MS 104

Quantitative Techniques in Management

Block - I
Block Title : Introduction to Statistics

UTTARAKHAND OPEN UNIVERSITY
SCHOOL OF MANAGEMENT STUDIES AND COMMERCE
University Road, Teenpani By pass, Behind Transport Nagar, Haldwani- 263 139
Phone No: (05946)-261122, 261123, 286055
Toll Free No.: 1800 180 4025
Fax No.: (05946)-264232, e-mail: info@uou.ac.in, som@uou.ac.in
http://www.uou.ac.in
www.blogsomcuou.wordpress.com

Board of Studies

Programme Coordinator

Dr. Manjari Agarwal
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

Units Written By		Unit No.
Text material developed by	Devashish Dutta	
Typeset by	Goswami Associates, Delhi	

Editor(s)

ISBN : 978-93-85740-10-7

Edition : 2016 (Restricted Circulation)
Published by : Uttarakhand Open University, Haldwani, Nainital - 263 139
Printed at : Laxmi Publications (P) Ltd., New Delhi
DUO-8156-69.62-QUAN TECH MGMT B-I

CONTENTS

Course Credits: 6

Course Objective: The objective of this course is to provide students the knowledge of Quantitative Techniques tools and their application in various decision-making situations.

Block I:	**Introduction to Statistics**
Unit I:	Introduction to Statistics
	Business Statistics – Concept, Significance and Limitations
Unit II:	Data – Types and Classification
	Primary and Secondary Data, Classification and Tabulation
Unit III:	Frequency Distribution and Graphical Representations
Unit IV:	Measures of Central Tendency
	Mean, Median, Mode and Quartile
Unit V:	Measures of Dispersion
	Range, Mean Deviation, Standard Deviation
Block II:	**Measurement of Variation, Correlation and Regression**
Unit VI:	Measures of Skewness, Kurtosis and Moments
Unit VII:	Correlation
	Correlation–Karl Pearson and Rank Correlation-Partial-Multiple
Unit VIII:	Regression Analysis and Properties of Regression Coefficients
	Properties of Regression Coefficients and Relationship between Regression and Correlation
Unit IX:	Times Series Analysis
Block III:	**Probability and Distribution**
Unit X:	Probability – Definition and Classification
	Probability Definition and Classification of Probability
Unit XI:	Laws of Probability
	Additive Law, Distribution and Multiplication Law, Joint Probability
Unit XII:	Probability Distribution
	Probability Distribution, Discrete and Continuous Distribution
Unit XIII:	Binomial Distribution
Unit XIV:	Normal and Poisson Distribution
Block IV:	**Operation Research**
Unit XV:	Linear Programming
	Graphical Solution Method-Simplex Method-Duality-Bounded Variables LP Problems-Parametric –Integer-Goal Programming
Unit XVI:	Transportation Problem
Unit XVII:	Assignment Problem
Unit XVIII:	Queueing Theory and Decision Theory
Unit XIX:	Replacement Theory and Sequencing Problems
Unit XX:	PERT and CPM

UNIT 1: INTRODUCTION TO STATISTICS

Structure

1.0 INTRODUCTION

Statistics is a set of decision making techniques which helps businessmen in making suitable policies from the available data. In fact, every businessman needs a sound background of statistics as well as of mathematics. The purpose of statistics and mathematics is to manipulate, summarize and investigate data so that the useful decision making results can be executed. The term 'statistics' has been derived from the Latin word '*status*' Italian word '*statista*' or German word '*statistik*'. All these words mean 'Political state'. In ancient days, the states were required to collect statistical data mainly for the number of youngmen so that they can be recruited in the Army. Also to calculate the total amount of land revenue that can be collected. Due to this reason, statistics is also called 'Political Arithmetic'.

Meaning of statistics. According to **Bowley**,

"Statistics are numerical statements of facts in any department of enquiry placed in relation to each other."

According to **Yule** and **Kendall,**

"By statistics, we mean quantitative data affected to a marked extend by multiplicity of causes."

According to **Horace Secrist,**

"By statistics, we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according

NOTES

to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other."

According to **Croxton** and **Cowden,**

"Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data."

Uses of Statistics and Mathematics in Business Decision Making

(a) ***Uses of Statistics in Business***

The following are the main uses of statistics in various business activities.

(i) With the help of statistical methods, a quantitative information about production, sale, purchase, finance, etc. can be obtained. This type of information helps the businessmen in formulating suitable policies.

(ii) By using the techniques of time series analysis which are based on statistical methods, the businessman can predict the effect of a large number of variables with a fair degree of accuracy.

(iii) In business decision theory, most of the statistics techniques are used in taking a business decision which helps us in doing the business without uncertainty.

(iv) Now a days, a large part of modern business is being organised around systems of statistical analysis and control.

(v) By using 'Bayesian Decision Theory', the businessmen can select the optimal decisions for the direct evaluation of the payoff for each alternative course of action.

(b) ***Uses of Mathematics for Decision Making***

(i) The number of defects in a roll of paper, bale of cloth, sheet of photographic film can be judged by means of Control Chart based on Normal distribution.

(ii) In statistical quality control, we analyse the data which are based on the principles involved in Normal curve.

(c) ***Uses of Statistics in Economics***

Statistics is the basis of economics. The consumer's maximum satisfaction can be determined on the basis of data pertaining to income and expenditure. The various laws of demand depend on the data concerning price and quantity. The price of a commodity is well determined on the basis of data relating to its buyers, sellers, etc.

1.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define the term 'Statistics' and its uses in business decision making
- Explain importance of statistics in business and industry
- Explain limitations of statistics

1.2 IMPORTANCE OF STATISTICS

Statistics in today's life has become an essential part of various business activities which is clear from the following points.

(a) ***Importance of Statistics in Business and Industry***

In past days, decisions regarding business were made only on personal judgement. However, in these days, they are based on several mathematical and statistical techniques and the best decision is arrived by using all these techniques. For example, by using the testing hypothesis, we can reject or accept the null hypothesis which are based upon the assumption made from the population or universe.

By using 'Bayesian Decision Theory' or 'Decision Theory', we can select the optimal decisions for the direct evaluation of the payoff for each alternative course of action. Mathematics and statistics have become ingredients of various decisions problems which is clear from the following :

(i) ***In Selecting Alternative Course of Action:*** The process of business-decisions involve the selection of a single action among some set of alternative actions. When there are two or more alternative courses of action, and we need only one course of action, statistical decisions theory helps us in selecting the required course of action by applying Bayesian decision theory and thus saves lot of time.

(ii) ***In Removing Uncertainty:*** In decision-making problems, uncertainty is very common in a situation, when the course of action is not known to us. When there are many possible outcomes of an event, we cannot predict with certainty that what will happen. By applying the concept of joint and conditional probability, the uncertainty about the event can be removed very easily.

(iii) ***In Calculating E.O.L., C.O.L., etc.:*** In business, the opportunity loss is very often, which can be defined as the difference between the highest possible profit for an event and the actual profit obtained for the actual action taken. The expected opportunity loss (E.O.L.) and conditional

NOTES

opportunity loss (C.O.L.) can be easily calculated by using the concept of maximum and minimum criteria of pay-off.

