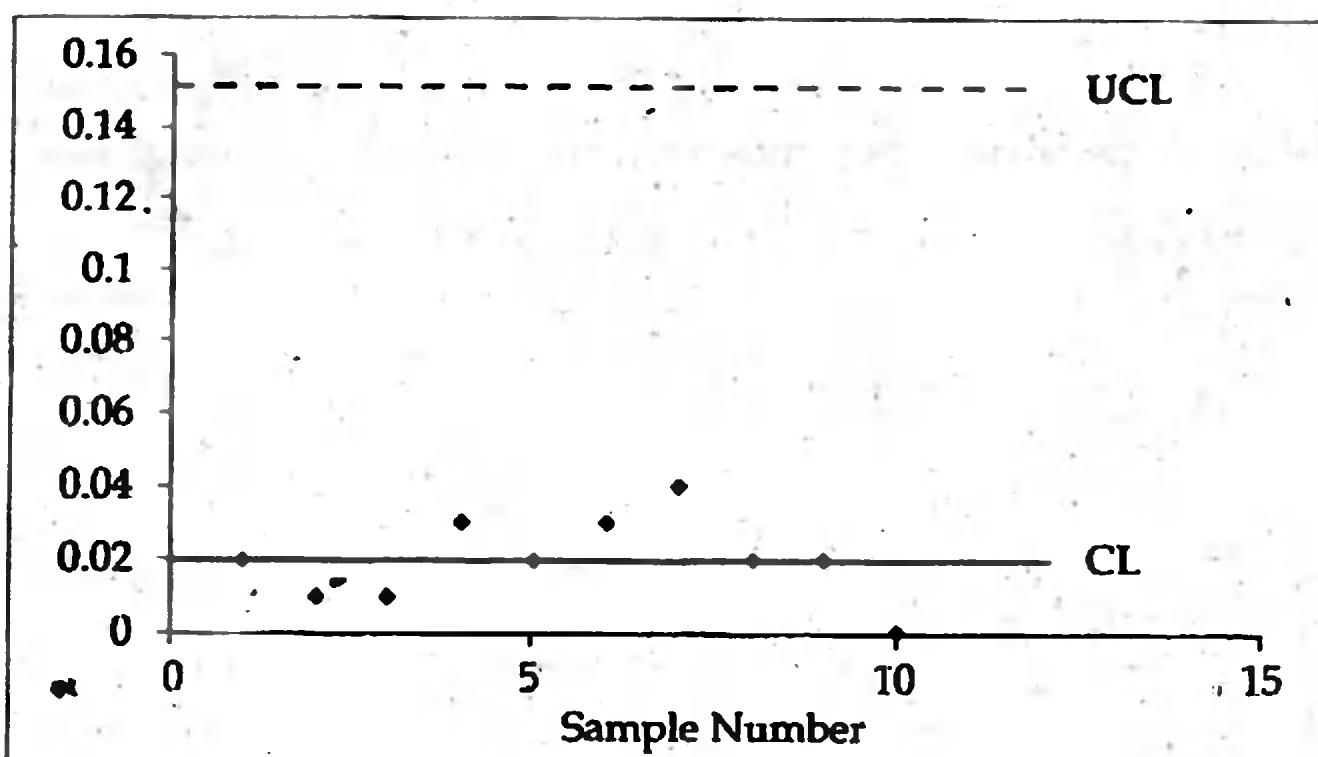


Fig. 17.12. Control chart for fraction defectives p .

It is observed that all the sample fraction defectives lie in the upper and lower control limits of the p -chart, so it can be concluded that the process is under statistical control.

Example 17.8.4. A manufacturer of textile produces bolts of cloth. Periodically, a bolt is carefully inspected, and the number of defects (imperfection) is recorded. Construct the c-chart and comment.

Number of defects in bolts of cloth are given below:

Cloth bolt	No. of defects	Cloth bolt	No. of defects	Cloth bolt	No. of defects
1	8	8	2	15	1
2	8	9	3	16	7
3	6	10	10	17	9
4	8	11	7	18	11
5	9	12	6	19	9
6	5	13	8	20	6
7	7	14	2		

Solution. Here, total number of defects is $8 + 8 + \dots + 9 + 6 = 132$

So, the average number of defects per bolt of cloth is

$$\bar{c} = (\text{Number of defects in all samples}) / (\text{Total number of samples})$$

$$= \frac{c_1 + c_2 + \dots + c_n}{N} = \frac{132}{20} = 6.6$$

The standard deviation of the number of bolts is estimated as $\sqrt{\bar{c}} = \sqrt{6.6} = 2.569$.

Therefore, $UCL = \bar{c} + 3\sqrt{\bar{c}} = 6.6 + 3\sqrt{6.6} = 14.31$, $CL = \bar{c} = 6.6$, and $LCL = \bar{c} - 3\sqrt{\bar{c}} = 0$, (because, it becomes negative, so we have to set it zero).

The control chart for number of defects (c) is shown in figure 17.13.

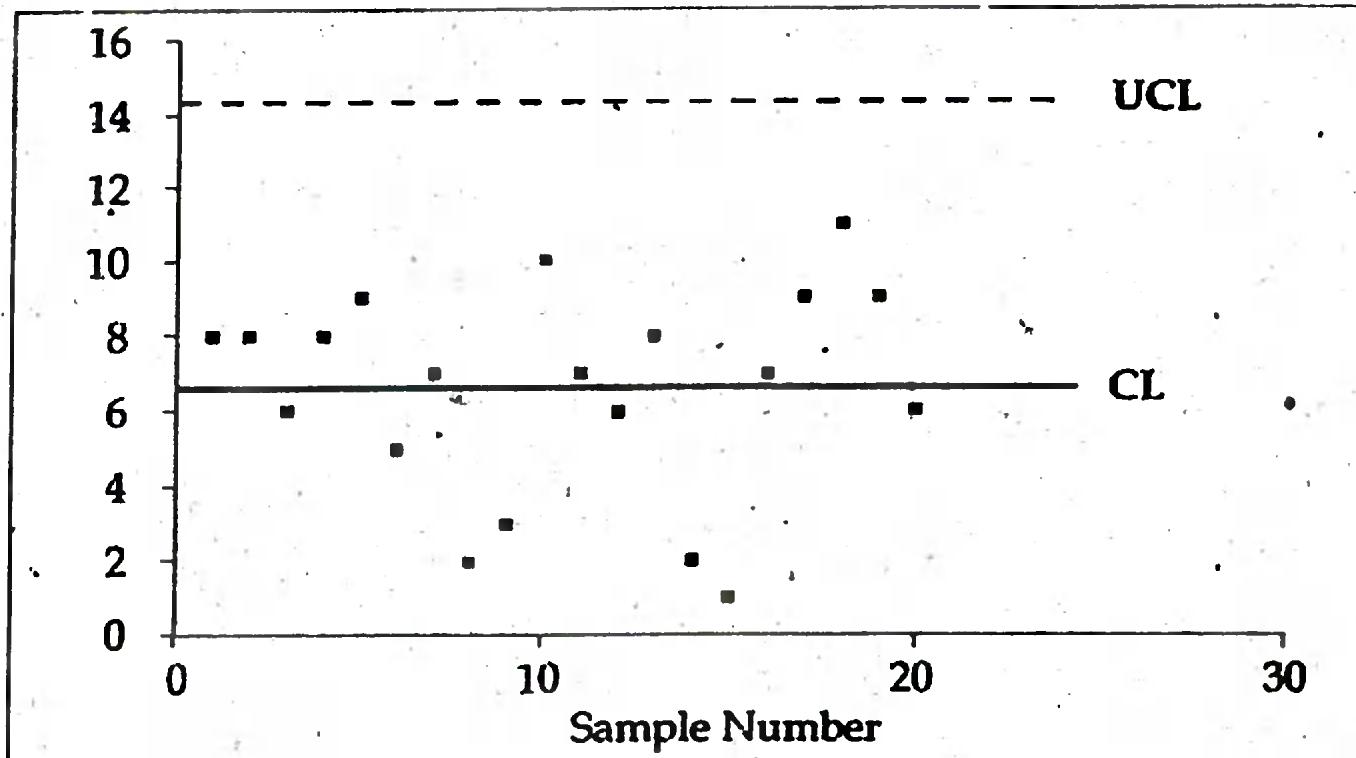


Fig. 17.13. Control chart for number of defects (c).

Inspection of c-chart suggests no cause for concern. The observations are all well below the upper control limit, and there is no evidence of an increasing number of imperfection over time. It appears, then, that the production process is under statistical control.

Example 17.8.5. Construct a control chart for c , that is, the number of defectives, from the following data pertaining to the number of imperfection in 20 pieces of cloth of equal length in a certain make of polyester and infer whether the process is in a state of control:

2, 3, 5, 8, 12, 2, 3, 4, 6, 5, 6, 10, 4, 6, 5, 7, 4, 9, 7, 3

Solution. Let c denotes the number of defectives per piece. Then,

$$\sum c = 2 + 3 + \dots + 3 = 111, \text{ and}$$

$$\bar{c} = (\text{Number of defects in all samples}) / (\text{Total number of samples})$$

$$= \frac{\sum c}{N} = \frac{111}{20} = 5.55$$

The standard deviation of the number of bolts is estimated as $\sqrt{\bar{c}} = \sqrt{5.55} = 2.356$.

Therefore, $UCL = \bar{c} + 3\sqrt{\bar{c}} = 5.55 + 3\sqrt{5.55} = 12.63$, $CL = \bar{c} = 5.55$, and $LCL = \bar{c} - 3\sqrt{\bar{c}} = -1.53 = 0$ (because, it becomes negative, so we have to set it zero).

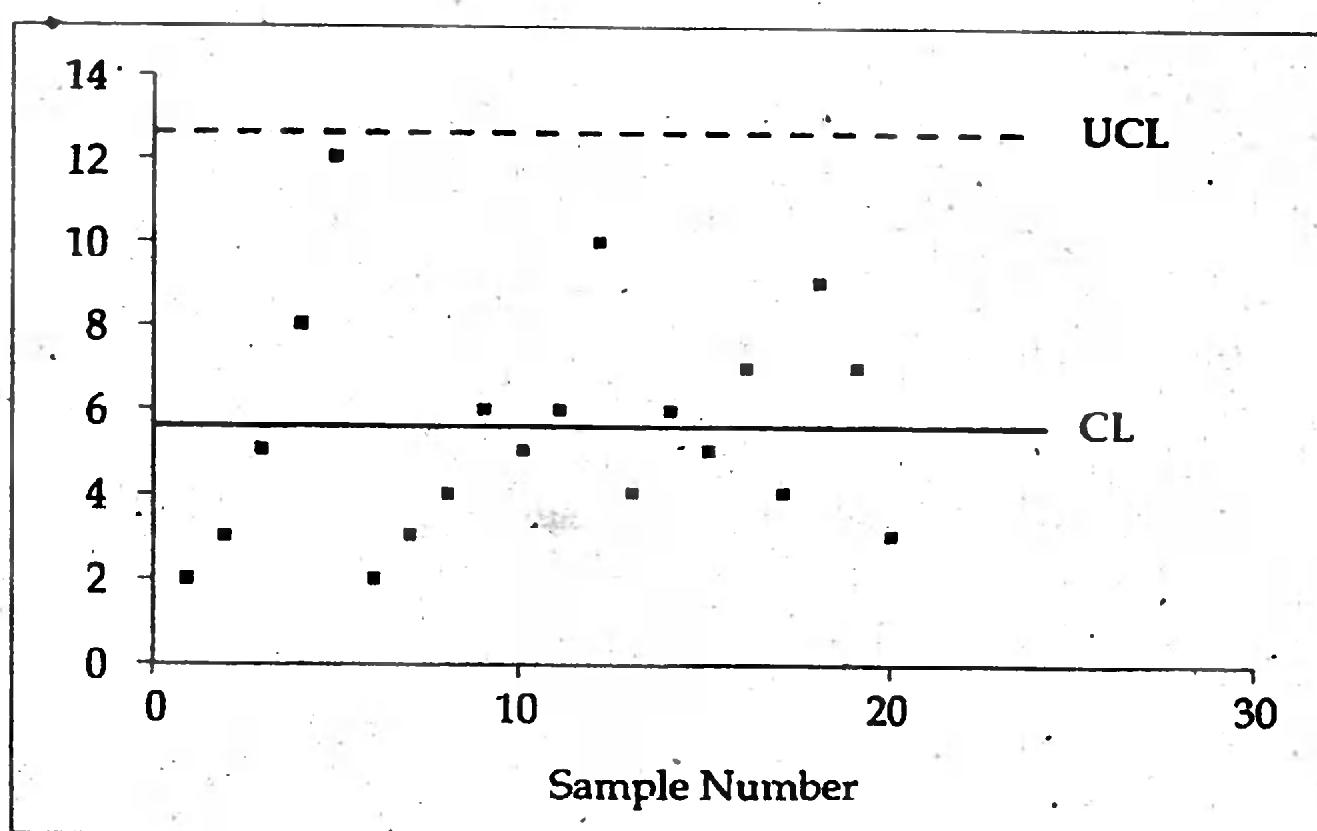


Fig. 17.14. Control chart for number of defects (c).

Since none of the sample points is falling outside the upper and lower control limits, the process is in control.

Example 17.8.6. The following data refers to visual defects found during the inspection of the first 10 samples of size 50 each from a lot of two-wheels manufactured by an automobile company:

Sample No.	1	2	3	4	5	6	7	8	9	10
No. Of defectives	4	3	2	3	4	4	4	1	3	2

Draw the p chart to show that the fraction defectives are under control.

Solution. Here, total number of defectives = $4 + 3 + \dots + 2 = 30$.

So, there are 30 defectives in 10 samples of 50 items each, hence, the average fraction defectives

$$\bar{p} = \frac{\text{Total number of defectives}}{\text{Total number of items inspected}} = \frac{30}{N} = \frac{30}{10 \times 50} = 0.06.$$

The control limits for the p-chart are as follows:

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 - 3\sqrt{\frac{0.06(1-0.06)}{10}} = -0.65 = 0$$

$$CL = \bar{p} = 0.06 \text{ and } UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 + 3\sqrt{\frac{0.06(1-0.06)}{10}} = 0.77.$$

The control chart based on sample data and the limits is constructed below:

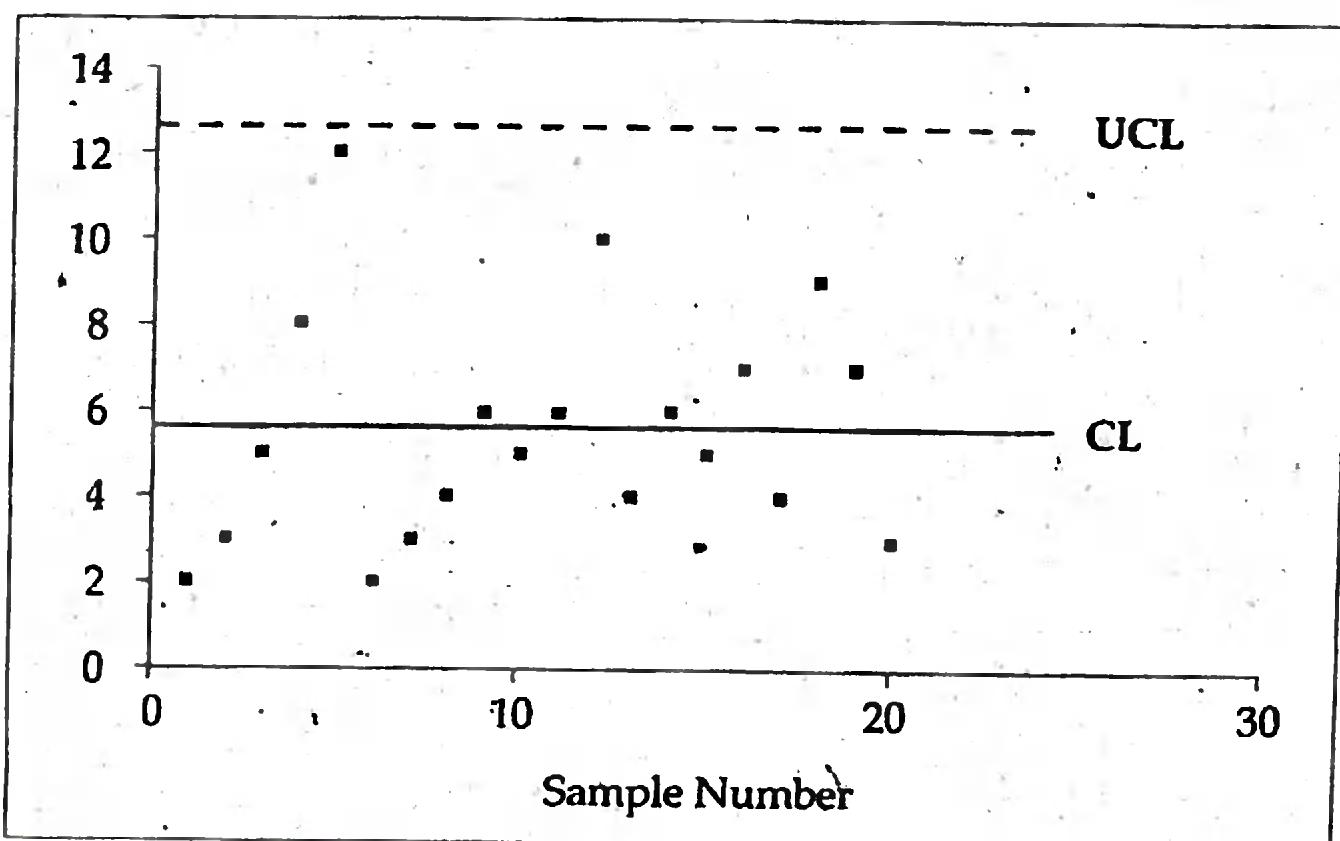


Fig. 17.15. Control chart for number of defects (c).