(b) ***Importance in the Field of Science and Research***

Statistics has great significance in the field of physical and natural sciences. It is widely used in verifying scientific laws and phenomenon. For eample, to formulate standards of body temperature, pulse rate, blood pressure, etc. Success of modern computers depends on the conclusions drawn on the basis of statistics.

(c) ***Importance in the Field of Banking***

In banking industry, the bankers have to relate demand deposits, time deposits, credit etc. It is on the basis of data relating to demand and time deposits that the bankers determine the credit policies. The credit policies are based on theory of probability.

Check Your Progress

Fill in the blanks:

1. Statistics is a set of ..... which helps business in making suitable policies from the available data.
2. With the help of statistical methods, ... about production, sale, purchase, finance, etc. can be obtained.
3. Statistics is also called .....
4. By using ....., we can select the optimal decisions for the direct evaluation of the payoffs for each alternative course of action.
5. The number of defects in a roll of paper, bale of cloth can be judged by means of ..... based on normal distribution.

1.3 LIMITATIONS OF STATISTICS

Statistics is considered to be a science as well as an art, which is used as an instrument of research in almost every sphere of our activities. There are limitations of statistics also. Care must be taken of these limitations while using statistical methods. *Newsholme* has well said, "It must be regarded as an instrument of research of great value but having several limitations which are not possible to overcome and as such they need our careful attention". Some of the limitations of statistics are as follows:

1. ***Statistics Suits to the Study of Quantitative Data Only*:** Statistics deals with the study of quantitative data only. By using the methods of statistics, the problems regarding production, income, price, wage, height, weight etc. can be studied. Such characteristics are quantitative in nature. The characteristics

like honesty, goodwill, duty, character, beauty, intelligence, efficiency, integrity etc. are not capable of quantitative measurement and hence cannot be directly dealt with statistical methods. These characteristics are qualitative in nature. In such type of characteristics, only comparison is possible. The statistical methods may be tried in studying qualitative characteristics only if they are expressed quantitatively. For example, the efficiency of workers in a hand made paper factory, may be studied by considering the number of paper sheets prepared daily by each worker. The use of statistical methods is limited to quantitative characteristics and those qualitative characteristics which are capable of being expressed numerically.

2. ***Statistical Results are not Exact*:** The task of statistical analysis is performed under certain conditions. It is not always possible, rather not advisable, to consider the entire population during statistical investigations. The use of samples is called for in statistical investigations. And the results obtained by using samples may not be universally true for the entire population. Data collected for a statistical enquiry may not be hundred percent true. Statistical results are true on an average. If we comment that the students of a particular class are intelligent, it does not necessarily imply that each and every student of the class is intelligent. The probability of getting a head in a single trial of an unbiased coin is 1/2, but we may not get exactly one head in two trials of the coin. That is why, statistics is not considered an exact science like Physics, Mathematics etc.
3. ***Statistics Deals with Aggregates Only*:** Statistics does not recognise individual items. Consider the statement, "The weight of Mr. X in the college is 70 kg". This statement does not constitute statistical data. Statistical methods are not going to investigate anything about this statement. Whereas, if the weights of all the students of the college are given, the statistical methods may be applied to analyse that data. According to *Tippett*, "Statistics is essentially totalitarian because it is not concerned with individual values, but only with classes". Statistics is used to study group characteristics of aggregates. If we are given the profit figure of a firm manufacturing a particular item, it does not help in commenting on the performance of the company. On the other hand, if we are given the profit figures of the firm for the ten or fifteen consecutive years, we can make use of statistical methods to comment on the performance of the firm.
4. ***Statistics is Useful for Experts Only*:** Statistics is both a science and an art. It is systematic and find applications in studying problems in Economics, Business, Astronomy, Physics, Medicines etc. Statistical methods are sophisticated in nature. Everyone is not expected to possess the intelligence required to understand and to apply these methods to practical problems. This is the job of

NOTES

an expert, who is well-versed with statistical methods. *W.I. King* says, "Statistics is a most useful servant but only of great value to those who understand its proper use".

A skilled statistician can never advise the public to walk in the middle of the road just on the plea that the number of padestrians who died in road accidents during 1975-86, on a particular road, was less for those who walked in the middle than those who walked on the road side. Such statements can only be given by those who cannot be considered as experts in using statistical methods. If such a type of decision is to be taken, then we must also consider the number of persons, who walk on the middle and on the sides of the road. In fact, if some body, accepts the advice of such an expert and start walking in the middle of the road, he is not hoped to survive much longer. *Yule* and *Kendall* has rightly said that "statistical methods are most dangerous tools in the hands of inexperts".

The methods of statistics must be tried by experts only. The results derived by using statistical methods would very much depends upon the skill of the user. According to *King*, "Statistics are like clay of which one can make a god or devil as one pleases".

5. ***Statistics does not Provide Solutions to the Problems:*** The statistical methods are used to explore the essentials of problems. It does not find use in inventing solutions to problems. For example, the methods of statistics may reveal the fact that the average result of a particular class in a college is deteriorating for the last ten years, *i.e.*, the trend of the result is downward, but statistics cannot provide solution to this problem. It cannot help in taking remedial steps to improve the result of that class. Statistics should be taken as a means and not as an end. The methods of statistics are used to study the various aspects of the data.

Check Your Progress

State whether the following statements are True or False:

6. Statistics is only a science.
7. Statistical results are not exact.
8. Statistical methods are sophisticated in nature.
9. Statistics deals with quantitative as well as qualitative data both.
10. The credit policies are based on theory of probability.

1.4 SUMMARY

- Statistics is a set of decision making techniques which helps businessmen in making suitable policies from the available data.
- "Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data."
- With the help of statistical methods, a quantitative information about production, sale, purchase, finance etc. can be obtained.
- Statistics is the basis of economics. The consumer's maximum satisfaction can be determined on the basis of data pertaining to income and expenditure.
- By using 'Bayesian Decision Theory' or 'Decision Theory', we can select the optimal decisions for the direct evaluation of the payoff for each alternative course of action.
- Statistics is considered to be a science as well as an art, which is used as an instrument of research in almost every sphere of our activities.
- *W.I. King* says, "Statistics is a most useful servant but only of great value to those who understand its proper use."
- *Yule* and *Kendall* has rightly said that "statistical methods are most dangerous tools in the hands of inexperts."
- The statistical methods are used to explore the essentials of problems.

1.5 GLOSSARY

- ***Statistics:*** Statistics are numerical statements of facts in any department of enquiry placed in relation to each other.
- ***Opportunity loss:*** Opportunity loss can be defined as the difference between the highest possible profit for an event and the actual profit obtained for the actual action taken.
- ***Conditional opportunity loss (E.O.L.):*** Conditional opportunity loss (C.O.L.) can be easily calculated by using the concept of maximum and minimum criteria of pay-off.
- ***Demand deposits, time deposits, credit:*** It is on the basis of data relating to demand and time deposits that the bankers determine the credit policies.

1.6 ANSWERS TO CHECK YOUR PROGRESS

1. Decision Making Techniques
2. Quantitative Information
3. Political Arithmetic
4. Decision Theory
5. Control Charts
6. False
7. True
8. True
9. False
10. True

1.7 TERMINAL AND MODEL QUESTIONS

1. Discuss the role of mathematics and statistics in business decisions.
2. List any five uses of mathematics and statistics in business.
3. Statistics is said to be both science and art. Comment.
4. In what sense do we say that a basic knowledge of statistics is essential to become an efficient businessman in a modern world.
5. Discuss the importance and limitations of statistics.

1.8 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 2: DATA-TYPES AND CLASSIFICATION

Structure

2.0 INTRODUCTION

The first step in a statistical investigation is the planning of the proposed investigation. After planning, the next step is the collection of data, keeping in view the object and scope of investigation. There are number of methods of collecting data. The mode of collection of data also depends upon the availability of resources. The collected data would be edited, presented, analysed and interpreted. If the job of data collection is not done sincerely and seriously, the results of the investigation is bound to be inaccurate and misleading. And so the resources used in performing the other steps would be wasted and the purpose of the investigation would be defeated.

Types of Data

I. Primary Data
II. Secondary Data

2.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Differentiate between primary and secondary data
- Explain various methods of collection of primary data

- Explain various methods of collection of secondary data
- Define classification of data
- Explain various types of classification of data
- Define tabular presentation or tabulation of data and its significance
- Explain various types of tabulation

2.2 PRIMARY DATA

Definition

Data is called **primary,** if it is originally collected in the process of investigation. Primary data are original in nature. Primary data are generally used in case of some special purpose investigation. The process of collecting primary data is time consuming. For example suppose we want to compare the average income of employees of two companies. This can be done by collecting the data regarding the incomes of employees of both companies. The data collected would be edited, presented and analysed by taking averages of both groups of data. On the basis of the averages, we would be able to declare as to the average income for which company is more. The data used in this investigation is primary, because the data regarding the income of employees was collected during the process of investigation.

Methods of Collecting Primary Data

(i) Direct personal investigation

(ii) Indirect oral investigation

(iii) Through local correspondents

(iv) Through questionnaires mailed to informants

(v) Through schedules filled by enumerators.

Now we shall discuss the process of collecting primary data by these methods. We shall also discuss the suitability, merits and demerits regarding the above mentioned methods of collecting primary data.

(i) ***Direct Personal Investigation***

In this method of collecting data, the investigator directly comes in contact with the informants to collect data. The investigator himself visits the different informants, covered in the scope of the investigation and collect data as per the need of the investigation. Suppose an investigator wants to use this method to collect data regarding the wages of the employees of a factory then he would have to contact each and every employee of the factory in order to collect the required data. In the

context of this method of collecting primary data, *Professor C.A. Moser* has remarked, "In the strict sense, observation implies the use of the eyes rather than of the ears and the voice". The suitability of this method depends upon the personality of the investigator. The investigator is expected to be tactful, skilled, honest, well behaved and industrious. It is suitable when the area to be covered is small. This is also suitable when the data is to be kept secret.

Merits

(a) The degree of accuracy in data collected is very high.

(b) Because of the personal visit of the investigator, the response of informants is expected to be very encouraging.

(c) The data collected is reliable.

(d) This method is flexible in the sense that the investigator can modify the nature of data to be collected in accordance with the prevailing circumstances.

(e) Consistency and homogeneity is present in the data collected.

(f) Data regarding complex and sensitive questions can also be gathered by twisting the questions, as per the need.

Demerits

(a) This method is very expensive and time consuming.

(b) This method is not useful when the informants are spread over a wide area.

(c) This method is not useful in case the investigator is expected to be biased.

(ii) ***Indirect Oral Investigation***

In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the data about the informants is collected from some selected persons who are expected to be acquinted with the informants as well as the object of the investigation. A person giving data about the informants is called 'witness.'

The suitability of this method depends upon the personalities of the investigator and the witnesses. The investigator is expected to be tactful, skilled, honest and well behaved. At the same time a person giving information about the informants is expected to be unbiased and well acquainted with the object of the investigation. This method is suitable in the cases where the informants are not expected to give data frankly, when contacted directly. Suppose an investigator wants to collect data regarding the level of intelligence of the students in different classes of a college. The investigator may not be able to get the required data by contacting the students and asking them about their intelligence. The required information in this case can be obtained by contacting the class teachers or the Head of the Institution. This method of collecting data is also useful when the area to be covered is very large.

Merits

(a) This method is economical in respect of money, labour and time.

(b) This method can be easily used even if the area to be covered is widely spread.

(c) In this method, the data is collected from the third person and so the data is not affected by the bias on the part either the informants or the investigator.

(d) In this method, the investigator can use the views and knowledge of experts by contacting them during the process of collecting data.

Demerits

(a) The data is expected to be highly affected by the bias of the witnesses. They can misguide the investigator by distorting the data.

(b) The data collecting is not expected to be accurate due to the carelessness of the witnesses.

(iii) ***Through Local Correspondents***

In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the data about the informants is collected and sent to the investigator by the local correspondents, appointed by the investigator. Newspaper agencies collect data by using this method. They appoint their correspondents area wise. The correspondents themselves send the desired data to the offices of their respective newspaper. The suitability of this method depends upon the personality of the correspondent. He is expected to be unbiased, skilled and honest. To eliminate the bias of the correspondents, it is advisable to appoint more than one correspondent in each area.

Merits

(a) This method of collecting data is most economical in respect of money, labour and time.

(b) This method can be easily used even if the area to be covered is widely spread.

(c) This method is specially recommended for investigations, in which data is to be collected on regular basis.

Demerits

(a) The data collected is not expected to be very accurate.

(b) The data collected is expected to be affected by the bias of the correspondents.

(c) The data collected is not expected to be very reliable.

(iv) ***Through Questionnaires Mailed to Informants***

In this method of collecting data, the informants are not directly contacted by the investigator but instead the investigator send questionnaires by post to the informants with the request of sending them back after filling the same. The suitability of this method depends upon the quality of the 'questionnaire' and the response of the informant. This method is useful when area to be covered is widely spread. This method would not work in case the informants are illiterate or semi-literate.

Merits

(a) This method is economical in respect of money, labour and time.

(b) This method can be very easily used if the area to be covered is widely spread.

Demerits

(a) This method is not useful in case the informants are illiterate or semi-literate.

(b) The collected data is not expected to be very accurate due to lack of seriousness in the informants.

(c) The response of the informant is expected to be poor because people generally avoid giving reply in written statement form.

(d) The reliability of the data collected cannot be judged by the investigator.

(e) The data may be unduly affected by the expected bias of the informants.

(v) ***Through Schedules Filled by Enumerators***

In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the enumerators are deputed to contact the informants and to fill the schedules on the spot, after collecting data as per the need of the schedule. The basic difference between this method and the previous method is that, in this method the schedules are filled by the enumerators after getting information from the informants, whereas in the previous method, the questionnaires were to be filled by the informants themselves. The suitability of this method depends upon the enumerators. The enumerators are expected to be skilled, honest, hard working, well-behaved and free from bias. This method of collecting data is suitable in case the informants are illiterate or semi-literate. In our country, census data about all the citizens is collected after every ten years by using this method.

Merits

(a) In this method of collecting data, the degree of accuracy is expected to be very high.

(b) The data collected is very reliable.

NOTES

(c) Because of the direct contact between the enumerators and the informants, data can also be gathered about sensitive questions by twisting the questions accordingly.