From the chart it is clear that all the sample points lie inside two limits, hence, it can be concluded that the quality of the products are under control.

17.9. Product Control

We have already discussed a statistical tool of controlling quality by process control system through control chart. When it is not possible or economically not feasible to exercise direct control on a process, we have another option to distinguish between good or bad products or incoming materials to test the quality of parts that a producer has to buy or procure from other sources. This statistical way of controlling quality of product is termed as product control. Product control means controlling the quality of the product by critical examination at strategic points and this is achieved through sampling inspection plans. Product control aims at guaranteeing a certain quality level to the consumer regardless of what quality the producer is maintaining level. In other words, it attempts to ensure that the product marketed by sales department does not contain a large number of defective items. This process deals with the quality of the incoming material known as acceptance sampling. For example, let us assume that a computer assembling company buys monitor of certain size for their computers from various suppliers in lots of 100 each. They are willing to accept the entire lot if there are less than 2% (say) defective monitors in the lot. This can be done by developing a sampling inspection plan to ensure that the accepted products meet the conditions.

Definition. Inspection sampling. When a sample is taken from a lot and inspected for determining the proportion of defective item in it. Thus the decision is taken on the basis of decision rule whether the lot is to be accepted or rejected.

17.9.1. Types of sampling inspection procedures. The acceptance or rejection of a given lot is based on inspection of a sample drawn from a the lot. Usually the decision is based on the number or proportion of defectives present in the sample according to attributes under consideration. But the items can also be classified as defective or non-defectives on the basis of measurable quality characteristics as well. For example, an item heavier than a fixed weight or shorter than a fixed length can be classified as defectives. Hence, there are two types product control or sampling inspection procedures that are generally used; these are (i) sampling inspection by attributes and (ii) sampling inspection by variables. Although both types of inspection lead to the same conclusion and each method possesses some inherent limitations and advantages over other, but the sampling inspection by attributes is mostly and widely practiced. In other words, sampling inspection plan for variables do exist, but are not so prevalent as for attributes. That is why only the acceptance sampling by attributes is discussed here. In this type of inspection, items are classified as defective or non-defective. Also the acceptance or rejection of a lot depends on the fraction defectives in the sample. The decision about the lot can be of two types, viz. (i) acceptance-rejection type, and (ii) acceptance-rectification type. In the first type; the lot is either accepted or rejected, if the lot is accepted, then the defective items in the sample are replaced by non-defectives and then the lot is passed for marketing. This is called Dodge and Romig convention (H F Dodge and H G Romig) that like Walter Shewhrtz did a lot of works on product control.

The points by which inspection by attributes contrasted with inspection by variables are listed below:

- i) The major advantage of inspection by variables over inspection by attributes is that inspection by variables needs fewer item for inspection for a given degree of accuracy.
- ii) The decision about the acceptance or rejection of a lot is more reliable if it is based on inspection by variables than by attributes for the fixed sample size.
- iii) The main advantage of inspection by attributes over inspection by variables is that it requires less accuracy of measurement, less skill, less time and less sophisticated instrument for measurements.
- iv) In case of inspection by attributes, selection and installation of sound sampling plan is easier than by variables.

- v) Percent defective is usually an appropriate measure and its appropriateness is not affected by shifting between variables and attributes, hence, inspection by attributes is widely used.

17.9.2. Acceptance Sampling. It is true that in most production situations, complete inspection of an entire batch of raw materials, purchased parts or finished products is impractical from the view points of time and cost. Instead a sample of the batch is inspected, and the decision to accept or reject the entire batch is based on the quality of sample. Thus, acceptance sampling involves taking random samples from a lot batch for inspecting their quality against predetermined standards to take decision whether to accept or reject the lot. Rejected lots may either be returned to the supplier or be inspected 100 percent at the producer's expense, followed by replacement of defective units by good ones. However, since the judgment is based on a sample, there is always a risk of making an error of accepting a bad lot or rejecting a good lot.

Definition. Acceptance sampling. This is a procedure for deciding whether to accept or to reject a lot of product on the basis of quality of a sample taken form that lot.

The flowchart or steps of acceptance sampling are as follows:

- i) Receive a lot of products
- ii) Select a sample from the lot
- iii) Inspect the sample for quality
- iv) Compare the results with specified quality characteristics, and
 - a. Accept the lot if the sample conforms the specified quality
 - b. Otherwise, reject the lot and decide whether to return to the supplier or be inspected 100% followed by the replacement of all defectives by good ones.

19.9.3. Advantages of Acceptance sampling. It has been mentioned that there is always a risk of making error of accepting a bad lot by acceptance sampling, however, there are some advantages of this sampling too, these are:

- i) It minimizes the total expected cost resulting from sampling errors,
- ii) It is only the sampling procedure when testing is destructive,
- iii) This procedure involves less product damage due to less handing and testing, so it is less expensive
- iv) Less manpower is required for quality inspection
- v) Corrective measures may be taken for an ongoing process as and when required
- vi) This presumes an agreement between producer and consumer as to what constitutes 'good' or 'bad' quality and the acceptable risk of error for each quality level.

17.10. Types of Sampling Plan

There are so far three types of sampling plans which are commonly in acceptance sampling procedure, these are (i) single sampling plan, (ii) double sampling plan and (iii) multiple or sequential sampling plan. Brief descriptions of each of these plans are provided below.

17.10.1. Single sampling plan. When a decision whether to accept or reject a lot is made on the basis of only one sample, the acceptance sampling plan is called a single sampling plan. This is the simplest type of sampling plan. The following three parameters are to be specified in a single sampling plan (i) number of items N in the lot from which the sample would be drawn, (ii) sample size n to be drawn from the lot, (iii) the acceptance number of defectives c , this acceptance number is the maximum allowable number of defective items in the sample. After specifying the above mentioned three parameters, a single sampling plan is undertaken following the rule as:

- i) Inspect a sample of size n from the lot
- ii) Count the number of defectives, let d be the number of defectives found in the sample,
- iii) Accept the lot if $d \leq c$, or
- iv) Reject the lot if $d > c$

Usually, a single sampling plan is specified, for example, as $N = 100$, $n = 15$, $c = 2$. The interpretation of the numbers is 'Take a random sample of size 15 from a lot of 100 items, if the sample contains more than 2 defectives, reject the lot, otherwise accept the lot'.

17.10.2. Double sampling plan. In a single sampling plan, decision is taken on the basis of a single sample, while in case of double sampling plan, the decision whether to accept or reject a lot is taken on the basis of a second sample, if necessary, in addition to the first sample. In this plan, a lot may be accepted at once if the first sample is good enough or reject at once if the first sample is bad. Again, if the first sample is neither good enough nor bad enough, a second is taken and the decision is taken on the basis of the results of the first and second sample combined. In a double sampling plan, the following parameters are to be specified:

- i) Total number of items in a lot N
- ii) Size of the first sample n_1
- iii) Acceptance number for the first sample or the maximum number of defectives to be allowed in the first sample, c_1
- iv) Size of the second sample n_2
- v) Size of two samples combined $n_1 + n_2$
- vi) Acceptance number of the two samples combined or the maximum number of defectives to be allowed in the two samples, c_2 .

After specifying the above-mentioned five parameters, a double sampling plan is undertaken following the rule as:

- i) Inspect a sample of size n_1 from a lot of N items
- ii) Count the number of defectives, let d_1 be the number of defectives found in the first sample,
- iii) Accept the lot if $d_1 \leq c_1$, or reject the lot if $d_1 > c_2$
- iv) If $c_1 < d_1 \leq c_2$, inspect a second sample of size n_2 .

Let d_2 be the total number of defectives in the combined sample of $n_1 + n_2$ items, then,

- v) Accept the lot if $d_2 \leq c_2$, or
- vi) Reject the lot if $d_2 > c_2$.

Advantage of double sampling plan. A double sampling plan has two possible advantages over a single sampling plan:

- i) It may reduce the total cost of inspection. Consequently, in all cases, in which a lot is accepted or rejected on the first sample, there may be considerable saving in total inspection cost. It is also possible to reject a lot without completely inspecting the entire second sample.
- ii) A double sampling plan has the psychological advantage of giving a second chance to inspect the second lot of items because to some people, especially the producer, it may seem unfair to reject a lot on the basis of a single sample.

17.10.3. Sequential sampling plan. Just as double sampling plan may postpone the decision on acceptance or rejection until a second sample has been taken, other plans may permit to draw a few more number of samples before a decision is reached. Thus the plans that permit three or more number of samples to take decision about acceptance or rejection of a lot are referred to as a multiple or sequential sampling plan. However, since such plans are quite complicated and rarely used in practice, it is not described here.

17.11. Factors Related to An Acceptance Sampling Plan

The factors related to an acceptance sampling plan which play important role in taking decision under a sampling plan are (i) The operating characteristic curve, (ii) Producer's risk and Consumer's risk, (iii) Acceptance quality level (AQL) and Lot tolerance percent defective (LTPD), (iv) Average outgoing quality (AOQ).

17.11.1. The Operating Characteristic (OC) Curve. The OC curve describes how well an acceptance plan discriminates between a good lot and a bad lot. This is a curve of the combination of n (sample size) and c (acceptance number) pertaining to a specific plan. The curve shows the

percentage of lots that would be accepted if a large number of lots of any specified quality are inspected.

Construction of OC curve. Construction of OC curve requires the agreement between the producer and consumer regarding the 'good' or 'bad' quality and the amount of risk that each side is ready to accept as a result of sampling error. This agreement should be set before a sampling plan is undertaken. The information of this agreement shall help in determining the sample size n and acceptance number c of a sampling plan that can be applied to incoming lots to distinguish between.

A family of curves can be drawn for every set of sample of size n and corresponding acceptance level c . For any particular set of n and c values, OC curve shows as to how well the sampling plan is able to distinguish between good and bad lots. The OC curve is illustrated below with two examples.

17.11.2. Producer's risk and Consumer's risk. A producer is a person who produces goods for the purpose of consumption by others while a consumer is a person who purchases goods for his own consumption and not for sale to others directly or indirectly. Since under an acceptance sampling plan, a decision is made as to whether to accept or reject a lot on the basis of a sample, there is possibility of (i) rejecting a lot which was actually acceptable according to the specified quality standard, such risk is termed as producer's risk, while, (ii) the risk of accepting a lot with poor quality by the buyer is termed consumer's risk or buyer's risk. Suppose, the process is set for a fraction defective \bar{p} , known as producer's process average, thus, the probability of rejecting a lot having $100\bar{p}$ as process average percent defectives is known as producer's risk P_p , usually denoted by α , which corresponds to a type I error in hypothesis testing. Again, if P_t is the proportion of defectives in the lot which can be tolerated, thus, the probability of accepting a lot with fraction defectives P_t is known as consumer's risk P_c , usually denoted by β , which corresponds to a type II error in hypothesis testing. Dodge and Romig suggested that β can be considered as 10%. Rejecting a satisfactory lot creates a risk for the producer due to unwarranted inspection and replacement costs. Thus a large sample may be helpful to minimize the producer's risk. On the other hand, accepting poor quality lots creates a risk for the consumer of the lot because he bears the cost.

Definition. OC curve. The graph showing the probability an acceptance-sampling scheme will accept a lot as a function of input quality of the lot.

Definition. Producer's Risk: The probability of rejecting a good lot through sampling inspection, i.e.

$$P_p = P(\text{rejecting a lot having } 100\bar{p} \text{ defectives}) = \alpha$$

Definition. Consumer's Risk. The probability of accepting a bad lot through sampling inspection is

$$P_c = P(\text{accepting a lot having } P_d \text{ defectives}) = \beta$$

Definition. Acceptance Quality Level. The percent of defective items in a lot under inspection that leads to acceptance of the lot.

Definition. Average Outgoing Quality. The average fraction of defective items remaining in a lot accepted by a sampling plan

An inspection plan can easily be constructed if the consumers and producers specify these probabilities and also the proportion of defectives above which a lot is considered to be bad and the proportion of defectives below which a lot is considered to be good.

17.11.3. Acceptance quality level (AQL) and Lot tolerance percent defective (LTPD). Acceptance quality level is the minimum level of percent defectives in a lot that can be considered satisfactory by the consumer, that means, a lot should be accepted by the consumer if the percent of defectives found in a lot is below this level. For example, if the acceptable quality is 40 defectives in a lot of 2000 items, then the AQL is $40/2000 \times 100 = 2$ percent. AQL is used to measure the producer's risk P_p .

On the other hand, in order to measure the consumer's risk, maximum percentage of defective items in a lot that the consumer wishes to accept is also defined. This is the quality level of a lot that is considered as bad. This is expressed as percentage defectives in a lot that is considered as the most unsatisfactory or bad quality the consumer can tolerate and known as lot tolerance percent defective (LTPD). For example, if an unacceptable quality level is 70 defectives in a lot of 1000, then LTPD is $70/1000 \times 100 = 7$ percent. LTPD is used to measure the consumers risk P_c .

The actual level of AQL and LTPD are decided by the negotiation between the consumer and the producer.

17.11.4. Average Outgoing Quality (AOQ). The average fraction of defectives remaining in the lot finally accepted is known as the average outgoing quality and the maximum average fraction of defective items in a product finally accepted is known as average outgoing quality limit (AOQL). Suppose, a sample of n items is drawn from a lot of size N , and p_a be the probability of accepting the lot of average outgoing quality level p ,

$$\text{then AOQL} = \frac{p(N-n)p}{N}$$

However, if the sampling fraction n/N is negligible, then $\text{AOQL} = p \times p_a$

The curve obtained by plotting AOQL against p is known as AOQL curve.

Example 17.11.1. Compute producers risk for the following single sampling plan from lots of 2000 items with $AQL = 0.005$. Also use binomial distribution to approximate the consumer's risk of the schemes. (i) $n = 150, c = 1$, (ii) $n = 150, c = 2$.

Solution. Given $AQL = 0.005$.

(i) We know the producer's risk is the probability of rejecting lot for a given AQL. The probability of accepting the lot can be obtained by using binomial distribution, which is given by

$$Pr(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}; \quad r = 0, 1, \dots, n$$

For $c = 1$, the lot will be accepted by the consumer, if he finds 0 or 1 defectives.

So, the probability of accepting the lot is $P(X \leq 1) = P(x = 0) + P(x = 1)$.

Example 17.11.2. Suppose the quality control manager of a big firm is negotiating a contract for 5000 tube lights with a company which is reputed for its high quality products, but sometimes do not found as perfect. The company claims that it can produce bulbs with rates of defects below 1%, a level that is acceptable to the manager of firm. Now, the manager decides to pick 100 bulbs randomly and considers acceptance number as $c = 1$. Compute consumer's risk and producer's risk.

Solution. If $p = 0.01$ is producer's true rate of defects, the probability that the lot will be rejected can be computed by using distribution for given $n = 100$ and $p' = 0.01$. We know the binomial probability distribution is

$$P(X = c) = \binom{n}{c} p^c (1-p)^{n-c}; \quad c = 0, 1, \dots, n$$

The lot will be accepted if number of defectives found are less than or equal to $c = 1$, so the probability of accepting the lot is given by.,

$$\begin{aligned} P(X \leq 1) &= P(x = 0) + P(x = 1) = \frac{100!}{0!100!} 0.01^0 0.99^{100} + \frac{100!}{1!99!} 0.01^1 0.99^{99} \\ &= 0.366 + 0.369 = 0.735 \end{aligned}$$

Example 17.11.3. Draw an OC curve for a sample of size 15 with acceptance level 0 considering 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 25% defectives in the lot.

Solution. For the given values of n , p and c , the probability of acceptance is computed using binomial distribution as

$$P(x = c) = \binom{n}{c} p^c (1-p)^{n-c}; \quad c = 0, 1, \dots, 15$$

Thus, for $n = 15$, $p = 0.01$, $c = 0$, the probability is given by

$$P(x = 0) = \binom{15}{0} 0.01^0 (1 - 0.01)^{15} = 0.860058 = 0.8601 \text{ (approx)}$$

In this way, the values of probabilities acceptance computed using binomial distribution for different values of $n = 15$, and $c = 0$ for the given values of p (percentage of defective lots) are shown in following table.