(d) The data collected is least affected by the bias of the enumerators and the informants.

(e) This method can also be used in case the informants are illiterate or semi-literate.

Demerits

(a) This method of collecting data is very expensive.

(b) Much time is taken in collecting data.

(c) The enumerator have to be trained before they are deputed to start collecting data.

(vi) ***Requisites of a Good 'Questionnaire' and 'Schedule'***

In the last two methods of collecting primary data, we discussed the method of questionnaires to be filled by the informants and the method of filling schedules by the enumerators. In fact, there is no fundamental difference between a questionnaire and a schedule. Both questionnaire and schedule contain some questions. The only difference between the two is that the former is filled by the informants themselves, whereas in the case of later, the data concerning the informants is filled by the enumerators. The success of collecting data by using either questionnaire or schedule depends upon the quality of itself. Preparation of questionnaire and schedule is an art. Now we shall discuss in detail the requisites of a good questionnaire and schedule.

(i) ***Forwarding letter:*** The investigator must include a forwarding letter in case of sending questionnaires to the informants. The investigator must request the informants to fill in the same and to return it back after filling it. The object of the investigation should also be mentioned in the latter. The informants should also be ensured that the filled questionnaires would be kept secretly, if desired. To encourage the response of informants, special concessions and free gifts may be offered to the informants.

(ii) ***Questions should be minimum in number:*** The number of questions in a questionnaire or a schedule should be as small as possible. Unnecessary questions should never be included. Inclusion of more than 20 or 25 questions would be undesirable.

(iii) ***Questions should be Easy to understand:*** The questions included in a questionnaire or a schedule should be Easy to understand. The questions should not be confusing in nature. The language used should also be simple and the use of highly technical terms should also be avoided.

(iv) ***Questions should be logically arranged:*** The questions in a questionnaire or a schedule also be logically arranged. The questions should be arranged so that there is natural and spontaneous reaction of the informants to the questions. It is not fair to ask the informant whether he is employed or unemployed after asking his monthly income. Such sequence of questions create bad impression on the mind of the informants.

(v) ***Only well-defined terms should be used in questions:*** In drafting questions for a questionnaire or a schedule, only well defined terms should be used. For example, the term 'income' should be clearly defined in the sense whether it is to include allowances etc. along with the basic income or not. Similarly, in case of businessman, whether the informants are to inform about their gross profits or net profits etc.

(vi) ***Prohibited questions should not be included:*** No such question should be included in the questionnaire or schedule which may immediately agitate the mind of the informants. Question like, "Have you given up the habit of telling a lie" or "How many times in a month, do you quarrel with your wife", would immediately mar the spirit of the informants.

(vii) ***Irrelevant questions should be avoided:*** In questionnaire or schedule, only those questions should be included which bears direct link with the object of the investigation. If the object is to study the problem of unemployment, then it would be useless to collect data regarding the heights and weights of the informants.

(viii) ***Pilot survey:*** Before the questionnaire is sent to all the informants for collecting data, it should be checked before hand for its workability. This is done by sending the questionnaire to a selected sample and the replies received are studied thoroughly. If the investigator finds that most of the informants in the sample have left some questions un-answered then those questions should be modified or deleted altogether, provided the object of the investigation permits to do so. This is called *Pilot Survey*. Pilot Survey must be carried out before the questionnaire is finally accepted.

Check Your Progress

Fill in the blanks:

1. Primary data are generally used in case of some .....
2. A person giving data about the informants is called .....
3. The method 'Through Questionnaires Mailed to Informants' would not work, in case the informants are ..... or .....
4. Pilot survey must be carried out before the ..... is finally accepted.
5. In our country, ... about all the citizens is collected after every 10 years by using the method of collected data.

NOTES

SPECIMEN OF QUESTIONNAIRE

Department of Industries,
.....

Dear student,

We are conducting this investigation to know about the future planning of our students, studying at present in colleges. The Govt. has decided to allocate funds for helping deserving students in establishing their career. You are requested to fill in the enclosed questionnaire and send it to back. A stamped self-addressed envelope has also been enclosed for your convenience.

Thanking you,

Yours sincerely,
.....

QUESTIONNAIRE

Department of Industries

Object: Future Planning of Students (1986)

1. Name .....
2. Class .....
3. Occupation of your father .....
4. Address .....
5. Which of the following occupation would you like to choose after completing your education
   - (*a*) Business ( )
   - (*b*) Service ( )
   - (*c*) Manufacturing concern ( )
   - (*d*) Doctor ( )
   - (*e*) Any other ( )
6. Why do you want to choose this occupation ?
   - (*a*) Less competition ( )
   - (*b*) Greater possibility of money making ( )
   - (*c*) Future is bright ( )
   - (*d*) Parental occupation ( )
7. If you plan to establish a manufacturing concern
   - (*a*) What amount of foreign exchange you can earn .....
   - (*b*) Which of the countries can be the expected buyers of your product .....
8. Sources of raw material .....
9. Sources of machinery .....
10. Gestation period .....
11. What type of help would you seek from the Govt.
    - (*a*) Financial ( )
    - (*b*) Technical knowledge ( )
    - (*c*) Imported machinery ( )
    - (*d*) Rent free land ( )

(*Contd.*)

12. State the amount of financial help which you would like to have
.....
13. In what way, your product would be superior to those of producing the same product .....
14. In what way, you would be helping the nation to prosper
.....

SPECIMEN OF SCHEDULE

Government of Haryana

Ration Card

Sr. No. .....

Name of City/Village .....

Name of Head of Family .....

Name of Father/Husband .....

House No. .....

Ward/Sector No. .....

Detail of Family

No. of units

Cereal ( ) Sugar ( )

....

Sig./Thumb Impression of Head of Family

Signature and Designation of authorised officer

Sr. No.	*Name*	*Relation with Head of Family*	*Age*
____	____	____	____
____	____	____	____

Total number of members:

Above 12 Years ( ) Between 2 Years and 12 Years ( )

Below 2 Years ( )

To be filled in by depot holder

S. No. and Address of the Depot .....

Registration No. .....

.....

Sig. and stamp of the Depot Holder

NOTES

2.3 SECONDARY DATA

Definition

Data is called **secondary** if it is not originally collected in the process of investigation, but instead, the data collected by some other agency is used for the purpose. If the investigation is not of very special nature, then the use of secondary data may be made provided that can serve the purpose. Suppose we want to investigate the extent of poverty in our country, then this investigation can be carried out by using the national census data which is obtained regularly after every 10 years. The use of secondary data economise the money spent. It also reduces the time period of investigation to a great extent. If in an investigation some secondary data could be made use of, then we must use the same. The secondary data are ought to be used very carefully. In this context, *Connor* has remarked, "Statistics, especially other peoples' statistics are full of pitfalls for the user."

Methods of Collecting Secondary Data

(i) Collection from Published Data

(ii) Collection from Un-published Data.

(i) ***Collection From Published Data***

There are agencies which collect statistical data regularly and publish it. The published data is very important and is used frequently by investigators. The main sources of published data are as follows:

(a) ***International publications:*** International Organisations and Govt. of foreign countries collect and publish statistical data relating to various characteristics. The data is collected regularly as well as on ad-hoc basis. Some of the publications are:

(i) U.N.O. Statistical Year Book

(ii) Annual Reports of I.L.O.

(iii) Annual Reports of the Economic and Social Commission for Asia and Pacific (ESCAP)

(iv) Demography Year Book

(v) Bulletins of World Bank.

(b) ***Goverment publications:*** In India, the Central Govt. and State Govt. collects data regarding various aspects. This data is published and is found very useful for investigation purpose. Some of the publications are:

(i) Census Report of India

(ii) Five-Year Plans

(iii) Reserve Bank of India Bulletin

(iv) Annual Survey of Industries

(v) Statistical Abstracts of India.

(c) ***Report of commissions and committees:*** The Central Govt. and State Govt. appoints Commissions and Committees to study certain issues. The reports of such investigations are very useful. Some of these are:

(i) Reports of National Labour Commission

(ii) Reports of Finance Commission

(iii) Report of Hazari Committee etc.

(d) ***Publications of research institutes:*** There are number of research institutes in India which regularly collect data and analyse it. Some of the agencies are:

(i) Central Statistical Organisation (C.S.O.)

(ii) Institute of Economic Growth

(iii) Indian Statistical Institute

(iv) National Council of Applied Economic Research etc.

(e) ***Newspapers and Magazines.*** There are many newspapers and magazines which publish data relating to various aspects. Some of these are:

(i) Economic Times

(ii) Financial Express

(iii) Commerce

(iv) Transport

(v) Capital etc.

(f) ***Reports of trade associations:*** The trade associations also collect data and publish it. Some of the agencies are:

(i) Stock Exchanges

(ii) Trade Unions

(iii) Federation of Indian Chamber of Commerce and Industry.

(ii) ***Collection from Un-published Data***

The Central Government, State Government and Research Institutes also collect data which is not published due to some reasons. This type of data is called un-published data. Un-published data can also be made use of in Investigations. The data collected by research scholars of Universities is also generally not published.

NOTES

Precautions in the Use of Secondary Data

The secondary data must be used very carefully. The applicability of the secondary data should be judged keeping in view the object and scope of the Investigation. *Prof. Bowley* has remarked, "Secondary data should not be accepted at their face value." Following are the basis on which the applicability of secondary data is to be judged.

(i) ***Reliability of data:*** Reliability of data is assessed by reliability of the agency which collected that data. The agency should not be biased in any way. The enumerators who collected the data should have been unbiased and well trained. Degree of accuracy achieved should also be judged.

(ii) ***Suitability of data:*** The suitability of the data should be assessed keeping in view the object and scope of Investigation. If the data is not suitable for the investigation, then it is not to be used just for the sake of economy of time and money. The use of unsuitable data can lead to only misleading results.

(iii) ***Adequacy of data*****:** The adequacy of data should also be judged keeping in view the object and scope of the investigation. If the data is found to be inadequate, it should not be used. For example, if the object of investigation is to study the problem of unemployment in India, then the data regarding unemployment in one state say U.P. would not serve the purpose.

2.4 CLASSIFICATION OF DATA

Definition

Classification is defined as the process of arranging data in groups (or classes), according to some common characteristic which separate them into different, but related parts. For example, we may classify the students of a college according to their age, by using the classes 16–18, 18–20, 20–22, 22–24 years etc. Here the students having common characteristics like age between 18 and 20 would be counted in the class 18–20. Here the classes are all different but still these are related in the sense that age is the common base of classification.

Objects of Classification

Collected data is classified in order to achieve the following objectives:

1. Data is classified to condense it into some classes formed according to the magnitude of the data.
2. Data is classified to bring out the points of similarities and dissimilarities in the data. For example, the data of population census can be classified according to the attributes male, female, literate, illiterate, rich, poor etc.

3. Data is classified to facilitate comparison. For example, performance of students of two colleges can be easily compared if they are classified according to classes of percentage of marks like 0–10, 10–20, ..., 90–100.
4. Data is classified so that statistical methods may be applied easily on the data.

Requisites of a Good Classification

1. The raw data must be classified, keeping in view the object of the investigation.
2. The classes must be exhaustive. In other words, there must exist some class for each and every item.
3. The classes must be mutually exclusive. It means that there must exist exactly one class for each item. For example, the classes 10–20 and 15–25 are not mutually exclusive, because the item 18 can be entered in any of the classes.
4. The classes must be homogeneous in the sense that the units of classes must be the same.
5. The classes must be flexible. It means that the classes may be decreased or increased as per the need of the situation.

Variable

The value of each item in the collected data is based on certain characteristics. The characteristics like height, weight, income, expenditure, population, marks etc. are measurable in nature. Such characteristics which are measurable in nature are called **quantitative variables.** A variable which can theoretically assume any value between two given values is called a **continuous variable,** otherwise it is called a **discrete variable.** The characteristics like beauty, honesty, intelligence, colour etc. are non-measurable in nature. Such characteristics which are non-measurable are called **qualitative variables** or **attributes.**

Statistical Series

The statistical data arranged according to some logical order is called a **statistical series.** *Prof. L.R. Connor* has defined statistical series as, "If two variable quantities can be arranged side by side so that measurable differences in the one corresponds with the measurable differences in other, the result is said to form a statistical series".

2.5 TYPES AND OBJECTIVES OF CLASSIFICATION

The collected data can be classified as follows:

1. Geographical Classification
2. Chronological Classification
3. Qualitative Classification
4. Quantitative Classification.

Geographical Classification

In this type of classification, the statistical data is arranged according to geographical or locational differences. This type of classification is also called **spatial classification.** The series obtained by using this type of classification is called a **spatial series.**

Example

States	No. of Units of '*X*' produced in 1986 (in millions)
Haryana	10
Punjab	6
U.P.	15
Maharashtra	12
Other States	20
Total	63

Chronological Classification

In this type of classification, the statistical data collected over a period of time is arranged according to time. This type of classification is also called **temporal classification.** The series obtained by using this type of classification is called a **'time series'.**

Example

Year	Profit of firm '*X*' (in thousands of ₹)
1980	120
1981	100
1982	150
1983	155
1984	170
1985	170
1986	180

Qualitative Classification

In this type of classification, the statistical data is classified according to some qualitative variable (attribute). In this type of classification only the presence or absence of a particular attribute is observed. The series obtained by using this type of classification is called **condition series.** If only one attribute is used in classification, then it is called **simple classification.** If more than one attribute is used then the classification is called **manifold classification.**

Following is an example of qualitative classification on the basis of two attributes:

Workers in Factory 'X'

	Skilled	Unskilled	Total
Male	92	27	119
Female	20	6	26
Total	112	33	145

Quantitative Classification

In this type of classification, the statistical data is classified according to some quantitative variable. The variable may be either discrete or continuous. The series obtained by using this type of classification is called a **frequency distribution.** The collection of values of items according to some quantitative variable is called an **individual series.** For example, the data regarding the heights of students of a class would be an individual series. If the values of the items are repeated, then the data is classified according to the different values of the variable. For example, if there are 5 students having weights 56 kg, then we shall say that the **frequency** of 56 is 5. Let us suppose that the weights (in kg) of students are as follows:

45	54	65	59	62	54	40	46
46	49	64	46	62	46	46	40
62	46	50	71	47	47	56	49
56	47	52	69	49	46	56	56
59	49	62	70	56	49	49	47