Table 17.6. Binomial probabilities of accepting the lot of sample of size $n = 15$ and $c = 0$

Percentage of defective in a lot	Probability of accepting the lot	Percentage of defective in a lot	Probability of accepting the lot
1	0.8601	10	0.2059
2	0.7386	15	0.0874
3	0.6333	20	0.0352
4	0.5421	25	0.0134
5	0.4633		

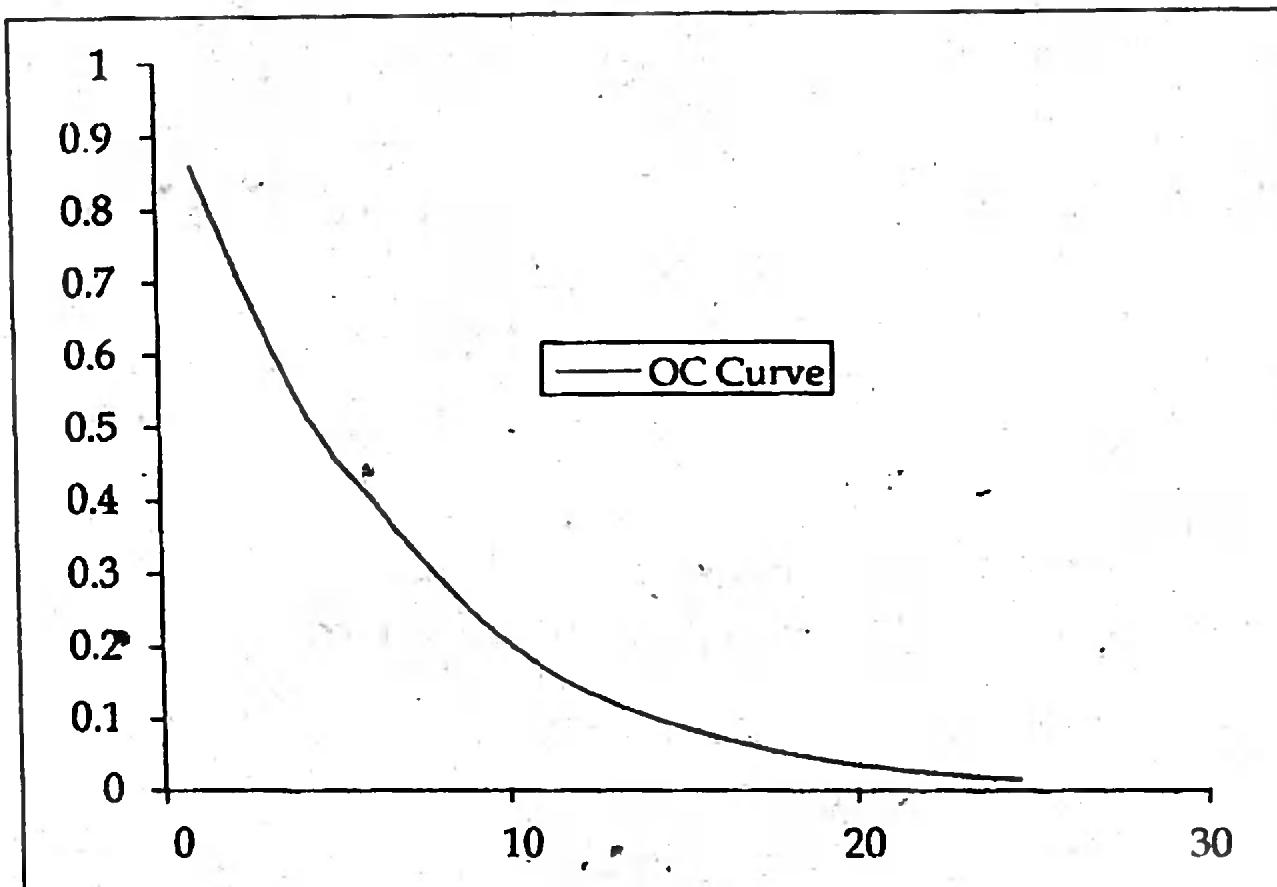


Fig. 17.16. OC curve.

From figure it can be seen that for 2% defectives, the probability of acceptance is about 0.74, similarly, if the percentage of defectives in the lots were 15 percent, then the probability of accepting the lot is about 0.10

(approx). In other words, if the percentage of defective acceptable is increased from 2 percent to 15 percent, then probability of accepting a lot is only 10 percent.

Example 17.11.4. For the sampling plan $N = 1200$, $n = 64$, $c = 1$, determine the probability of acceptance of the lots with (i) 0.5% defectives, (ii) 0.8% defectives, (iii) 1% defectives, (iv) 2% defectives, (v) 4% defectives, (vi) 10% defectives. Also draw an OC curve.

Solution. Here, the sample size is large, and the percentage of defectives is small, so the Poisson approximation of binomial distribution would be the appropriate distribution.

The probability function of Poisson distribution with parameter λ is given by

$$P(x = c) = \frac{e^{-\lambda} \lambda^c}{c!}; \quad c = 0, 1, \dots, \infty$$

Hence, if the lot contains 0.5% defectives, the samples from it will also have an average of 0.5% defectives. Thus, in a sample of 64, the average number of defectives will be $(64 \times 0.5)/100 = 0.32$, if the lot contains 0.8% defectives, the average number of defectives will be $(64 \times 0.8)/100 = 0.512$, and so on. Since, $c = 1$, the lot is to be accepted under the sampling plan if the lot contains 0 or 1 defective. The cumulative probability obtained by using Poisson distribution containing 0 or 1 defectives with different averages (on drawing a sample of 64) are shown in following table.

Table 17.7. Poisson probabilities of accepting the lot of sample of size $n = 64$ and $c = 0$ or 1 .

Percentage of defective in the lot	Average number of defectives	$p(c = 0)$	$p(c = 1)$	Prob of acceptance $p(c = 0) + p(c = 1) = p_a$
0.5	0.320	0.73	0.23	0.96
0.8	0.512	0.60	0.31	0.91
1	0.640	0.53	0.35	0.88
2	1.280	0.28	0.36	0.64
4	2.560	0.08	0.21	0.29
10	6.400	0.002	0.01	0.012

The value 0.96 represents the probability of drawing a sample of 64 with 0 or 1 defectives from a lot known to have 0.5% defectives. Thus, we can say that such a sample will enable acceptance of 96% of lots containing 0.5% defectives. Which again mean that, if 1000 such lots are submitted for inspection under the given sampling plan, on an average 960 of the lots will be accepted and the rest 40 will be rejected.

If we plot the probabilities of acceptance p_a on the y-axis and percentage of defectives in the lots along x-axis, and join the various points, the curve so found is called operating characteristics curve of sampling plan. From the OC curve we can easily obtain the probability of rejection of lot ($1 - p_a$) for a specified proportion of percent defectives. The OC curve showing the probabilities of acceptance is presented below.

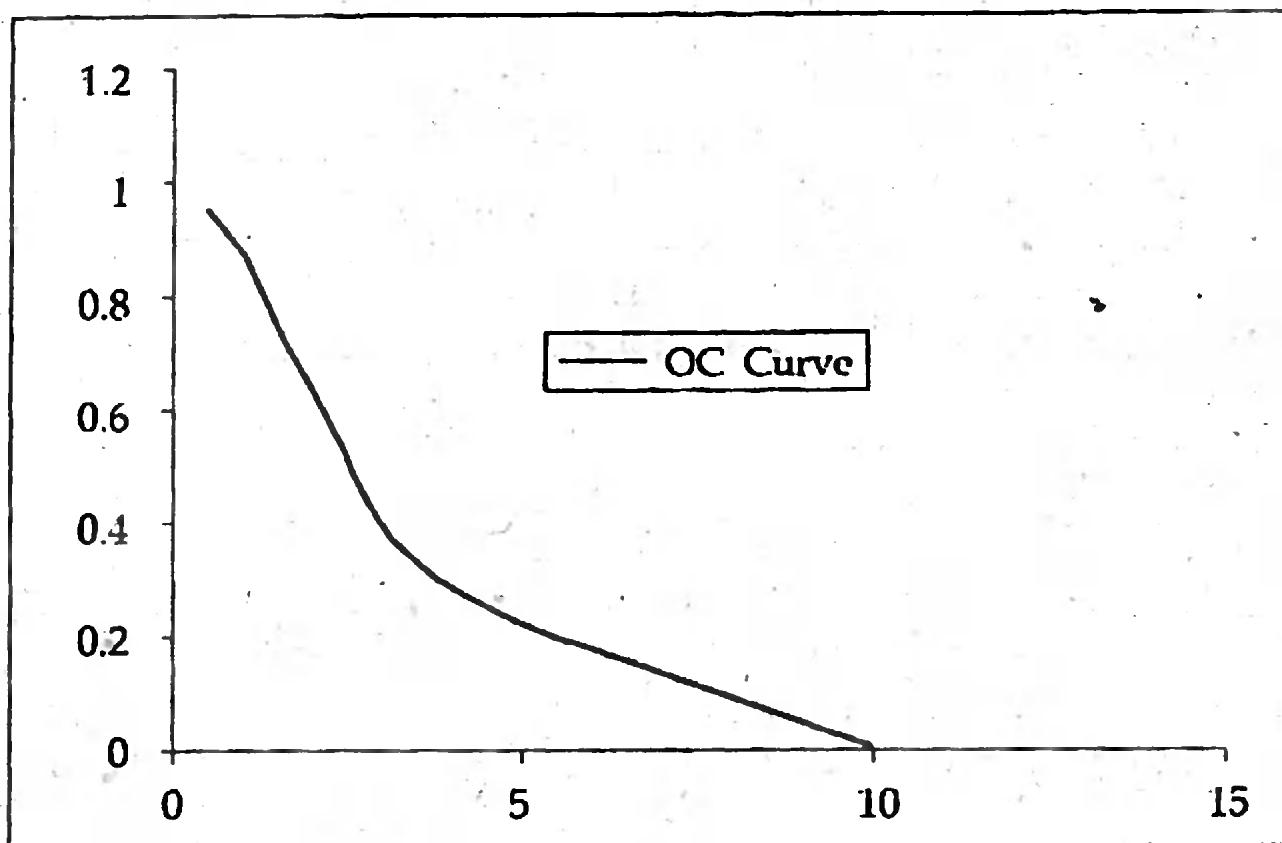


Fig. 17.17. OC curve.

Questions

1. What do you mean by statistical quality control? Discuss briefly its use in industry. Also discuss the causes of variations in industrial products.
2. What is control chart? Why is it called Schewartz control chart?
3. Explain the justification of using 3σ limits in a control chart. Also discuss how can you construct the control chart for mean and range.
4. Differentiate between random variations and assignable variations.
5. What is process control? Why R-chart is prepared?
6. What is statistical quality control? State the uses and State the limitations of statistical quality control.
7. What do you mean by control chart? Explain the methods of construction of control chart.
8. Explain the term Statistical Quality Control. How is the process control achieved with the help of control chart? What are the fundamental steps in the construction of quality control chart?
9. What are the control limits for mean, range, fraction defectives and number of defects.
10. How do control charts reveal that the process is out of control?

11. Distinguish between process control and product control.
12. Define defectives and defects with examples. Also explain how can you construct control chart for fraction defectives and number of defects.
13. Give four types of patterns that indicate that a process is out of control. Give examples where each might arise.
14. Distinguish between chance causes and assignable causes of variations. Also distinguish between control limits and tolerance limits.
15. What are the various types of control charts known to you? Briefly explain any two of them.
16. How do you set the control limits for \bar{x} , R, σ , c, d and p charts (in usual notations)? Also explain, with the help of example, why \bar{x} and R should be used simultaneously for taking decision.
17. Differentiate between p chart and c chart in context of quality control.
18. What is acceptance sampling? Point out the role of operating characteristics curve. Discuss different types of acceptance sampling plan.
19. How does single sampling plan differs from double sampling plan. Explain the concept of single sampling plan in details.
20. Explain how would you set up the control limits for a control chart. Also explain the rationale behind setting of control limits.
21. What do you mean by OC curve? Also explain how would construct an OC curve.
22. Distinguish between producer's risk and consumer's risk. Also distinguish between AQL and LTPD.
23. What do you mean by average outgoing quality. Also discuss how can you draw an OC curve.
24. Briefly explain the different types of acceptable sampling plans known to you.
25. Why is it impractical to inspect an entire lot of input from a producer?
26. What is the significance of the acceptance number c in single sampling plan?
27. Distinguish between
 - a. Assignable causes of variation and chance causes of variation
 - b. Upper control limit, lower control limit and central limit
 - c. Control chart for variables and control chart for attributes
 - d. R chart and σ chart
 - e. Acceptance sampling and rejection sampling
 - f. Single sampling plan and double sampling plan
 - g. Acceptance quality level and lot tolerance percent defectives.
 - h. Average outgoing quality
28. Explain how can you construct:
 - a. Control chart for Mean

- b. Control chart for Range
 - c. Control chart for standard deviation
 - d. Control chart for fraction defectives
 - e. Control chart for number of defects
 - f. Control chart for number of defectives
29. Explain the following terms in connection with sampling inspection plans
- a. Average outgoing quality level (AOQL)
 - b. Lot tolerance percent defectives (LTPD)
 - c. Producer's risk
 - d. Consumer's risk.
30. Distinguish between the following with examples,
- a. Chance causes and assignable causes of variations
 - b. Defects and defectives
 - c. Control chart for variables and control chart for attributes
 - d. P-chart and c-chart

Exercises

31. On the basis of the information given below, find LCL, CL and UCL for mean chart and range chart, where possible, for each case
- a. $n = 6, \bar{x} = 546, \bar{R} = 84$ (Ans. 505.43, 546, 586.57)
 - b. $n = 9, \bar{x} = 26.7, \bar{R} = 5.3$
 - c. $n = 17, \bar{x} = 138.6, \bar{R} = 15.1$
 - d. $n = 4, \bar{x} = 84.2, \bar{R} = 9.6$
 - e. $n = 22, \bar{x} = 8.1, \bar{R} = 7.4$
 - f. $n = 5, \bar{x} = 1367.5, \bar{R} = 427.5$
 - g. $n = 12, \bar{x} = 16.4, se\ of\ \bar{x} = 1.2$
 - h. $n = 7, \bar{x} = 22.7, \sigma = 12.6$
 - i. $R = 6.0, LCL = 3.0$, find UCL
32. For each of the following cases, find the LCL, CL and UCL for a p chart based on the given information:
- a. $n = 144, p = 0.10$
 - b. $N = 2000, n = 60, p = 0.9$
 - c. $n = 125, p = 0.36$
 - d. $N = 150, n = 30, \bar{p} = 0.25$
 - e. $n = 65, \bar{p} = 0.15$
 - f. $N = 1100, n = 65, \bar{p} = 0.15$
 - g. $N = 2000, n = 22, \bar{p} = 0.16$

33. Compute the producer's risks for the following single sampling plan from lots of 1500 items with AQL = 0.02

- a. $n = 175, c = 1$
- b. $n = 175, c = 2$
- c. $n = 175, c = 3$
- d. $n = 250, c = 1, 2, 3$

Also compute consumer's risk for all of above schemes for LTPD = 0.03

34. Draw OC curve for following sampling plans

- a. $N = 50, n = 10, c = 1$
- b. $N = 150, n = 15, c = 2$
- c. $N = 200, n = 15, c = 1$

Applications

35. A company sells its tires of special brand with a 50,000 kms life warranty. A quality control engineer of the company runs simulated road test to monitor the life of the output from the production process. From each of the last 12 batches of 1000 tires, he has tested 5 tires and recorded the following results, with \bar{x} and R measured in thousand kms.

Batch	1	2	3	4	5	6	7	8	9	10	11	12
\bar{x}	50.5	19.7	50.0	50.7	50.7	50.6	49.8	51.1	50.2	50.4	50.6	50.7
R	1.1	1.6	1.8	0.1	0.9	2.5	0.3	0.8	2.3	1.3	2.0	2.1

- i) Use the given information to help the manager construct an \bar{X} chart
- ii) Comment on the state of production process with explanation

36. The following are the mean and ranges of 20 samples each of size 5. The data pertain to the overall length of a product manufactured during a certain period. Obtain the control limits for \bar{X} chart and R charts and comment on the status of control of the production.

Group No.	Mean	Range	Group No.	Mean	Range
1	0.8372	0.010	11	0.8380	0.006
2	0.8324	0.009	12	0.8322	0.002
3	0.8318	0.008	13	0.8356	0.013
4	0.8344	0.004	13	0.8322	0.005
5	0.8346	0.005	13	0.8404	0.008
6	0.8332	0.011	16	0.8372	0.011
7	0.8340	0.009	17	0.8282	0.006
8	0.8344	0.003	18	0.8346	0.006
9	0.8308	0.002	19	0.8360	0.004
10	0.8350	0.006	20	0.8374	0.006

37. The following data give the weight (in gram) of an automobile part. Five samples of four items each were taken on a random sample basis

(at an interval of one hour each). Draw the control chart for mean and range, and find out if the production process is in control.

Sample No.	1	2	3	4	5
Observations	10	10	10	11	12
	12	12	10	10	12
	10	13	9	9	12
	12	13	11	14	12

38. A machine is set to deliver packets of a give weight. Ten sample of size 5 each were recorded as shown below:

Sample number	1	2	3	4	5	6	7	8	9	10
Sample mean	12.8	13.1	13.5	12.9	13.2	14.1	12.1	15.5	13.9	14.2
Sample range	2.1	3.1	3.9	2.1	1.9	3.0	2.5	2.8	2.5	2.0

Calculate the values for the central line and the control limits for mean chart and range chart and comment on the state of control.

39. A director of 300 emergency medical services in a big city is concerned about response time, the amount of time that elapses between the receipt of a call at the 300 switchboard and the arrival of a municipal rescue squad crew at the calling location. For the last three weeks he has randomly sampled response times for 9 calls each day and obtained the following results of sample mean and range measured in minutes.