This is an individual series, we can also write the above data as follows:

Weight (in kg.)	Tally bars	No. of students (*f*)
45	|	1
46	𝍸 ||	7
62	||||	4
56	𝍸	5
59	||	2
54	||	2
49	𝍸 |	6
47	||||	4
65	|	1
64	|	1
50	|	1
52	|	1
71	|	1
69	|	1
70	|	1
40	||	2
		N = 40

This type of series is called a **frequency distribution.** Here tally bars have been used to count the number of times, the values of the variable has occurred. The fifth, tenth etc. bar is marked oblique, just to facilitate counting. By doing so, we shall get blocks of 5 bars. In this series, we see that 46 has occurred 7 times, we shall say that the frequency of 46 is 7. The frequency is denoted by '*f* '. The variable is generally denoted by '*x*'. In this frequency distribution, we arrange the values of the variable in order of magnitude. Thus, we get the above frequency distribution as follows:

Sr. No.	**Weight (in kg) *x***	**No. of students (*f*)**
1	40	2
2	45	1
3	46	7
4	47	4
5	49	6
6	50	1
7	52	1
8	54	2
9	56	5
10	59	2
11	62	4
12	64	1
13	65	1
14	69	1
15	70	1
16	71	1
		N = 40

2.6 TABULATION

Tabulation is a statistical method in which data is *presented* systematically using rows and columns. It is a fact that unarranged data cannot be analysed effectively. Tabulation helps to bring items with common characteristics to come together. *Tuttle* defined a statistical table exhaustively in the following words, "A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers, with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and notes to make clear the full meaning of the data and their origin."

Objects of Tabulation

The process of tabulation is carried to achieve the following objectives:

(i) Data is tabulated to condense it by using rows and columns.

(ii) Data is tabulated to bring out the points of similarities and dissimilarities in the data.

(iii) Data is tabulated to facilitate comparison.

(iv) Data is tabulated so that statistical methods may be used effectively.

(v) Data is tabulated so that classified data may be arranged systematically in rows and columns.

(vi) Data is tabulated to economise space and time.

General Rules for Tabulation

The following points must be kept in mind while presenting the data in tables:

(i) It should be presented in an attractive form.

(ii) The data may be approximated to a reasonable standard.

(iii) The totals of different columns and rows should also be shown.

(iv) Title must be provided with every table.

(v) Source of data should also be mentioned in the table.

(vi) It should not be made unduly complex by introducing too many rows and columns.

Main Parts of a Table

A good table is expected to have the following parts:

(i) ***Title:*** Every table must possess a suitable title. Title should be self-explanatory in nature. It should also be brief.

(ii) ***Table number:*** Every table should be provided with a table number. It helps in locating it.

(iii) ***Captions:*** The headings of columns are called **captions**. A table may have more than one caption. The captions should be brief and self-explanatory. Captions should also be numbered.

(iv) ***Stubs:*** The headings of rows are called **stubs**. A table may have more than one stub. The stubs should also be brief and self-explanatory. The stubs should also be numbered.

(v) ***Body:*** The part of the table which contains the numerical figures is called the **body** of the table. If the entries in a particular row or column are in the same units, then instead of mentioning units with every entry, the units are mentioned in the corresponding stub or column.

NOTES

(vi) ***Ruling and spacing:*** For the table to look attractive and clear, it is very necessary to make use of ruling between different stubs and columns. The different stubs and columns should also be preferably at equal distances.

(vii) ***Footnotes:*** Any other necessary information regarding the contents of the table which are not covered in the stubs, columns, title etc. may be given in the footnotes.

(viii) ***Source of data:*** The source of data in the table should also be mentioned therein.

Specimen of Table

The following is a specimen of table showing the places for its different parts:

TITLE

Table No.

	Caption	**Caption**	**⟶**	**Caption**	**Total**
Stub					
Stub					
↓			Body		
Stub					
Total					

Footnotes .....

Source .....

Types of Tables

Tables are classified as:

(i) Simple Table

(ii) Complex Table

(i) ***Simple Table:*** When the data in a table is tabulated on the basis of only one characteristic, it is called a **simple table**. Simple table is also called **one-way table**. For example, the table showing the number of students in different faculties in a University is a simple table.

No. of Students in the University in the Year 1986–87

Faculties	No. of students
Arts	...
Commerce	...
Law	...
Non-Medical	...
Medical	...
Total	...

(ii) ***Complex Table:*** When the data in a table is tabulated on the basis of more than one characteristic, it is called a **complex table**. A complex table is also called a **manifold table**. When two characteristics are used in the table, then it is called a **double table** or **two-way table**. In the above example, if we also tabulate the number of boys and girls in different faculties, it would be a two-way table:

When three characteristics are used in a table, then it is also called a **triple table** or **three-way table**. In the above example, if we also show the number of hostellers and day scholars separately, it would become a triple table.

No. of Students in the University in the Year 1986–87

Faculties	No. of students		
	Boys	**Girls**	**Total**
Arts	...	...	...
Commerce	...	...	...
Law	...	...	...
Non-Medical	...	...	...
Medical	...	...	...
Total	...	...	...

No. of students in the University in the year 1986–87

Faculties	No. of students				Total		Total
	Boys		Girls				
	Hostellers	Day scholars	Hostellers	Day scholars	Hostellers	Day scholars	
Arts	...	...	...	...	...	...	...
Commerce	...	...	...	...	...	...	...
Law	...	...	...	...	...	...	...
Non-Medical	...	...	...	...	...	...	...
Medical	...	...	...	...	...	...	...
Total	...	...	...	...	...	...	...

The above table can also be extended to show the fourth characteristic "Age Groups". When four characteristics are used in a table then it may also be called a **four-way table**.

No. of Students in the University in the Year 1986–87

Faculties	Age Group (in years)	No. of students						Total		
		Boys			Girls					
		Hostellers	Day scholars	Total	Hostellers	Day scholars	Total	Hostellers	Day scholars	Total
Arts	Less than 20	—	—	—	—	—	—	—	—	—
	20–25	—	—	—	—	—	—	—	—	—
	25 and above	—	—	—	—	—	—	—	—	—
Commerce	Less than 20	—	—	—	—	—	—	—	—	—
	20–25	—	—	—	—	—	—	—	—	—
	25 and above	—	—	—	—	—	—	—	—	—

(*Contd.*)

Law	Less than 20	—	—	—	—	—	—	—	—	—
	20–25	—	—	—	—	—	—	—	—	—
	25 and above	—	—	—	—	—	—	—	—	—
Non-medical	Less than 20	—	—	—	—	—	—	—	—	—
	20–25	—	—	—	—	—	—	—	—	—
	25 and above	—	—	—	—	—	—	—	—	—
Medical	Less than 20	—	—	—	—	—	—	—	—	—
	20–25	—	—	—	—	—	—	—	—	—
	25 and above	—	—	—	—	—	—	—	—	—
Total		—	—	—	—	—	—	—	—	—

Check Your Progress

State wheather the following statements are True or False:

6. Reserve Bank of India Bulletin is an international publication.
7. The data collected by research scholars of Universities is generally published.
8. The value of each item in the collected data is based on certain characteristics.
9. Geographical classification is also called temporal classification.
10. Chronological classification is also called spatial classification.

Example 1: *Prepare a blank table to show the following characteristics regarding the employees of establishment 'X'*:

(*i*) *Employees in the years 1985, 1986.*

(*ii*) *Male and Female.*

(*iii*) *Grade II, III, IV employees.*

Solution: **Table showing employees of establishment 'X' (1985–86)**

Year	No. of employees								Total			
	Male				Female							
	Grade II	Grade III	Grade IV	Total	Grade II	Grade III	Grade IV	Total	Grade II	Grade III	Grade IV	Total
1985												
1986												
Total												

Example 2: *A college organised a trip of 65 persons, out of which 60 were students of different classes. One servant and one lady lecturer also joined with other male lecturers accompanying the trip. 15 girls students also participated. Present the above data in the form of a table.*

Solution: Total number of participants = 65

Number of students = 60

Therefore, others = 65 – 60 = 5

Number of servants = 1

Number of lady lecturers = 1

Therefore, number of male lecturers = 5 – 1 – 1 = 3

Girls students = 15

Therefore, boys students = 60 – 15 = 45

Participants	Sex		Total
	Males	Females	
Students	45	15	60
Lectures	3	1	4
Servants	0	1	1
Total	48	17	65

2.7 SUMMARY

- Data is called **primary,** if it is originally collected in the process of investigation. Primary data are original in nature.
- In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the data about the informants is collected from some selected persons who are expected to be acquinted with the informants as well as the object of the investigation. A person giving data about the informants is called 'witness.'
- The success of collecting data by using either questionnaire or schedule depends upon the quality of itself.
- If the investigator finds that most of the informants in the sample have left some questions un-answered then those questions should be modified or deleted altogether, provided the object of the investigation permits to do so. This is called *Pilot Survey*.
- Data is called **secondary** if it is not originally collected in the process of investigation.
- International Organisations and Govt. of foreign countries collect and publish statistical data relating to various characteristics.
- In India, the Central Govt. and State Govt. collects data regarding various aspects.
- Such characteristics which are measurable in nature are called **quantitative variables.** A variable which can theoretically assume any value between two given values is called a **continuous variable,** otherwise it is called a **discrete variable.**
- In this type of classification, the statistical data is arranged according to geographical or locational differences. This type of classification is also called **spatial classification.**
- In this type of classification, the statistical data collected over a period of time is arranged according to time. This type of classification is also called **temporal classification.**
- In this type of classification, the statistical data is classified according to some quantitative variable. The variable may be either discrete or continuous.
- **Tabulation** is a statistical method in which data is *presented* systematically using rows and columns.
- The headings of columns are called **captions**.
- The headings of rows are called **stubs**.
- When the data in a table is tabulated on the basis of only one characteristic, it is called a **simple table**.
- When the data in a table is tabulated on the basis of more than one characteristic, it is called a **complex table**.

- When three characteristics are used in a table, then it is also called a **trable table** or **three-way table**.

NOTES

2.8 GLOSSARY

- ***Un-published data:*** The Central Government, State Government and Research Institutes also collect data which is not published due to some reasons. This type of data is called un-published data.
- ***Quantitative variables:*** The characteristics which are measurable in nature are called quantitative variables.
- ***Continuous variable:*** A variable which can theoretically assume any value between two given values is called a continuous variable.
- ***Attributes:*** The characteristics like beauty, honesty, intelligence, colour etc. are non-measurable in nature. Such characteristics which are non-measurable are called attributes.
- ***Temporal classification:*** The statistical data collected over a period of time is arranged according to time. This type of classification is also called temporal classification.
- ***Condition series:*** The series obtained by using qualitative classification is called condition series.
- ***Frequency distribution:*** The series obtained by using quantitative classification is called a frequency distribution.
- ***Individual series:*** The collection of values of items according to some quantitative variable is called an individual series.

2.9 ANSWERS TO CHECK YOUR PROGRESS

1. special purpose investigation
2. witness
3. illiterate, semi-literate
4. questionnaire
5. census data
6. False
7. False
8. True
9. False
10. False

2.10 TERMINAL AND MODEL QUESTIONS

1. Distinguish between primary and secondary data.
2. Explain the various methods used in collection of primary data pointing out their merits and demerits.
3. Discuss the sources, advantages and limitations of secondary data. What precautions should be taken while using secondary data?
4. Describe the points which you would consider in drafting a questionnaire.
5. Why we prefer primary data than secondary data? Explain.
6. What is a statistical table? Mention the rules of the construction of a table.
7. Distinguish between classification and tabulation. Discuss the purpose, methods and importance of classification.
8. Define classification and tabulation and show their importance in statistical studies.
9. Prepare a blank table to show the following characteristics regarding the population in city 'X':
   (i) Male, Female
   (ii) Age groups, less than 30–45, 45 and above
   (iii) Employed and unemployed.
10. Draft a blank table, showing the distribution of students of a college according to classes and sex.
11. Draft a blank table, showing the distribution of students of a college according to classes, sex and age.
12. There are 1000 employees in a factory, out of which 200 are females. There are only 50 class-II female employees and all other female empolyees are class-III. The number of class-II male employees is 200 and the rest are class-III employees. Represent the above data in the tabular form.
13. Submit a proforma of table which could explain the distribution of students of a college on the following basis: Faculty wise, Rural or Urban, Male or Female, hostlers or day scholar.

2.11 REFERNCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

NOTES

UNIT 3: FREQUENCY DISTRIBUTION AND GRAPHICAL REPRESENTATIONS

Structure

3.0 INTRODUCTION

Statistical data recorded in an arbitrary manner after their collection from the field or enquiry are called Raw data.

Definition of Frequency Distribution

A classification showing different values of a variable and their respective frequencies (*i.e.,* number of times each value occurs) side by side is called a frequency distribution.

Croxton and Cowden defined frequency distribution as a statistical table which shows the sets of all distinct values of the variable arranged in order of magnitude either individually or in groups with their corresponding frequencies side by side.

Tally Mark

Tally mark is the upward slant stroke (/) which is put against a value when it occurs once in the raw data. When the value occurs more than four times the fifth occurence is represented by cross tally mark (–) running diagonally across the first four tally marks.

The total of the tally marks against each value is its frequency, such a frequency distribution is known as simple (or ungrouped) frequency distribution.

Let us consider the marks obtained by 60 students of a class in statistics.

15	55	18	25	56	39	26	18	32	15
25	25	22	25	46	46	25	2	36	35
35	68	35	32	38	56	32	22	46	48
10	75	42	36	24	64	39	35	64	42
45	40	56	48	18	78	42	54	47	54
50	20	63	45	35	26	54	58	35	68

Solution:

Arranging the data in ascending order.

2	20	25	32	39	45	48	55	64
10	22	25	32	39	45	48	56	64
15	22	25	32	40	46	50	56	68
15	24	26	35	42	46	54	56	68
18	25	26	35	42	46	54	58	75
18	25		35	42	47	54	63	78

Simple frequency distribution of marks of 60 students.