Day	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
Mean	11.6	17.4	14.8	13.8	13.9	22.7	16.6
Range	14.1	19.1	22.9	18.2	14.6	23.7	21.0
Day	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
Mean	9.5	12.7	17.7	16.3	10.5	22.5	12.6
Range	12.6	17.0	12.0	15.1	22.1	24.1	21.3
Day	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
Mean	11.4	16.0	11.0	13.3	9.3	21.5	17.9
Range	12.1	21.1	13.5	20.3	16.8	20.7	23.2

- a. Construct an \bar{x} chart to help the director see whether the response time process is in control.
b. What aspect of the chart should disturb him? What action might he take to address the problem?
c. Excluding the data identified as outlying in part (b), is the process under control? Explain.
40. A plant produces paper for newsprint, and rolls of paper are inspected for defects. The results of 25 rolls of paper are given below:

Roll No.	No. of defects								
1	10	6	15	11	16	16	2	21	15
2	20	7	25	12	14	17	3	22	18
3	8	8	7	13	6	18	6	23	20
4	12	9	13	14	5	19	8	24	10
5	13	10	18	15	4	20	9	25	5

Draw control chart for defects and determine whether inspection results indicate stability.

41. Sample of 50 calculators is drawn randomly from the output of a process that produces that product several thousand units daily. Sampled items are inspected for quality, and faulty calculators are rejected. The results to a series of samples are given below:

Sample results of 15 lots of 50 calculators					
Lot No.	No. of defectives	Lot No.	No. of defectives	Lot No.	No. of defectives
1	4	6	7	11	8
2	5	7	3	12	5
3	8	8	2	13	12
4	10	9	5	14	4
5	6	10	6	15	2

Draw an appropriate control chart and interpret it.

42. With a view to examine the quality of an engineering product, 10 samples of 200 items each were taken from a day's production and the number of defective items in each sample was recorded

Sample No. :	1	2	3	4	5	6	7	8	9	10
No. of defectives :	14	20	36	42	22	18	26	2	12	8

Draw control chart s for number of defectives and fraction defectives and comment whether the process is under control or not.

43. Each of the 20 lots of rubber belts contains 2000 belts. Numbers of defective rubber belts in those lots are 410, 420, 324, 292, 310, 282, 300, 320, 296, 392, 432, 294, 324, 220, 400, 400, 258, 226, 460, 280.

Draw a control chart for fraction defectives and comment.

- 44.. Suppose 20 milk packets each of weight 1 litre are selected at random from the packaging process in a dairy firm. The numbers of air bubbles (defects) observed from the bottles are given below

Packet No.	No. of bubbles						
1	4	6	5	11	3	16	5
2	5	7	6	12	5	17	3
3	7	8	2	13	4	18	7
4	3	9	4	14	3	19	6
5	3	10	8	15	4	20	13

Draw a control chart for average number of defects and comment whether the production process is within control or not.

45. The following data are pertaining to the number of imperfection in 20 pieces of cloth of equal length in a certain product of cotton

2, 3, 5, 8, 12, 2, 2, 3, 4, 6, 5, 6, 10, 4, 5, 7, 4, 9, 7, 3

Compute LCL, CL and UCL for a control chart for C, the number of defects, and draw control chart and infer whether the process is in a state of control.

46. An industry that produces steel tubes, the thickness of walls to be controlled. Every hour a sample of 6 tubes is taken and after measurement average thickness in centimeters and the range for each sample is noted, the summary of the measurements are given below:

Sample No.	1	2	3	4	5	6	7	8	9	10
Average thickness	0.25	0.32	0.42	0.22	0.28	0.10	0.25	0.40	0.06	0.29
Range	0.25	0.28	0.46	0.13	0.15	0.10	0.08	0.44	0.10	0.32

Draw mean chart and range chart and give your comments whether any measure is to be taken by the authority.

47. The following information have been gathered from the production process regarding the weights of certain types of chocolates from a factory (every sample contains 5 chocolates, weights measured in gm)

Sample No.	1	2	3	4	5	6	7	8	9	10
Sample mean	12.8	13.1	13.5	12.5	13.2	14.1	12.1	15.5	13.9	14.2
Range	2.1	3.1	2.9	2.1	1.9	3.0	2.5	2.8	2.5	2.0

Draw mean chart and range chart and state your conclusion.

48. A quality control manager of an industry thinks that the production process would be considered under control if the average of the defectives per sample of 10 items is found as 12. What limits would he set in control chart in order to check if the undergoing process is under control or not?
49. The number of defects found in inspection of a particular brand of television at the assembling stage by 20 inspection units, each unit inspects five sets, are given below:

Unit	1	2	3	4	5	6	7	8	9	10
No. of defects	2	40	38	63	92	45	18	65	45	38
Unit	11	12	13	14	15	16	17	18	19	20
No. of defects	40	63	60	80	61	65	56	72	40	50

Set up a control chart to be considered at the time of assembling production of television in future.

50. A company purchases small bolts in cartons that usually contain several thousands of bolts. Each shipment consists of a number of cartons. As part of acceptance procedure for these bolts, 400 bolts are selected at random from each carton and are subjected to visual inspection for certain types of defects. In a shipment of 10 cartons the respective fraction of defectives in the sample from each carton are 0.02, 0.05, 0.07, 0.20, 0.45, 0.32, 0.25, 0.16, 0.23, 0.15. Plot the appropriate control chart and draw your conclusion.

51. The certain electric bulbs company manufactures electric bulbs under an improved process. The Production Engineer takes a random sample of 100 bulbs during production process for inspection. The number of defective pieces is determined by applying a high voltage test. It is supposed the process is supposed to admit as under control if fraction of defectives found as 0.05. Determine the control limits on the p-chart.
52. Number of defects observed in circuit panels used in a computer are as follows:

Panel	1	2	3	4	5	6	7	8	9	10	11	12	13	14
No. of defects	4	0	1	4	5	3	6	2	2	0	5	10	3	4

- Prepare an appropriate control chart to control the quality of the circuit panels. Also assess whether the process is under control on the basis of given information.
53. The following data are related to the length (in cm) of bolts used for special purpose, 10 samples each of size 50 have been measured

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean	11.4	12.0	11.0	11.8	11.2	9.8	10.9	9.7	10.8	9.0
Range	7	1	8	5	7	4	8	4	7	9

- Draw control charts for mean and range and determine whether the process is under control or not.
54. Sample of 50 calculators are drawn randomly from the output of a process that produces several thousands units daily. Sampled items are inspected for quality, and faulty calculators are rejected. The results to a series of samples are given below:

Sample results of 15 lots of 50 calculators

Lot No.	No. inspected	No. of defectives	Lot No.	No. inspected	No. of defectives	Lot No.	No. inspected	No. of defectives
1	50	4	6	50	7	11	50	8
2	50	5	7	50	3	12	50	5
3	50	8	8	50	2	13	50	12
4	50	10	9	50	5	14	50	4
5	50	6	10	50	6	15	50	2

CHAPTER - 18

INTERPOLATION AND EXTRAPOLATION

18.1. Introduction

In earlier chapters, we have already discussed a few of the statistical methods of business, marketing or management data analysis. For proper decision-making, one should have the full set of data at hand. Sometimes the decision maker might face the problem of lack of intermediate data. In that case, the decision may not be efficient. For example, the businessman has to take decision on the basis of monthly data on sales for the past three years, say for January 2009 to December 2011, however, unfortunately sales data for July 2010 or some other period may not be available. In order to come to a proper decision, he has to consider the sales figure of that period. One way of getting rid of this problem is that he can use a logical estimated value of this sale. The method of estimation of this intermediate value for a given period is known as interpolation.

Again, for some particular purpose, the businessman or manager may need the past value of sales which was not recorded, or a future value which is not yet been happened, in this case too, the businessman or manager can use a logical estimation of these values, such type of estimation which is beyond the available range of data, is termed as extrapolation.

Suppose, the sales values y (popularly known as function of argument or entries) as a function of time x (popularly known as argument), are given to us, perhaps it is not necessary to note that here the argument x is independent variable and y is dependent variable. Thus interpolation is defined as the technique of estimating the value of y for any intermediate value of x . Usually the functional relationship between x and y is not known, if it is known, then we can find y for any given value of x by simple substitution. But in case, it is unknown or if it is quite complicated, then the problem of determining the nature of functional relationship arises.

Definition. Interpolation. The technique of estimating the value of a dependent variable or entry for a given intermediate value of an independent variable or argument is called interpolation.

Definition. Extrapolation. The technique of estimating the value of a dependent variable or entry for any value of an independent variable or argument outside the given range is called extrapolation.

The task of interpolation or extrapolation is done using algebraic formula which are based on following two assumptions, such as:

- i) There is no sudden jumps or falls in values of entries for the period under consideration. That means, the given data are free from any abnormal period, such as famines, wars, epidemics, etc. More clearly, all the formulae of interpolation are based on the fundamental assumption that The dependent and independent variables have definite mathematical relationship of the type $y = f(x)$, with fair degree of accuracy.
- ii) In the absence of any evidence to the contrary, the ups or downs in the data have been uniform. That means we assume that the rate of growth of entries over time has been uniform.
- iii) There is regularity in variation of variate values.

18.2. Methods of Interpolation

The following methods are used for interpolation of some entries for given value of argument

- i) Graphic method
- ii) Method of curve fitting
- iii) Methods based on calculus of finite differences

18.2.1. Graphic method. This is the simplest of all methods. For any given values of entries y and argument x , we can easily plot graph. Thus from the graph so obtained, it is possible to determine the value of entry for any given value of argument. For illustration of this graphic method, let us consider the following census data of population:

Year (x)	1951	1961	1971	1981	1991	2001	2011
Population (y) (In million)	55	64	75	95	106	125	159

It is clear from the given information that the population for the year 1995 is not given, but at the moment the decision maker needs this information badly. In this case, he has to interpolate the population for the year 1995. Now the question is how can he interpolate the population for 1995 using graph. For this he has to follow the steps given below.

Steps in Graphic method:

- i) Take a suitable scale for the values of x and y and plot the various points on the graph paper, for given values of x and y .
- ii) Draw a free hand curve passing through the plotted points, and
- iii) Find the point on the curve corresponding to $x = 1995$ by drawing a perpendicular from $x = 1995$ and draw another perpendicular on y -

axis parallel from the point of intersection, find the required value of y where this line cuts the y -axis.

This method is not popularly used for interpolation purposes, because, free hand is purposive, so the value obtained by this method may not be reliable.

Example 18.2.1. From the following data estimate the population for 1985 using graphical method.

Year (x)	1951	1961	1971	1981	1991	2001	2011
Population (y) (In million)	55	64	75	95	106	125	159

Solution. The given data is plotted in graph and shown below.

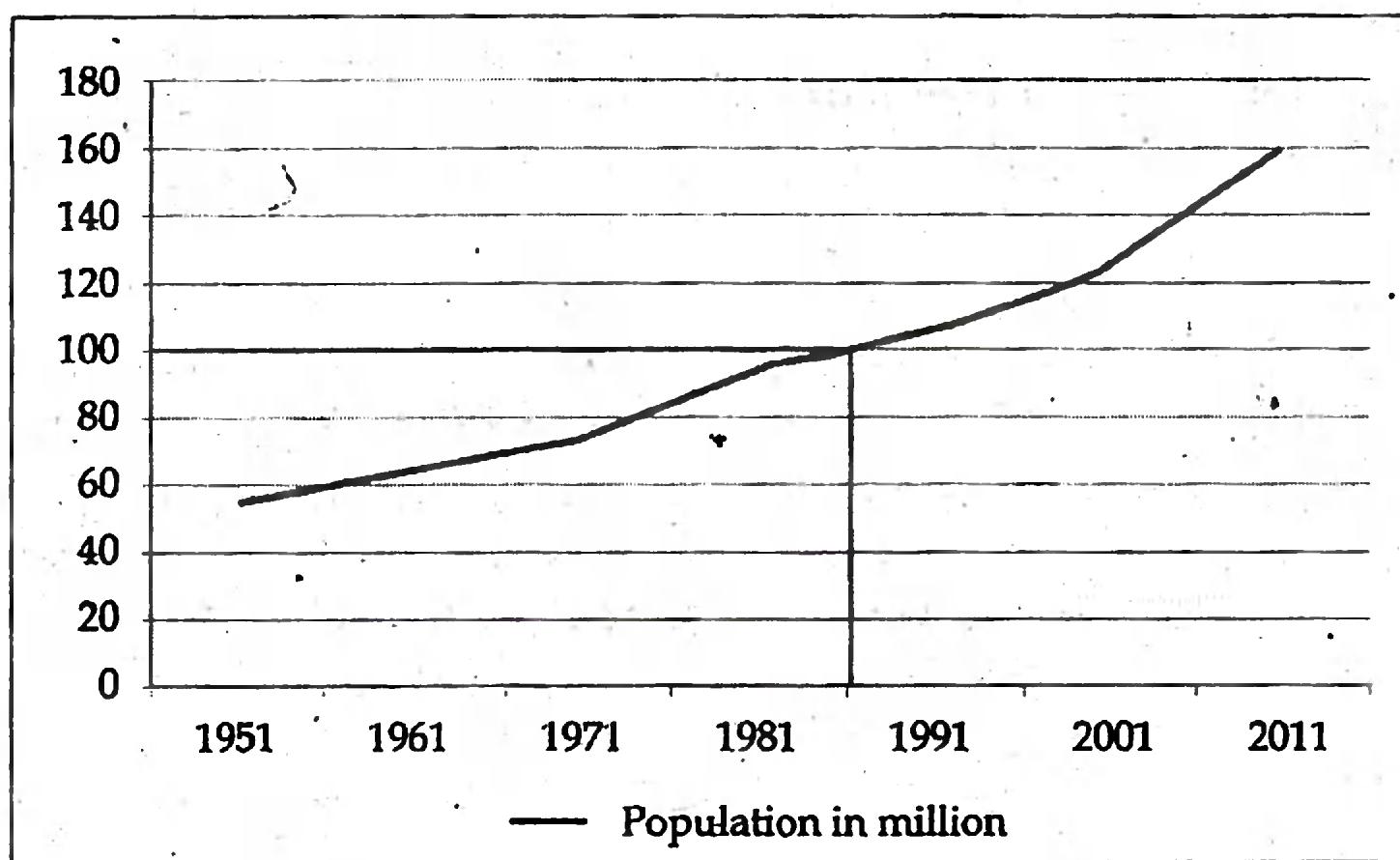


Fig. 18.1. Yearly population in million.

Let us draw a perpendicular on the curve from $x = 1985$. Again, draw another perpendicular on the y -axis parallel to x -axis from the point of intersection. The point where it touches y -axis gives the desired value, which is approximately 100. Thus the estimated population for 1985 is 100 million.

Example 18.2.2. From the following data estimate the profits (Taka in lacs) for the year 1995 by using graphic method.

Year (x)	1992	1993	1994	1995	1996	1997	1998	1999	2000
Profit (Taka in Lac)	30	34	55	?	48	62	65	60	70

Solution. The given data is presented on graph paper as follows.

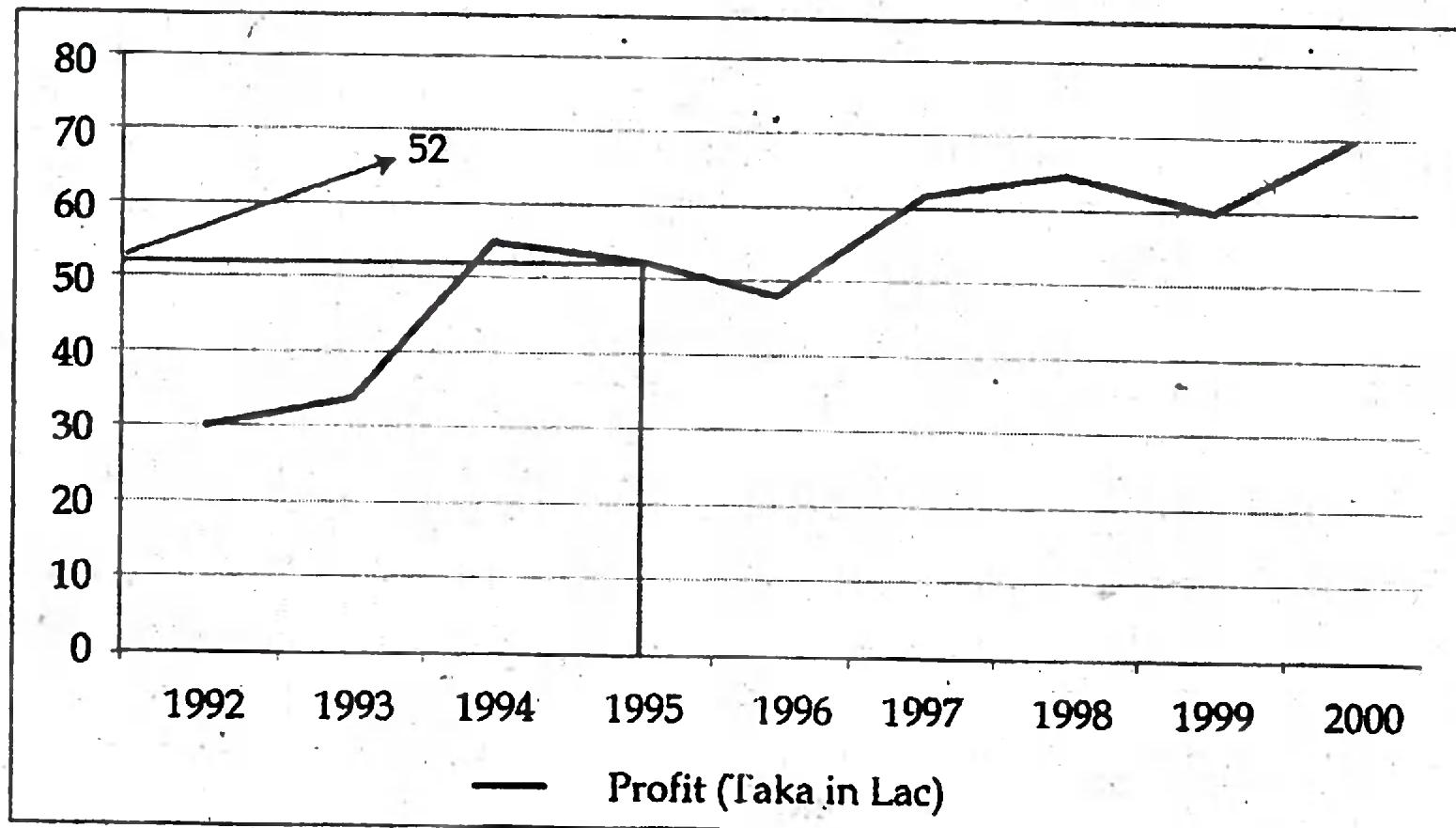


Fig. 18.2. Profits in different years.

Let us draw a perpendicular on the curve from $x = 1995$. Again, draw another perpendicular on the y-axis. The point where it touches the y-axis gives the desired value. This value is 52 in this case. Thus the profit for 1995 is Taka 52 lacs.

18.2.2. Method of curve fitting. This method is used only in those cases where the form of the function is known. The function may be of liner form such as $y = a + bx$ (as we have used in case simple linear regression model where we have estimated the expected value of y for given value of x) or parabolic form such as $y = a + bx + cx^2 + dx^3$, or exponential form such as $y = a + b e^x$ or of any other non-linear form. The curve can be fitted by the method of least squares. Then, the unknown value of y for any given value of x can be calculated from the fitted curve.

However, the following are the merits and demerits of curve fitting method:

Merits

- i) It is applicable in almost all situations.
- ii) The value of entry y can be estimated for any given value of argument x .
- iii) The curve is a good fit if passes through all the points.

Demerits

- i) It needs strong background in mathematics.
- ii) The calculation becomes very lengthy if the observed values of entries and arguments are more.

- iii) The calculation becomes tedious if interpolation or extrapolation has to be carried out for more than one value.

Interpolation or extrapolation using linear curve fitting is the same as done for linear regression model. In spite of above mentioned merits in favour of curve fitting method, considering the demerits of this method, interpolation or extrapolation using non-linear method is not discussed here.

18.3. Methods Based on Calculus of Finite Difference

This method uses some formulae, which are functions of differences of entries. There are two types of differences, viz. forward differences and backward differences. Again, on the basis of presentation of these differences in table, there are three types of tables, viz. diagonal difference table, central difference table and divided difference table. Suppose, $y = f(x)$ is a function of x , x being given at an equal interval. Let the values of x be x_0, x_1, x_2, x_3 , and so on, and the corresponding values of y be y_0, y_1, y_2, y_3 , and so on. The tables where the differences shown are called difference tables. Then the forward and backward differences can be calculated from the forward difference table and backward difference table respectively. The forward difference is denoted by Δ and backward difference is denoted by ∇ . The forward and backward difference tables are illustrated below:

Table 18.1. Forward Difference Table.

Argument x	Entry $y = f(x)$	First differences $\Delta f(x)$ (or Δy)	Second differences $\Delta^2 f(x)$ (or $\Delta^2 y$)
x_0	y_0		
		$\Delta y_0 = y_1 - y_0$	
x_1	y_1		$\Delta^2 y_0 = \Delta y_1 - \Delta y_0$
		$\Delta y_1 = y_2 - y_1$	
x_2	y_2		$\Delta^2 y_1 = \Delta y_2 - \Delta y_1$
		$\Delta y_2 = y_3 - y_2$	
x_3	y_3		$\Delta^2 y_2 = \Delta y_3 - \Delta y_2$
		$\Delta y_3 = y_4 - y_3$	
x_4	y_4		

In the same way, we can calculate third and higher order differences from the table. Note that in case of linear relationship between x and y , the first differences will be constant and second differences will be zero; in case of quadratic relationship between x and y , the second differences will be constant and the third differences would be zero and so on. Also note that

$$\Delta^2 y_1 = \Delta y_2 - \Delta y_1 = (y_3 - y_2) - (y_2 - y_1) = y_3 - 2y_2 + y_1,$$

$$\Delta^2 y_2 = \Delta y_3 - \Delta y_2 = (y_4 - y_3) - (y_3 - y_2) = y_4 - 2y_3 + y_2 \text{ and so on.}$$

$\Delta^3 y_1 = \Delta^2 y_2 - \Delta^2 y_1 = (y_4 - 2 y_3 + y_2) - (y_3 - 2 y_2 + y_1) = y_4 - 3y_3 + 3y_2 - y_1$
and so on.

Table 18.2. Backward Difference Table.

Argument x	Entry $y = f(x)$	First differences $\nabla f(x)$ (or ∇y)	Second differences $\nabla^2 f(x)$ (or $\nabla^2 y$)
x_0	y_0		
		$\nabla y_1 = y_2 - y_1$	
x_1	y_1		$\nabla^2 y_2 = \nabla y_2 - \nabla y_1$
		$\nabla y_2 = y_3 - y_2$	
x_2	y_2		$\nabla^2 y_3 = \nabla y_3 - \nabla y_2$
		$\nabla y_3 = y_4 - y_3$	
x_3	y_3		$\nabla^2 y_4 = \nabla y_4 - \nabla y_3$
		$\nabla y_4 = y_4 - y_2$	
x_4	y_4		

Similarly, we can calculate the backward differences of third and higher order from this table.

However, the following three cases may arise in case of interpolation based on calculus of finite difference:

- i) The case of interpolation with equal intervals
- ii) The case of interpolation with unequal intervals
- iii) The case of central differences.

These methods are also approximate. However, the use of these methods have distinct advantages over the methods of graphs and curve fitting.

Merits

- i) These methods do not assume the form of the function to be known
- ii) These methods are less approximate than the method of graphs
- iii) The calculation remains simple even if some additional observations are included in the given data.

Demerits

There is no definite rule to verify whether the assumptions for the application of the finite difference calculus are valid for the given set of observations.

18.3.1. Interpolation with equal intervals (Newton's Forward Formula). There are two different formulae used for forward and backward interpolation in case of equal intervals of argument. Both formulae are given by Newton and Gregory, and popularly known as Newton's forward and backward formula. Forward formula is based on forward difference table, while backward difference formula is based on backward difference

table. Because of easier approach, here we will discuss only forward formula only.

Suppose, $y = f(x)$ is a function of x , x being given at an equal interval h (that means $x_1 - x_0 = h$, and so on), where the values of x be x_0, x_1, x_2, x_3 , and so on, and the corresponding values of y be y_0, y_1, y_2, y_3 , and so on. Thus, for interpolation of y_x for any given value of x , is given by

$$y_x = y_0 + u \Delta y_0 + \frac{u(u-1)}{2!} \Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!} \Delta^3 y_0 + \dots$$

$$\text{where } u = \frac{x - x_0}{h}$$

Although this formula is mainly designed for interpolating the values of y near the beginning of a set of tabulated values and for extrapolating values of y a short distance backward of y , however, this formula is also useful for interpolation of any value of y for given value of x .

Example 18.3.1. The price per kg of a commodity in a particular period along with demand are given below:

Price/kg (in Taka)	20	25	30	35	40
Demand (in tons)	33.0	29.8	26.6	23.6	20.5

Use Newton's formula to estimate the demand of the commodity at price Taka 32/kg.

Solution. Since there are five observations, we can calculate up to fourth difference. Thus, the Newton's interpolation formula is given by

$$y_x = y_0 + u \Delta y_0 + \frac{u(u-1)}{2!} \Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!} \Delta^3 y_0 + \frac{u(u-1)(u-2)(u-3)}{4!} \Delta^4 y_0$$

$$\text{where } u = \frac{x - x_0}{h}$$

Here, we have to estimate y_{32} for $x = 32$, given $x_0 = 20$, $h = 5$ and $y_0 = 33.0$,

So, $u = \frac{32 - 20}{5} = 2.4$, and the calculation of differences are shown in following difference table:

Price x	Demand $y = f(x)$	Differences			
		First Δy	Second $\Delta^2 y$	Third $\Delta^3 y$	Fourth $\Delta^4 y$
20	33.0	y_0			
		-3.2	Δy_0		
25	29.8	y_1	0	$\Delta^2 y_0$	
		-3.2	Δy_1	+0.1	$\Delta^3 y_0$
30	26.6	y_2	+0.1	$\Delta^2 y_1$	-0.1
		-3.1	Δy_2	0	$\Delta^3 y_1$
35	23.5	y_3	+0.1	$\Delta^2 y_2$	
		-3.0	Δy_3		
40	20.5	y_4			

Thus, the estimated value of y is

$$\begin{aligned}
 y_{32} &= 33 + 2.4(-3.2) + \frac{2.4(2.4-1)}{2} \times 0 + \frac{2.4(2.4-1)(2.4-2)}{6} \times 0.1 \\
 &\quad + \frac{2.4(2.4-1)(2.4-2)(2.4-4)}{24} \times (-0.1) \\
 &= 33 - 7.68 + 0 + 0.0224 + 0.0003 = 25.34 \text{ years (approx).}
 \end{aligned}$$

Example 18.3.2. From the following data of insurance premium (in thousand Taka) to be paid at the particular age at next birthday in years, interpolate the value of premium to be paid by the person if his (her) age at next birth is 17 years using Newton's forward formula.

Age at next birthday	15	25	35	45	55
Premium	11.5	12.6	14.3	16.1	18.3

Solution. Since there are five observations, we can calculate up to fourth difference. Thus, the Newton's interpolation formula is given by

$$\begin{aligned}
 y_x &= y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 \\
 &\quad + \frac{u(u-1)(u-2)(u-3)}{4!}\Delta^4 y_0
 \end{aligned}$$

$$\text{where } u = \frac{x - x_0}{h}$$

Here, we have to estimate y_{17} for $x = 17$, given $x_0 = 15$, $h = 10$ and $y_0 = 11.5$

So, $u = \frac{17-15}{10} = 0.2$, and the calculation of necessary differences are shown in following difference table:

Age x	Premium $y = f(x)$	Differences			
		First Δy	Second $\Delta^2 y$	Third $\Delta^3 y$	Fourth $\Delta^4 y$
15	11.1	y_0			
		1.5	Δy_0		
25	12.6	y_1	0.2	$\Delta^2 y_0$	
		1.7	Δy_1	-0.1	$\Delta^3 y_0$
35	14.3	y_2	0.1	$\Delta^2 y_1$	0.4
		1.8	Δy_2	0.3	$\Delta^3 y_1$
45	16.1	y_3	0.4	$\Delta^2 y_2$	
		2.2	Δy_3		
55	18.3	y_4			

Thus, the estimated value of y is

$$\begin{aligned}
 y_{32} &= 11.1 + 0.2(1.5) + \frac{0.2(0.2-1)}{2} \times 0.2 + \frac{0.2(0.2-1)(0.2-2)}{6} \times (-0.1) \\
 &\quad + \frac{0.2(0.2-1)(0.2-2)(0.2-3)}{24} \times (0.4) \\
 &= 11.1 + 0.3 - 0.16 - 0.048 = 11.40 \text{ (approx).}
 \end{aligned}$$

Thus, value of premium to be paid by the person of age at next birth 17 years is Taka 11.40 thousand.

Example 18.3.4. Suppose the year productions of a firm are recorded as in following table. For some reason, the firm needs the production figure for 1985, estimate the production for 1985 using Newton's formula:

(Note that the production for 1985 for the same data has been estimated by graphical method in Example 18.1)

Year (x)	1951	1961	1971	1981	1991	2001	2011
Production (y) (In tons)	55	64	75	95	106	125	159

Solution. Since there are seven observations, we have to calculate up to sixth difference. Thus, the Newton's interpolation formula is given by

$$\begin{aligned}
 y_x &= y_0 + u\Delta y_0 + \frac{u(u-1)}{2!} \Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!} \Delta^3 y_0 \\
 &\quad + \frac{u(u-1)(u-2)(u-3)}{4!} \Delta^4 y_0 + \frac{u(u-1)(u-2)(u-3)(u-4)}{5!} \Delta^5 y_0 \\
 &\quad + \frac{u(u-1)(u-2)(u-3)(u-4)(u-5)}{6!} \Delta^6 y_0
 \end{aligned}$$

$$\text{where } u = \frac{x - x_0}{h}$$

Here, we have to estimate y_{1985} for $x = 1985$, given $x_0 = 1951$, $h = 10$ and $y_0 = 55$.

So, $u = \frac{1985 - 1951}{10} = 3.4$, and the calculation of necessary differences are

shown in following difference table:

Year x	Produc- tion $y = f(x)$	Differences					
		First Δy	Second $\Delta^2 y$	Third $\Delta^3 y$	Fourth $\Delta^4 y$	Fifth $\Delta^5 y$	Sixth $\Delta^6 y$
1951	55	y_0					
		9	Δy_0				
1961	64	y_1	2	$\Delta^2 y_0$			
		11	Δy_1	7	$\Delta^3 y_0$		
1971	75	y_2	9	$\Delta^2 y_1$	-25	$\Delta^4 y_0$	
		20	Δy_2	-18	$\Delta^3 y_1$	60	$\Delta^5 y_0$
1981	95	y_3	11	Δy_3	-9	$\Delta^2 y_2$	35
		19	Δy_4	8	$\Delta^2 y_3$	$\Delta^4 y_1$	-105
1991	106	y_4			17	$\Delta^3 y_2$	$\Delta^6 y_0$
					-10	$\Delta^4 y_2$	
2001	125	y_5			7	$\Delta^3 y_3$	
		34	Δy_5	15	$\Delta^2 y_4$		
2011	159	y_6					

Thus, the estimated value of y is

$$\begin{aligned}
 y_{1985} &= 55 + 3.4 \times 9 + \frac{3.4(3.4-1)}{2} \times 2 + \frac{3.4(3.4-1)(3.4-2)}{6} \times 7 + \\
 &\quad + \frac{3.4(3.4-1)(3.4-2)(3.4-3)(3.4-4)}{120} \times (60) \\
 &\quad + \frac{3.4(3.4-1)(3.4-2)(3.4-3)(3.4-4)(3.4-5)}{720} \times (-105)
 \end{aligned}$$

$$y_{1985} = 55 + 30.6 + 8.16 + 13.33 - 4.76 - 1.37 - 0.64 = 100.32$$

So, the estimated production for 1985 is 100.32 tons.

18.3.2. Interpolation with unequal intervals (Lagrange's Formula). The Lagrange's formula is merely a relationship between the y 's and x 's, either of which may be taken as the independent variable. Let entry y denotes a polynomial of n th degree which takes the values $y_0, y_1, y_2, \dots, y_n$ when the argument x takes the values $x_0, x_1, x_2, \dots, x_n$, respectively which may or

may not be equidistant. Then, the Lagrange's formula for interpolation of y_x for any given value of x is given by

$$y_x = \frac{(x - x_1)(x - x_2) \dots (x - x_n)}{(x_0 - x_1)(x_0 - x_2) \dots (x_0 - x_n)} y_0 + \frac{(x - x_0)(x - x_2) \dots (x - x_n)}{(x_1 - x_0)(x_1 - x_2) \dots (x_1 - x_n)} y_1 \\ + \frac{(x - x_0)(x - x_1)(x - x_3) \dots (x - x_n)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3) \dots (x_2 - x_n)} y_2 + \dots \\ + \frac{(x - x_0)(x - x_1)(x - x_3) \dots (x - x_{n-1})}{(x_n - x_0)(x_n - x_1)(x_n - x_3) \dots (x_n - x_{n-1})} y_n$$

Lagrange's formula has wider application than other methods due to following advantages:

- i) This method has no restriction on the argument whether it should be equally spaced or not
- ii) This method can be used for any value of argument either for interpolation or extrapolation
- iii) This formula can also be used to estimate the argument for a given value of y . It means, this formula can be used for inverse interpolation.
- iv) The important advantage of this formula is that it does not require any difference table.

18.3.3. Lagrange's Formula for inverse interpolation. The formula for interpolation of any value of x for given value of y can be obtained by interchanging x and y in the formula given above. This type of interpolation is also known as inverse interpolation. Thus, the formula for inverse interpolation of x_y for given value of y is given by

$$x_y = \frac{(y - y_1)(y - y_2) \dots (y - y_n)}{(y_0 - y_1)(y_0 - y_2) \dots (y_0 - y_n)} x_0 + \frac{(y - y_0)(y - y_3) \dots (y - y_n)}{(y_1 - y_0)(y_1 - y_2) \dots (y_1 - y_n)} x_1 \\ + \frac{(y - y_0)(y - y_1)(y - y_3) \dots (y - y_n)}{(y_2 - y_0)(y_2 - y_1)(y_2 - y_3) \dots (y_2 - y_n)} x_2 + \dots \\ + \frac{(y - y_0)(y - y_1)(y - y_3) \dots (y - y_{n-1})}{(y_n - y_0)(y_n - y_1)(y_n - y_3) \dots (y_n - y_{n-1})} x_n$$

Example 18.3.5. The investment and profit, measured in million Taka, of a firm for different periods are given below. Now the firm wants to estimate the expected profit of the firm, if Taka 33 million is invested.

Find the expected profit of the firm by using suitable formula.

Investment (in million Taka)	20	30	38
Profit (in million Taka)	1.3010	1.4771	1.5798

Solution. Here we have to estimate dependent variable profit (y) for the given value of independent variable investment (x), which is not equidistant, so for interpolation of profit we have to use Lagrange's interpolation formula that is given by

$$y = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2$$

Here, we have to find y for $x = 33$, given y -values are $y_0 = 1.3010$, $y_1 = 1.4771$, $y_2 = 1.5798$ and corresponding x 's are $x_0 = 20$, $x_1 = 30$, $x_2 = 38$ respectively.

Substituting the values in the formula, we have,

$$\begin{aligned} y_{33} &= \frac{(33 - 30)(33 - 38)}{(20 - 30)(20 - 38)} \times 1.3010 + \frac{(33 - 20)(33 - 38)}{(30 - 20)(30 - 38)} \times 1.4771 \\ &\quad + \frac{(33 - 20)(33 - 30)}{(38 - 20)(38 - 30)} \times 1.5798 \\ &= -0.1084 + 1.2001 + 0.4278 = 1.5196 \end{aligned}$$

Thus, the estimated profit is Taka 1.5196 million.

Example 18.3.6: The following table gives the monthly average sales (in thousand Taka) of 4 salesmen according to their age (in approximate years). Estimate the expected average sales of the salesman who is 35 years old.

Age	25	30	40	50
Sales	48.0	54.0	80.6	92.2

Solution. Since the intervals of argument age (x) are unequal, so Lagrange's formula would be appropriate for interpolation of entry (y). The formula is given by

$$\begin{aligned} y &= \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} y_0 + \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} y_1 \\ &\quad + \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_0 - x_3)} y_2 + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} y_3 \end{aligned}$$

We have to find y_x for $x = 35$,

Substituting the values of x 's and y 's in the formula, we have

$$\begin{aligned} y_{35} &= \frac{(35 - 30)(35 - 40)(35 - 50)}{(25 - 30)(25 - 40)(25 - 50)} \times 48 + \frac{(35 - 25)(35 - 40)(35 - 50)}{(30 - 25)(30 - 40)(30 - 50)} \times 54 \\ &\quad + \frac{(35 - 30)(35 - 30)(35 - 50)}{(40 - 25)(40 - 30)(25 - 50)} \times 80.6 + \frac{(35 - 25)(35 - 40)(35 - 40)}{(50 - 25)(50 - 30)(50 - 40)} \times 92.2 \end{aligned}$$

$$= -9.6 + 40.5 + 40.3 - 4.61 = 66.59 = 67 \text{ (approx.)}$$

Thus the average monthly sales of the salesman who is 35 years old is estimated as Taka 67 thousand (approximately).

Example 18.3.7. Using the data given in example 18.6, estimate the amount to be invested in order to make a profit of Taka 1.50 million.

Solution. We know, the Lagrange's formula for inverse interpolation is given by

$$\begin{aligned} x_y = & \frac{(y - y_1)(y - y_2) \dots (y - y_n)}{(y_0 - y_1)(y_0 - y_2) \dots (y_0 - y_n)} x_0 + \frac{(y - y_0)(y - y_3) \dots (y - y_n)}{(y_1 - y_0)(y_1 - y_2) \dots (y_1 - y_n)} x_1 \\ & + \frac{(y - y_0)(y - y_1)(y - y_3) \dots (y - y_n)}{(y_2 - y_0)(y_2 - y_1)(y_2 - y_3) \dots (y_2 - y_n)} x_2 + \dots \\ & + \frac{(y - y_0)(y - y_1)(y - y_3) \dots (y - y_{n-1})}{(y_n - y_0)(y_n - y_1)(y_n - y_3) \dots (y_n - y_{n-1})} x_n \end{aligned}$$

Where x stands for investment and y stands for profit.

Substituting the values of x and y , we have

$$\begin{aligned} x_{1.50} = & \frac{(1.50 - 1.4771)(1.50 - 1.5798)}{(1.3010 - 1.4771)(1.3010 - 1.5798)} \times 20 \\ & + \frac{(1.50 - 1.3010)(1.50 - 1.5798)}{(1.4771 - 1.3010)(1.4771 - 1.5798)} \times 30 \\ & + \frac{(1.50 - 1.3010)(1.50 - 1.4771)}{(1.5798 - 1.3010)(1.5798 - 1.4771)} \times 38 \\ = & -0.744 + 26.34 + 6.048 = 31.645. \end{aligned}$$

That means in order to make a profit of 1.50 million, the firm has to invest Taka 31.645 million.

Example 18.3.8. Suppose data given in example 18.7 are related to age and sales of four salesmen working in a big company. The company wants to employ a salesman who can sale Taka 70 thousand per month. Determine the expected age of the salesman to be employed.

Solution:

Solution. We know, the Lagrange's formula for inverse interpolation is given by

$$x_y = \frac{(y - y_1)(y - y_2) \dots (y - y_n)}{(y_0 - y_1)(y_0 - y_2) \dots (y_0 - y_n)} x_0$$

$$\begin{aligned}
 & + \frac{(y - y_0)(y - y_3) \dots (y - y_n)}{(y_1 - y_0)(y_1 - y_2) \dots (y_1 - y_n)} x_1 \\
 & + \frac{(y - y_0)(y - y_1)(y - y_3) \dots (y - y_n)}{(y_2 - y_0)(y_2 - y_1)(y_2 - y_3) \dots (y_2 - y_n)} x_2 + \dots \\
 & + \frac{(y - y_0)(y - y_1)(y - y_3) \dots (y - y_{n-1})}{(y_n - y_0)(y_n - y_1)(y_n - y_3) \dots (y_n - y_{n-1})} x_n
 \end{aligned}$$

Where x stands for investment and y stands for profit, and we have to find x_y for $y = 70$.

Substituting the values of x and y , we have

$$\begin{aligned}
 x_{70} &= \frac{(70.0 - 54.0)(70.0 - 80.6)(70.0 - 92.2)}{(48.0 - 54.0)(48.0 - 80.6)(48.0 - 92.2)} \times 25 \\
 &\quad + \frac{(70.0 - 48.0)(70.0 - 80.6)(70.0 - 92.2)}{(54.0 - 48.0)(54.0 - 80.6)(54.0 - 92.2)} \times 30 \\
 &\quad + \frac{(70.0 - 48.0)(70.0 - 54.0)(70.0 - 92.2)}{(80.6 - 48.0)(80.6 - 54.0)(80.6 - 92.2)} \times 40 \\
 &\quad + \frac{(70.0 - 48.0)(70.0 - 54.0)(70.0 - 80.6)}{(92.2 - 48.0)(92.2 - 54.0)(92.2 - 80.6)} \times 50 \\
 &= -10.887 + 25.475 + 31.074 - 9.525 = 36.136 = 36 \text{ (approx)}
 \end{aligned}$$

So, in order to achieve the sales figure of the company, it has to appoint a salesman of age about 36 years.

Note. It is mentioned above, that one of the advantage of Lagrange's formula that it can be used successfully used for extrapolation. Let us illustrate the matter with following illustrations.

Example 18.3.9. (Extrapolation). Consider the data given in example 18.6 and suppose the firm wants to estimate its profit to be incurred if it invests Taka 45 million. Help the firm by estimating its expected profit from the investment of Taka 45 million using suitable formula.

Solution. The expected profit of the firm can be suitably estimated by using Lagrange's formula given by

$$y_x = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2$$

Here, we have to find y for $x = 45$, given y -values are $y_0 = 1.3010$, $y_1 = 1.4771$, $y_2 = 1.5798$ and corresponding x 's are $x_0 = 20$, $x_1 = 30$, $x_2 = 38$ respectively.

Substituting the values in the formula, we have,

$$\begin{aligned}
 y_{45} &= \frac{(45 - 30)(45 - 38)}{(20 - 30)(20 - 38)} \times 1.3010 + \frac{(45 - 20)(45 - 38)}{(30 - 20)(30 - 38)} \times 1.4771 \\
 &\quad + \frac{(45 - 20)(45 - 30)}{(38 - 20)(38 - 30)} \times 1.5798 \\
 &= 0.7589 - 3.2310 + 4.1141 = 1.6420
 \end{aligned}$$

Thus, the estimated profit is Taka 1.6420 million.

Example 18.3.10. (Extrapolation). Consider the data given in example 18.7, suppose the authority is in a position to appoint a salesman of age 22. Estimate the average monthly sales of this new salesman using suitable formula.

Solution. Here, Lagrange's formula would be appropriate for extrapolation of entry (y). The formula is given by

$$\begin{aligned}
 y &= \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} y_0 + \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} y_1 \\
 &\quad + \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_0 - x_3)} y_2 + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} y_3
 \end{aligned}$$

We have to find y_x for $x = 22$,

Substituting the values of x 's and y 's in the formula, we have

$$\begin{aligned}
 y_{35} &= \frac{(22 - 30)(22 - 40)(22 - 50)}{(25 - 30)(25 - 40)(25 - 50)} \times 48 + \frac{(22 - 25)(22 - 40)(22 - 50)}{(30 - 25)(30 - 40)(30 - 50)} \times 54 \\
 &\quad + \frac{(22 - 30)(22 - 30)(22 - 50)}{(40 - 25)(40 - 30)(25 - 50)} \times 80.6 + \frac{(22 - 25)(22 - 40)(22 - 40)}{(50 - 25)(50 - 30)(50 - 40)} \times 92.2 \\
 &= 103.219 - 84.648 + 36.109 - 7.966 = 46.714
 \end{aligned}$$

Thus the average monthly sales of the salesman who is 22 years old is estimated as Taka 46.714 thousands.

Questions

1. Distinguish between interpolation and extrapolation. Mention the names of some methods of interpolation known to you and describe any one of them.
2. State the advantages of calculus method over graphical method. Also state merits and demerits of parabolic method.
3. State the Newton's formula for forward interpolation with equal intervals along with underlying assumptions.
4. State Lagrange's formula for interpolation. Also state the advantages of this method over Newton's method.
5. State Lagrange's formula for inverse interpolation.
6. What are the assumptions on which the interpolation and extrapolation are based?
7. How can you differentiate interpolation from extrapolation. State Lagrange's formula for extrapolation.

Applications

8. Estimate the production (in tons) for the year 1999 from following data using graphical method

Year	1997	1998	2000	2001	2002
Production	320	300	280	278	250

(Ans. 284 tons)

9. The following table gives the total expenditure (in thousand taka) of a firm during the period 1996-2003 are as follows:

Year	1996	1997	1998	1999	2001	2002	2003
Production	265	320	300	280	278	250	335

Estimate the expenditure of the firm using graphical method.

10. From the following table of yearly profits (in lac Taka) of a company, find out the profits for the year 2002 using graphical method

Year	1999	2000	2001	2003	2004
Profits	173	149	145	131	141

(Ans. Taka 139.2 lac)

11. From the following of yearly premium for policies maturing at different age, estimate the premium for policies maturing at the age of 47 years using graphical method and Newton's interpolation formula:

Age (in years)		45	50	55	60	65
Premium (in '00 Tk)	287	240	208	186	171	

(Ans. Tk. 266 hundred)

12. Find the annual premium at the age of 26 from following information:

Age (in years)	20	25	30	35	40
Premium (in '000 Tk)	23	26	30	35	42

(use graphical method and Newton's forward formula)

(Ans. Tk. 26.73 thousand)

13. The following table gives the demand (in kg) of some commodity at different times for different price (in Tk/kg), estimate demand of the commodity for price Tk 12/kg using Newton's forward formula:

Price	10	15	20	25	30	35
Demand	35.4	32.2	29.1	26.0	23.1	20.4

(Ans. 34.01 kg)

14. The experience (in round years) and monthly sales (in thousand taka) of 4 salesmen are given below:

Experience	5	6	9	11
Sales	12	13	14	16

- i) Calculate the sales of the salesman who has 10 years of experience
 - ii) Calculate the expected sales of the salesman who has 14 years of experience.
 - iii) Also determine the experience required to make a sales of Taka 15.5 thousand.
15. The sales and profit, both are measured in thousand Taka, of 5 items are given below:

Sales	23	37	38	42	51
Profit	3	7	9	11	14

Calculate the profit of the item if its sales value is Tk. 40 by using Lagrange's formula.

16. Suppose typing speeds (w/m) of five Assistants of a firm according to their ages (in round years) are as follows:

Age	24	32	35	41
Speed	37	40	42	50

- i) Estimate the speed of the Assistant who is 28 years old
 - ii) Determine the expected speed of the Assistant whose age is 45 years
 - iii) If the firm needs an Assistant who has typing speed 45 words per minute, what should be age of Assistant?
17. The per unit cost and sales value of an item produced by a company for different periods are as follows:

Cost (in Tk.)	21	30	36	40	48	60
Sales (in Tk.)	25	35	40	46	56	72

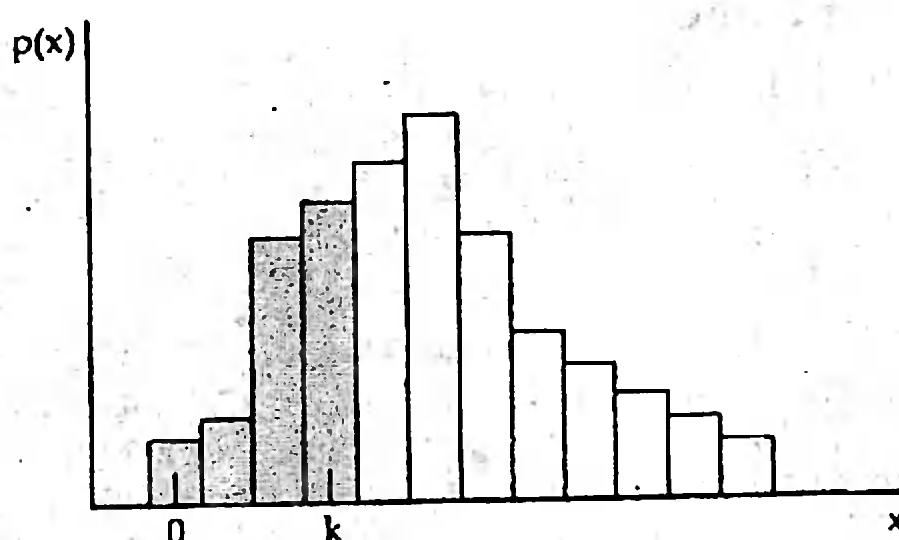
- i) Calculate the sales value of the item whose cost is Tk. 43.
- ii) Estimate the cost of the item if its sales value to be determined as Tk. 50.
- iii) Also determine the expected sales value of the item if their costs rise to Tk 70.

Appendix

Table 1. Random numbers.

27 76 74 35 84	85 30 18 80 77	20 49 06 97 14	73 03 54 12 07	74 69 90 93 10
13 02 51 43 38	54 06 61 52 43	47 72 46 67 33	47 43 14 39 05	31 04 85 66 99
80 21 73 62 92	98 52 52 43 35	24 43 22 48 96	43 27 75 88 74	11 46 64 60 82
10 87 56 20 04	90 39 16 11 05	57 41 10 63 68	53 85 63 07 43	08 07 08 47 41
54 12 75 73 26	26 62 91 87 24	47 28 87 79 30	54 02 78 78 86	61 73 27 54 54
60 31 14 28 24	37 30 14 26 78	45 99 04 32 42	17 37 45 20 03	70 70 77 02 14
49 31 14 28 24	37 30 14 26 78	45 99 04 32 42	17 37 45 20 03	70 70 77 02 14
78 62 65 15 94	16 45 39 46 14	39 01 49 70 66	83 01 20 98 32	25 57 17 76 28
66 69 21 39 86	99 83 70 05 82	81 23 24 49 87	09 50 49 64 12	90 19 37 95 68
44 07 12 80 91	07 36 29 77 03	76 44 74 25 37	98 52 49 78 31	65 70 40 95 14
41 46 88 51 49	49 55 41 79 94	14 92 43 96 50	95 29 40 05 56	70 48 10 69 05
94 55 93 75 59	49 67 85 31 19	70 31 20 56 82	66 98 63 40 99	74 47 42 07 40
41 61 57 03 60	64 11 45 80 60	90 85 06 46 18	80 62 05 17 90	11 43 63 80 72
50 27 39 31 13	41 79 48 68 01	24 78 18 96 83	55 41 18 56 67	77 53 59 98 92
41 39 68 05 04	90 67 00 82 89	40 90 20 50 69	95 08 30 67 83	28 10 25 78 16
25 80 72 42 60	71 52 97 89 20	72 68 20 73 85	90 72 65 71 66	98 88 40 85 83
06 17 09 79 65	88 30 39 80 41	21 44 34 18 08	68 98 48 36 20	89 74 79 88 82
60 80 85 44 44	74 41 28 11 05	01 17 62 88 38	36 42 11 64 89	18 05 95 10 61
80 94 04 48 93	10 40 83 62 22	80 58 27 19 44	92 63 84 03 33	67 05 41 60 67
19 51 69 01 20	46 75 97 16 43	13 17 75 52 92	21 03 68 28 08	77 50 19 74 27
49 38 65 44 80	23 60 42 35 54	21 78 54 11 01	91 17 81 01 74	29 42 08 04 38
06 31 28 89 40	15 99 56 93 21	47 45 86 48 09	98 18 98 18 51	29 65 18 42 15
60 94 20 03 07	11 89 79 26 74	40 40 56 80 32	96 71 75 42 44	10 70 14 13 93
92 32 99 89 32	78 28 44 63 47	71 20 99 20 61	39 44 89 31 36	25 72 20 86 64
77 93 66 35 74	31 38 45 19 24	85 56 12 96 71	58 13 71 78 20	22 75 13 65 18
38 10 17 77 56	11 65 71 38 97	95 88 95 70 67	47 64 81 38 85	70 66 99 34 06
39 64 16 94 57	91 33 92 25 02	92 61 38 97 19	11 94 75 62 03	19 32 42 05 04
84 05 44 04 55	99 39 66 36 80	67 66 76 06 31	69 18 19 68 45	38 52 51 16 00
47 46 80 35 77	57 64 96 32 66	24 70 07 15 94	14 00 42 31 53	69 24 90 57 47
43 32 13 13 70	28 97 72 38 96	76 47 96 85 52	62 34 20 75 89	08 80 90 59 85
64 28 16 18 26	18 55 56 49 37	13 17 33 33 65	78 85 11 64 99	87 06 41 30 75
66 84 77 04 95	32 35 00 29 85	86 71 63 87 46	26 31 37 74 63	55 38 77 26 81
72 46 13 32 30	21 52 95 34 24	92 58 10 22 62	78 43 86 62 76	18 39 67 35 38
21 03 20 10 50	13 05 81 62 18	12 47 05 65 00	15 29 27 61 39	59 52 65 21 13
95 36 26 70 11	06 65 11 61 36	01 01 60 08 57	55 01 85 63 74	35 82 47 17 08
40 71 29 73 80	10 40 45 54 52	34 03 06 07 26	75 21 11 02 71	36 63 36 84 24
58 27 56 17 64	97 58 65 47 16	50 25 94 63 45	87 19 54 60 92	26 78 76 09 39
89 51 41 17 88	68 22 42 34 17	73 95 97 61 45	30 34 24 02 77	11 04 97 20 49
15 47 25 06 69	48 13 93 67 32	46 87 43 70 88	73 46 50 98 19	58 86 93 52 20
12 12 08 61 24	51 24 74 43 02	60 88 35 21 09	21 43 73 67 86	49 22 67 78 37
19 61 27 84 30	11 66 19 47 70	77 60 36 56 69	86 86 81 26 65	30 01 27 59 89
39 14 17 74 00	28 00 06 42 38	73 25 87 17 94	31 34 02 62 56	66 45 33 70 16
64 75 68 04 57	08 74 71 28 36	03 46 95 06 78	03 27 44 34 23	66 67 78 25 56
92 90 15 18 78	56 44 12 29 98	29 71 83 84 47	06 45 32 53 11	07 56 55 37 71
03 55 19 00 70	00 48 39 40 50	45 93 81 81 35	36 90 84 33 21	11 07 35 18 03

Table 2. Cumulative Binomial Probabilities.



Tabulated values are $P(x \leq k) = p(0) + p(1) + \dots + p(k)$ (Computations are rounded at the third decimal place.)

Table 3. Cumulative Poisson Probabilities.

Tabulated values are $p(x \leq k) = p(0) + p(1) + \dots + p(k)$
 (Computations are rounded at the third decimal place.)

Table 4. Areas under the standard Normal Distribution.

$$\Phi = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = P(Z \leq z)$$

<i>z</i>	0	1	2	3	4	5	6	7	8	9
-3	.0013	.0010	.0007	.0005	.0003	.0002	.0002	.0001	.0001	.0000
-2.9	.0019	.0018	.0017	.0017	.0016	.0015	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0020	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.016	.0132	.0129	.0216	.0022	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0238	.0233
-1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0300	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0570	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3516	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Reprinted with permission of the Macmillan Company from INTRODUCTION TO PROBABILITY AND STATISTICS. Second edition, by B.W. Lindgren and G.W. McElarth. Copyright © 1966 by B.W. Lindgren and G.W. McElarth.

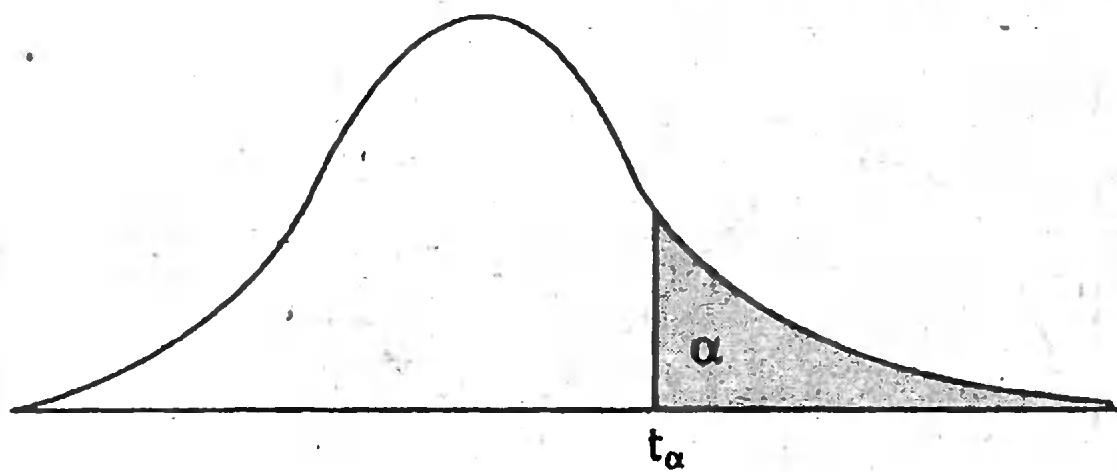
z	0	1	2	3	4	5	6	7	8	9
.0	.5000	.5040	.5080	.5120	.5160	.5190	.5239	.5279	.5319	.5359
+.1	.5398	.5438	.5478	.5517	.5557	.5596	.5363	.5675	.5714	.5753
+.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
+.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
+.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
+.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
+.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
+.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7974	.7823	.7549
+.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
+.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
+1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
+1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
+1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
+1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
+1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
+1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9430	.9441
+1.6	.9452	.9563	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
+1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
+1.8	.9641	.9648	.9656	.9664	.9671	.9678	.9686	.9693	.9700	.9706
+1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9762	.9767
+2.0	.9772	.9778	.9738	.9788	.9793	.9842	.9846	.9850	.9854	.9857
+2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
+2.2	.9861	.9846	.9868	.9871	.9874	.9878	.9881	.9884	.9887	.9890
+2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
+2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
+2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9963	.9964
+2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
+2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
+2.8	.9974	.9975	.9967	.9977	.9977	.9978	.9979	.9979	.9980	.9981
+2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
+3	.9987	.9990	.9993	.9995	.9997	.9998	.9998	.9999	.9999	1.0000

Note 1 : If a normal variable X is not "standard". Its values must be

"standardized" : $Z = \frac{X - \mu}{\sigma}$. That is, $P[X \leq x] = \Phi\left(\frac{x - \mu}{\sigma}\right)$.

Note 2 : For $z \geq 4$, $\Phi(z) = 1$ to four decimal places; for $z \leq -4$, $\Phi(z) = 0$ to four decimal places.

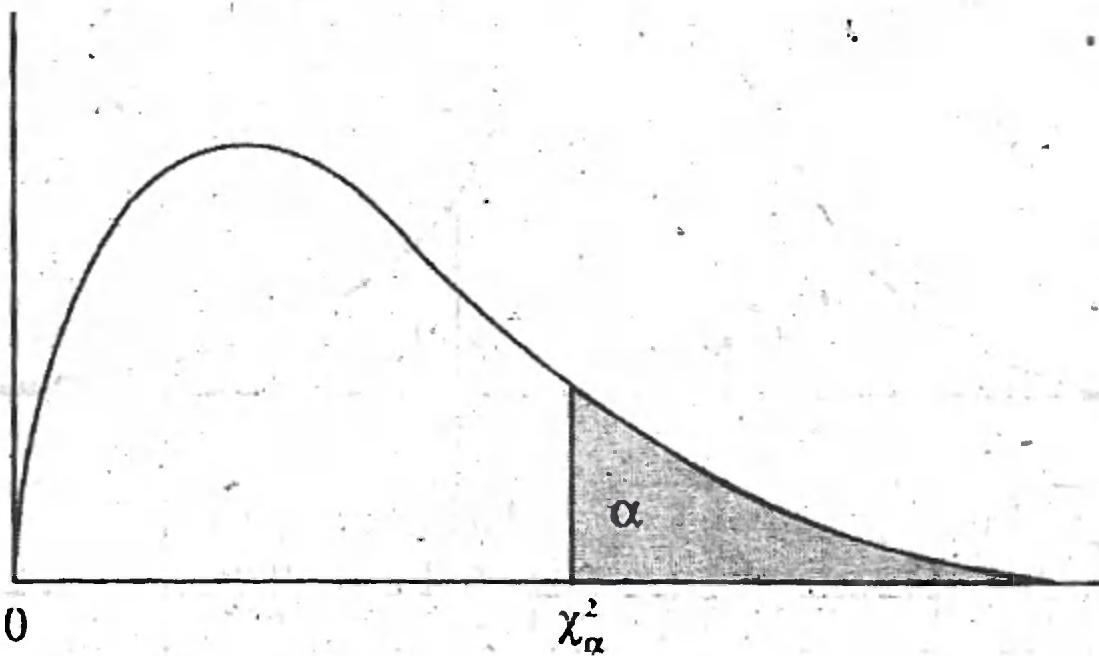
Note 3 : The entries opposite to $z = 3$ are for 3.0, 3.1, 3.2, etc.

Table 5. Critical Values of t.

dt	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.196	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
∞	1.282	1.645	1.960	2.326	2.576	∞

Source : From "Table of percentage Points of the t-Distribution," *Biometrika* 32 (1941): 300 Reproduced by permission of the *Biometrika* Trustees.

Table 6. Critical Values of Chi-Square.

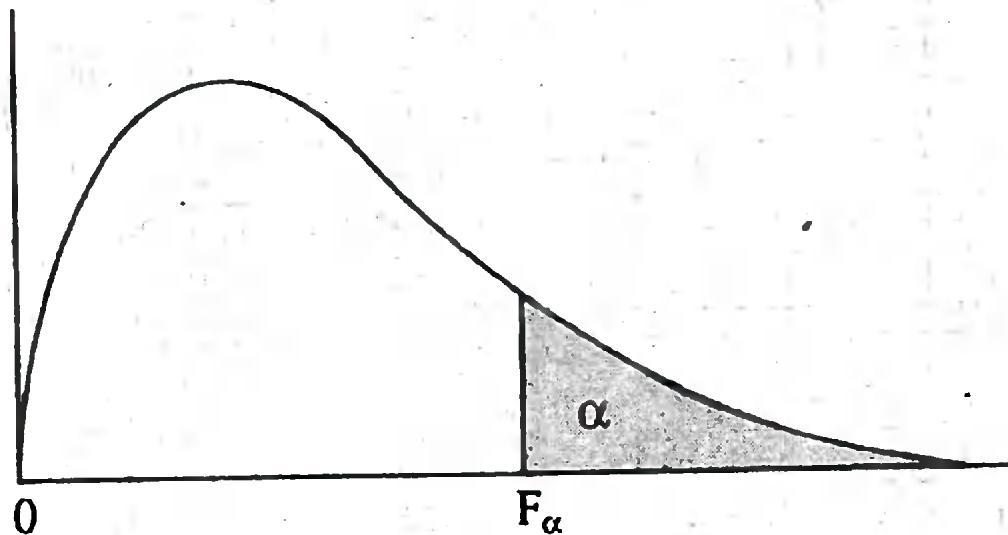


df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$
1	.0000393	.0001571	.0009821	.0039321	.0157908
2	.0100251	.0201007	.0506356	.1025870	.2107200
3	.0717212	.1148320	.2157950	.3518460	.5843750
4	.2069900	.2971100	.4844190	.7107210	1.0636230
5	.4117400	.5543000	.8312110	1.1454760	1.6103100
6	.6757270	.8720850	1.2373470	.6353900	2.2041300
7	.9892650	1.2390430	1.6898700	2.1673500	2.8331100
8	1.3444190	1.6464820	2.1797300	2.7326400	3.4895400
9	1.7349260	2.0879120	2.7003900	3.3251100	4.1681600
10	2.1558500	2.5582100	3.2469700	3.9403000	4.8651800
11	2.6032100	3.0534700	3.8157500	4.5748100	5.5777900
12	3.0738200	3.5705600	4.4037900	5.2260300	6.3038000
13	3.5650300	4.1069100	5.0087400	5.8918600	7.0415000
14	4.0746800	4.6604300	5.6287200	6.5706300	7.7895300
15	4.6009400	5.2293500	6.2621400	7.2609400	8.5467500
16	5.1422400	5.8122100	6.9076600	7.9616400	9.3122300
17	5.6972400	6.4077600	7.5641800	8.6717600	10.0852000
18	6.2648100	7.0149100	8.2307500	9.3904600	10.8649000
19	6.8439800	7.6327300	8.9065500	10.1170000	11.6509000
20	7.4338600	8.2604000	9.5908300	10.8508000	12.4426000
21	8.0336600	8.8972000	10.2829300	11.5913000	13.2396000
22	8.6427200	9.5424900	10.9823000	12.3380000	14.0415000
23	9.2604200	10.1956700	11.6885000	13.0905000	14.8479000
24	9.8862300	10.8564000	12.4011000	13.8484000	15.6587000
25	10.5197000	11.5240000	13.1197000	14.6114000	16.4734000
26	11.1603000	12.1981000	13.8439000	15.3791000	17.2919000
27	11.8076000	12.8786000	14.5733000	16.1513000	18.1138000
28	12.4613000	13.5648000	15.3079000	16.9279000	18.9392000
29	13.1211000	14.2565000	16.0471000	17.7083000	19.7677000
30	13.7867000	14.9535000	16.7908000	18.4926000	20.5992000

40	20.7065000	22.1643000	24.4331000	26.5093000	29.0505000
50	27.9907000	29.7067000	32.3574000	34.7642000	37.6886000
60	35.5346000	37.4848000	40.4817000	43.1879000	46.4589000
70	43.2752000	45.4418000	48.7576000	51.7393000	55.3290000
80	51.1720000	53.5400000	57.1532000	60.3915000	64.2778000
90	59.1963000	61.7541000	65.6466000	69.1260000	73.2912000
100	67.3276000	70.0648000	74.2219000	77.9295000	82.3581000

df	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8485	31.9999	34.2672
17	24.7690	27.8571	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169

Table 7. Percentage Points of the F Distribution.



df_2	A.	df ₁								
		1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3
	.010	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
	.005	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.99	99.36	99.37	99.39
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.64	27.49	27.35
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.63	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.005	3.59	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
7	.100	8.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	8.07	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68

	.025	12.25	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	16.24	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.005	3.46	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54

df_1											
10	12	5	20	24	30	40	60	120	∞	a	df_2
60.19	60.71	60.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33	.100	1
241.9	243.9	245.9	248.0	249.1	250.1	251.2	252.2	253.3	254.3	.050	
968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018	.025	
6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	.010	
24224	24426	24630	24836	24940	25044	25148	25253	25359	25465	.005	
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	.100	2
19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50	.050	
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50	.025	
99.40	99.42	99.43	99.46	99.46	99.47	99.47	99.48	99.49	99.50	.010	
199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.6	199.5	199.5	.005	
5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13	.100	3
8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53	.050	
14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90	.025	
27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13	.010	
43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83	.005	
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76	.100	4
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	.050	
8.82	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26	.025	
14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46	.010	
20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32	.005	
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10	.100	5
4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36	.050	
6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02	.025	
10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	.010	
13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14	.005	
2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72	.100	6
4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67	.050	
5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85	.025	
7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88	.010	
10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88	.005	

2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47	.100	7
3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	.050	
4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14	.025	
6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	.010	
8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08	.005	
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29	.100	8
3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93	.050	
4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67	.025	
5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	.010	
7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95	.005	
2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16	.100	9
3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71	.050	
3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33	.025	
5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	.010	
6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19	.005	

		df_1								
df_2	A	1	2	3	4	5	6	7	8	9
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.28	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54

16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	.010	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96

df ₁											
10	12	15	20	24	30	40	60	120	∞	a	df ₂
2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06	.100	10
2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54	.050	
3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08	.025	
4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	.010	
5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64	.005	
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97	.100	11
2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40	.050	
3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88	.025	
4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	.010	
5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23	.005	
2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90	.100	12
2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30	.050	
3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72	.025	
4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	.010	
5.09	4.19	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90	.005	
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85	.100	13
2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21	.050	
3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60	.025	
4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17	.010	
4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65	.005	
2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80	.100	14

2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13	.050	
3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49	.025	
3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	2.09	3.00	.010	
4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44	.005	
2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76	.100	15
2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07	.050	
3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40	.025	
3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	.010	
4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26	.005	
2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72	.100	16
2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01	.050	
2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32	.025	
3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	.010	
4.27	4.10	2.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11	.005	
2.00	1.96	1.91	1.86	1.87	1.81	1.78	1.75	1.72	1.69	.100	17
2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96	.050	
2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25	.025	
3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	.010	
4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98	.005	
1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66	.100	18
2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92	.050	
2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19	.025	
3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	.010	
4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87	.005	
1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63	.100	19
2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88	.050	
2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13	.025	
3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	.010	
3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78	.005	
1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61	.100	20
2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84	.050	
2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09	.025	
3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	.010	
3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69	.005	

df_1											
df_2	a	1	2	3	4	5	6	7	8	9	
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	
	.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	

	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	2.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	3.72	4.22	3.90	3.67	3.50	3.36	3.26
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45

df_1												
10	12	15	20	24	30	40	60	120	∞	a	df_2	
1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59	.100	21	
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81	.050		
2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04	.025		
3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	.010		
3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61	.005		
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57	.100	22	
2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78	.050		
2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00	.025		
3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	.010		
3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55	.005		
1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55	.100	23	
2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76	.050		
2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97	.025		
3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	.010		
3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48	.005		
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53	.100	24	
2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73	.050		
2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94	.025		
3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	.010		
3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43	.005		
1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52	.100	25	
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71	.050		
2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91	.025		
3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17	.010		
3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38	.005		
1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50	.100	26	
2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69	.050		
2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88	.025		
3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13	.010		
3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33	.005		
1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49	.100	27	
2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67	.050		
2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85	.025		
3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10	.010		
3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29	.005		
1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48	.100	28	
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	.050		
2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83	.025		
3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06	.010		
3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25	.005		
1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47	.100	29	
2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64	.050		
2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81	.025		
3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03	.010		
3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21	.005		
1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46	.100	30	

2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62	.050	
2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	-1.79	.025	
2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01	.010	
3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18	.005	

Source : A portion of "Table of percentage points of the inverted beta (E) distribution, "Biometrika, vol.33 (1943) by M. Merrington and C.M. Thompson and from Table 18 of Biometrika Tables for Statisticians, vol. 1, Cambridge University Press, 1954, edited by E.S. Pearson and H.O. Hartley. Reproduced with permission of the authors, editors and Biometrika trustees.

Table 8. Factors Useful in the Construction of Control Charts.

Mean chart				Standard deviation chart				Range chart						
Sam- ple size	Factors for control limits			Factor for central line		Factors for control limits		Factors for central line		Factors for control limits				
	n	A	A ₁	A ₂	C ₂	B ₁	B ₂	B ₃	B ₄	d ₂	D ₁	D ₂	D ₃	D ₄
2	1.121	3.760	1.880	0.5642	0	1.843	0	3.267	1.128	0	3.686	0	3.267	
3	1.732	2.394	1.023	0.7236	0	1.858	0	2.568	1.693	0	4.358	0	2.575	
4	1.500	1.880	0.729	0.7979	0	1.808	0	2.266	2.059	0	4.698	0	2.282	
5	1.342	1.596	0.577	0.8407	0	1.756	0	2.089	2.326	0	4.918	0	2.115	
6	1.225	1.410	0.483	0.8686	0.026	1.711	0.030	1.970	2.534	0	5.078	0	2.004	
7	1.134	1.277	0.419	0.8882	0.105	1.672	0.118	1.882	2.704	0.205	5.203	0.076	1.924	
8	1.061	1.175	0.373	0.9027	0.167	1.638	0.185	1.815	2.847	0.387	5.307	0.136	1.864	
9	1.000	1.094	0.337	0.9139	0.219	1.609	0.239	1.761	2.970	0.546	5.394	0.184	1.816	
10	0.949	1.028	0.308	0.9227	0.262	1.584	0.284	1.716	3.078	0.687	5.469	0.223	1.777	
11	0.905	0.973	0.285	0.9300	0.299	1.561	0.321	1.679	3.173	0.812	5.534	0.256	1.744	
12	0.866	0.925	0.266	0.9359	0.331	1.541	0.354	1.646	3.258	0.924	5.592	0.284	1.716	
13	0.832	0.884	0.249	0.9410	0.359	1.523	0.382	1.618	3.336	1.026	5.646	0.308	1.692	
14	0.802	0.848	0.235	0.9453	0.384	1.507	0.406	1.594	3.407	1.121	5.693	0.329	1.671	
15	0.775	0.816	0.223	0.9490	0.406	1.492	0.428	1.572	3.472	1.207	5.737	0.348	1.652	
16	0.750	0.788	0.212	0.9523	0.427	1.478	0.448	1.552	3.532	1.285	5.779	0.364	1.636	
17	0.728	0.762	0.203	0.9551	0.445	1.465	0.466	1.534	3.583	1.359	5.817	0.379	1.621	
18	0.707	0.738	0.194	0.9576	0.461	1.454	0.482	1.518	3.640	1.426	5.854	0.393	1.608	
19	0.688	0.717	0.187	0.9599	0.477	1.443	0.497	1.503	3.689	1.490	5.888	0.404	1.596	
20	0.671	0.697	0.180	0.9619	0.491	1.433	0.510	1.490	3.735	1.548	5.922	0.414	1.586	
21	0.655	0.679	0.173	0.9638	0.504	1.424	0.523	1.477	3.778	1.606	5.950	0.425	1.575	
22	0.640	0.662	0.167	0.9655	0.516	1.415	0.534	1.466	3.819	1.659	5.979	0.434	1.566	
23	0.626	0.647	0.162	0.9670	0.527	1.407	0.545	1.455	3.858	1.710	6.006	0.443	1.557	
24	0.612	0.632	0.157	0.9684	0.538	1.399	0.555	1.445	3.895	1.759	6.031	0.452	1.548	
25	0.600	0.619	0.153	0.9696	0.548	1.392	0.565	1.435	3.931	1.804	6.058	0.459	1.541	

References

- Agarwal, B.L.** *Programming Statistics*, New Age International (P) Limited, 1996.
- Aggarwal, Y.P.** *Better Sampling: Concepts, Techniques and Evaluation*, Sterling Publishers Private Limited, 1988.
- Chandan J.S., Singh Jagjit, Khanna K.K.** *Business Statistics*, Vikas Publishing House Pvt. Ltd. 2004.
- Chandan, J. S.** *Statistics for Business and Economics*, Vikas Publishing House Pvt. Ltd. 2008.
- Gupta S.P. and Gupta M.P.** *Business Statistics (New Edition)*, Sultan Chand and Sons, 2008-2009.
- Islam, M.N.** *An introduction to Statistics and Probability (2nd Edition)*, Book World, Dhaka.
- Mendenhall, W.; Beaver, R.B. and Beaver B.M.** *Introduction to Probability and Statistics*, (10th Edition) Duxbury Press, 1999.
- Paul, N., William L.C.; Betty T.** *Statistics For Business and Economics* (6th Edition), Prentice Hall, Upper Saddle River, New Jersey 07458, 2007.
- Richard, I.L.; David S.R.** *Statistics for Management*, 7th Edition, Prentice-Hall of India, New Delhi. 2007-2008.
- Ronald E.W.** *Introduction to Statistics* (3rd Edition), Macmillan Publishing Co. Inc, New York, Collar Macmillan Publishers, London, 1982.
- Roy, M.K. and Shil, R.N.** *Mowlik Parisankhayon Parichity*, Minati Shil and Olga Roy, Chittagong. 2011.
- Roy, M.K.** *Fundamentals of Probability and Probability Distributions*, Romax Publication, Chittagong, Eighth Edition ,2011
- Sharma, J.K.** *Business Statistics*, Pearson Education (Singapore), 2004
- Vasishtha, A.R. and Vasishtha V.** *Numerical Analysis*, Kader Nath Ram Nath, Delhi, 2004.

INDEX

A

acceptance quality level 698
acceptance region 575
acceptance sampling 694
acceptance sampling, factors 696

alternative hypothesis 571

arithmetic mean 102
arithmetic mean, short-cut method 108
arithmetic mean, weighted 111

assignable causes 663

autocorrelation co-efficient 498

average outgoing quality 698

B

base period 374

Base Shifting 409

Bayes theorem 254

Bernoulli trial 290

binomial distribution 291

box-and-whisker plot 214

business statistics 5

C

causes of variation 663

c-chart 684

census 509

central limit theorem 528

central tendency 101

chain index number 404

chance causes 663

Chebyshev's Theorem 187

chi-square distribution 529

chi-square test 580

chi-square, independence test 640

chi-square, variance test 638

circular test 402

class 41

class boundary 62

class limit 60

coding 36

coding time variable 444

co-efficient of

determination 357

co-efficient of variation

182

components of time series

data 432

composite hypothesis 572

conditional probability

241

confidence interval 551

confidence interval, width

553

consistency 550

construction of frequency distribution 59

construction of index

number 375

consumer's price index 415

consumer's risk 697

continuous variable 13

control chart 667

control chart, attribute 683

control chart for mean 670

control chart for range 670

control chart for SD 682

convenient sampling 523

correlation 312

correlation analysis 312

correlation and regression, difference 356

correlation co-efficient,

test 525

correlation coefficient 314

correlation co-efficient,

properties 316

correlation, rank 332

correlation, simple 314

C

cost of living index 415
critical value 576
critical region 575

cross section data 22
cross tabulation 95
cumulative frequency 63, 84
current period 374

curve method, interpolation 713

cyclical fluctuations 437
cyclical fluctuations,
measurements 487

D
data 9
data collection, methods 28
data processing 35
data, types 18
data, grouped 67
data, ungrouped 67
d-chart 684
deciles 137
defect, definiton 683
defective, definiton 683
deflating index number 412
dependent variable 349
descriptive statistics 23
deseasonalization 486
determinaiton of sample size 558
diagram, Venn 225
diagram, tree 225
diagram, component bar 49
diagram, multiple bar 48
diagram, pareto 53
diagram, pie 50
diagram, scatter 96, 319

diagram, simple bar 46
diagram, stem and leaf 67
difference table 714
distribution, frequency 57
discrete variable 13
dispersion 162

dispersion, relation 181
dispersion, measures 163
distribution of difference
between two sample means 544
distribution of proportion 547
distribution of sample mean
532
distribution, binomial 291
distribution, normal 301
distribution, poisson 297
distribution, standard normal
303
Dorbish and Bowley index 390
double sampling plan 695

E

econometric models 491
editing 36
efficiency 551
errors in decision 572
estimation 549
estimation, interval 551
estimation, point 550
estimator 549
event space 232
events 229
events, complementary 231
events, compound 229
events, impossible 230
events, independent 242
events, mutually exclusive 230
events, sure 230
exclusive method 60
expectation 281
expected frequency 641
experiment, random 229
experiments 228
exploratory data analysis 212
extrapolatin 710, 723
extrapolation 490,

F
factor reversal test 401
F-distribution 531
finite difference method 714
Fisher's Ideal Price Index 390
five-number summary 212

forecasting 487

forecasting, methods 491

frequency 39

frequency, cumulative 63, 84

frequency distribution 57

frequency distribution, construction 59

frequency polygon 77

frequency relative 42

F-test 581

G

graph, dot plot 71

graph, histogram 72

graphic method, interpolation 711

graphic method 439

graphic method, trend 439

H

histogram 72

hypothesis 570

hypothesis testing 569

hypothesis, alternative 571

hypothesis, null 570

hypothesis, statistical 570

I

inclusive method 62

independent events 242

independent variable 349

index number 373

index number, characteristics 374

index number, classification 378

index number, construction 375

index number, cost of living 415

index number, features 379

index number, Fisher's 390

index number, Laspeyeres 386

index number, Paasches 388

index number, tests 399

index number, unweighted 381

index number, uses 375

inferential statistics 23

interpolation 710

interpolation, calculus method 714

interpolation, curve method 713

interpolation, graphic method 711

inter-quartile range 165

interval scale 17

inverse interpolation 720

irregular variation 438

J

joint probability 240

judgment sampling 522

K

Kelly's price index 394

kurtosis 208

L

lack of control 672

Lagrange's Formula 719

Laspeyres price index 386

laws of probability 247

least squares method 351, 444

level of significance 573

line graph 88

linear trend 446

link relative method 481

lottery method 515

M

marginal probability 240

Marshall-Edgeworth index 393

mean chart 670

mean deviation 168

mean, geometric 148

mean, harmonic 152

measurement of trend 439

measurement of cyclical variation 487

 seasonal variation 466

median 116

median, from graph 121

methods of forecasting 491

mid point 62

mid-range 155

mode 128

mode, from graph 134

- models of time series, 433
moments 200
moving average method 454
- N**
Newton's forward formula 715
nominal scale 15
non-linear trend 454
non-sampling error 525
normal distribution 301
null hypothesis 570
- O**
observed frequency 641
OC curve 696
ogive 84
one tailed test 574
ordinal scale 16
outliers, causes 190
outcomes 228
outcomes, equally likely 232
outcomes, exhaustive 233
outcomes, favourable 232
outcomes, mutually exclusive 232
outliers 127
- P**
Paasche's Price Index 388
paired t-test 580, 613
parameter 10
p-chart 683
Pearson's correlation co-eff 314
percent 40
percentiles 138
- pictogram 56
point estimation 550
Poisson distribution 297
population 8, 509
population covariance 313
power of a test 573, 645
price index, simple average 381
price index 378
primary data 27
- probability definition 234
probability, multiplicative law 229
- probability density function 280
probability distribution 290
probability function 275
probability, statistical 238
probability, axiomatic 239
probability, classical 234
probability, joint 240
probability, marginal 240
probability, subjective 240
probable error 325
probability, addition law 247
probability, conditional 241
process control 667
producer's risk 697
product control 667, 692
proportion 40
proportion test, single 616
proportion test, two 621
p-value 577
- Q**
qualitative variable 11
quality control, definition 662
quantitative variable 12
Quantity index numbers 396
quantity index 379
quartile deviation 165
quartiles 135
questionnaire 30
quota sampling 521
- R**
ratio scale 17
random number table 516
random variable 274
random variable, continuous 279
random variable, discrete 274
random variable, mean 281
random variable, variance 283
range 164
range chart 670
rank correlation 332
rate 41
ratio 40
ratio to moving average method 477
ratio to trend method 472

regression analysis 348, 490
regression co-efficient test 527
regression co-efficient, properties 355
regression model, forms 353
relative frequency 42
relative standing 190
residual method 487
S
sampling 10, 511
sample 10
sample points 229
sample size, determination 558
sample space 229
sampling distribution 526
sampling error 524
sampling frame 511
sampling inspection 693
sampling methods 514
sampling plan, double 695
sampling plan, sequential 696
sampling plan, single 695
scale of measurement 14
scatter diagram 96, 319
schedule 31
seasonal variation 435
seasonal variation, measurement 466
secondary data 27, 34
secular trend 434
secular trend, measurement 439
semi-average method 441
sequential sampling plan 696
set theory 220
set theory, addition law 224
Sheppard's correction 202
simple average method 468
simple average price index 381
simple correlation 314

simple exponential method 492
simple hypothesis 571
simple random sampling 514
single mean test 584, 592
single sampling plan 695
skewness 204
smoothing technique 491
snowball sampling 523
Spearman's correlation co-eff 332
Splicing index number 410
standard deviation 173
standard error 527
standard normal distribution 303
statista 1
statistic 10
statistical hypothesis 570
statistics, definition 4
statistics, function 6
statistics, limitations 6
statistique 1
steps of hypothesis test 581
stratified random sampling 518
student's t distribution 530
sufficiency 551
systematic sampling 520
T
test of correlation co-efficient 525
test of independence of attributes 640
test of regression co-efficient 525
test statistic 575
theorem on total probability 253
time reversal test 399
time series data 22, 431
time series, component 432
translating time variable 444
trend, measurement 439
t-test 579

t-test , two mean 610

t-test, single mean 592

two tailed test 574

two-mean test 603

type I error 572

type II error 572

U

unbiasedness 550

unit test 402

unweighted index number 381

V

value index 379

value index number 399

variable, independent 349

variable, dependent 349

variable, random 274

variables, type 11

variance 173

variance test 638

W

**weighted aggregative price
index** 386

**Weighted Average of Price
Relatives** 394

Z

Z, standard normal 303

Z-test 579

Z-test, single mean 584

Z-test, single proportion 616

Z-test, two means 603

Z-test, two proportion 621

Z-test 579

Professor Manindra Kumar Roy has been teaching in the University of Chittagong for more than forty years. His teaching and research interest is on Mathematical Statistics, especially on Probability Distribution and Statistical Inference. A member of several National and International Learned Societies, Dr. Roy served in the Garyounis University, Libya, as a faculty in the mid-eighties. Author of a number of Statistics Texts, Professor Roy has many scientific papers published in national and international reputed journals to his credit.

Professor Jiban Chandra Paul has been teaching in the University of Chittagong for about 28 years. His teaching and research interest is on Econometrics, Time Series Analysis and Forecasting. A member of several National and International Learned Societies, Dr. Paul served in the Asmara University, Eritrea, as a Professor and Head of the Department of Statistics and Demography for more than three years. Professor Paul has many scientific papers published in national and international reputed journals to his credit.

OTHER BOOKS OF **PROFESSOR MANINDRA KUMAR ROY**

Fundamental of Probability &
Probability Distributions, 1991

Moulik Parisankhyan Pariciti, 1994

Parisankhyan Binyash Pariciti Vol 1, 1996

Parisankhyan Binyash Pariciti Vol 2, 1996

Uccha Madhyamik Parisankhyan 1st Paper, 1999

Uccha Madhyamik Parisankhyan 2nd Paper, 1998

Uccha Madhyamik Parisankhyan Babaharic, 2000