Marks	Tally marks	Frequency	Marks	Tally marks	Frequency
2	|	1	42	|||	3
10	|	1	45	||	2
15	||	2	46	|||	3
18	|||	3	47	|	1
20	|	1	48	||	2
22	||	2	50	|	1
24	|	1	54	|||	3
25	𝍸	5	55	|	1
26	||	2	56	|||	2
32	|||	3	58	|	1
35	𝍸	6	63	|	1
36	||	2	64	||	2
38	|	1	68	||	2
39	||	2	75	|	1
40	|	1	78	|	1
Total	**33**	**Total frequency 27 + 33 = 60**			

3.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- Define frequency distribution and various terms related with it
- Explain grouped frequency distribution and different ways of representing it
- Define graphical representation of frequency distribution
- Explain histogram and solve problems related to it
- Explain frequency polygon and solve problems related to it
- Explain bar graph and solve problems related to it
- Explain ogive and solve problems related to it

3.2 GROUPED FREQUENCY DISTRIBUTION

A tabular arrangement of raw data by putting the whole range of observations into a number of smaller groups or classes, showing the respective class-frequencies against the class interval is known as a grouped frequency distribution.

Example: The above table can be written in class interval form as follows:

Class intervals	Tally marks	Frequency
1–10	||	2
11–20	𝍸 |	6
21–30	𝍸 𝍸	10
31–40	𝍸 𝍸 𝍸	15
41–50	𝍸 𝍸 ||	12
51–60	𝍸 |||	8
61–70	𝍸	5
71–80	|||	2
Total		**60**

Class Limits

The two end values of a class-interval are called class limits.

The smaller of these two end-values is called the lower class limit and the larger one is the upper class limit.

Mid Value

The mid value of a class-interval is the value exactly at the middle of the class interval and is given by

$$\text{Mid value} = \frac{\text{lower class limit} + \text{upper class limit}}{2}$$

Class Boundaries

In most cases of measurement of continuous variables all data are recorded nearest to some unit. From the following grouped frequency table, the limits 119.5 and 129.5 are called the class boundaries of the class interval 120–129. The lower limit 119.5 being the lower class boundary and upper limit 129.5 being the upper class boundary.

If d is the gap between any two consecutive classes, then d is the common difference between upper class limit of any class and lower class limit of next higher class.

Lower class – boundary = Lower class limit – 0.5

Upper class – boundary = Upper class limit + 0.5

A grouped frequency table (inclusive type) class limit and class-boundaries is given below. Under the inclusive type of tabulation, the upper limit of a class is included in that class.

Class interval	Frequency	Class limits		Class boundaries	
Height	No. of students	Lower	Upper	Lower	Upper
100–109	7	100	109	99.5	109.5
110–119	15	110	119	109.5	119.5
120–129	30	120	129	119.5	129.5
130–139	20	130	139	129.5	139.5
140–149	8	140	149	139.5	149.5

In a grouped frequency distribution of exclusive type, the upper limit of one class is the lower limit of next class and in this case class limit of a class coincide with the class boundaries of that class.

Class Interval	Frequency	Class limits		Class boundaries	
Age in years	No. of persons	Lower	Upper	Lower	Upper
20–25	8	20	25	20	25
25–30	20	25	30	25	30
30–35	40	30	35	30	35
35–40	23	35	40	35	40

Width of a Class Interval

Width of a class = Upper class boundary – Lower class boundaries

If the class interval is not given then the width is difference between two successive class marks.

Percentage Frequency

Percentage frequency of a C.I. is the frequency of the class interval expressed as a percentage of the total frequency of the distribution.

$$\text{P.F. of a class} = \frac{\text{Frequency of the class}}{\text{Total frequency}} \times 100$$

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

Frequency Density

Frequency density of a C.I. is defined by

$$\text{Frequency density of a class} = \frac{\text{Frequency of the class}}{\text{Width of the class}}$$

If the classes of a frequency distribution are not of equal width, then Frequency density of such classes are used for comparing the concentration of frequencies in the classes.

Different Ways of Presenting Frequency Distribution

Type 1: Height of 100 students of a class.

Height in cms	No. of students
120 and above but below 125	10
125 and above but below 130	18
130 and above but below 135	25
135 and above but below 140	22
140 and above but below 145	16
145 and above but below 150	9

Here the upper limit of the class does not belong to the class, but it is included in next class. Sometimes, the words but less than or 'and under' are used in place of 'but below'.

Type 2: Frequency distribution of output of 120 workers

Output in units	No. of workers
300–309	9
310–319	28
320–329	42
330–339	30
340–349	11

Type 3: Weekly wages of 200 workers

Weekly wages in (₹)	30	50	70	90	110	130	150	170	190
No. of workers	12	23	35	48	36	21	15	10	

Here the classes are of equal width and we read them as 30 and above but below 50

Type 4: Yearly income of 600 employees of an industry

Yearly income (₹ 000)	**No. of employees**
Above 4.3 but not exceeding 5.2	15
Above 5.2 but not exceeding 6.1	80
Above 6.1 but not exceeding 7.0	120
Above 7.0 but not exceeding 7.9	180
Above 7.9 but not exceeding 8.8	110
Above 8.8 but not exceeding 9.7	70
Above 9.7 but not exceeding 10.6	25

The lower limit does not belong to this class but the upper limit belong to the class.

Type 5: Turnover of 250 firms in a State

Turnover	50–100	100–150	150–250	250–400	400–600	600–800
No. of workers	15	50	100	55	22	8

The classes are of unequal widths. Unequal C.I. are preferred only when there is greater fluctuation in the data *i.e.,* when there is sharp rise or fall in the frequency over a small interval.

Type 6: Age of 500 employees of a firm:

Age	**No. of employees**
Below 20	15
20 – 25	34
25 – 30	58
30 – 35	80
35 – 40	120
40 and above	193

Cumulative Frequency Distribution

In classification of statistical data it is sometimes necessary to find the number of observations less than or more than a given value. This is done by accumulating the frequencies upto or above the given value. This accumulated frequency is called cumulative frequency for that value.

Check Your Progress

Fill in the blanks:

1. The two end-values of a class interval are called .....
2. The ..... of a class-interval is the value exactly at the middle of the class interval.
3. When the value of frequency more than four times, the fifth occurance is represented by .....
4. The accumulated frequency running diagonally across the first is called ..... upto or above the given value for that value.

3.3 GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS

We know that a frequency distribution is obtained by classifying the data according to the values of some quantitative variable. The number of times a different value of the variable is repeated, is called its frequency. Now we shall consider the graphical representation of frequency distributions.

Types of Graphs of Frequency Distributions

(i) Histogram (ii) Frequency Polygon

(iii) Bar Chart (iv) Ogives.

We discuss each type of the above mentioned graphs of frequency distribution, in detail.

3.4 HISTOGRAM

A **histogram** is a set of rectangles with class intervals as basis and corresponding frequencies as heights, provided the classes in the distribution are of equal widths and are in ascending order of magnitude. In case, the classes are not of equal width, then before constructing histogram, the width of classes are made equal. If the width of a particular class is double than those of other classes, then it is divided into two classes with frequencies also equally divided. We know that the frequency of a class is always a whole number. If the frequency of the class which is to be broken into two classes is odd, say 15, then the frequencies of the classes are to be as 8, 7 or 7, 8. If the classes are given in inclusive form, then the actual class limits are taken into account while drawing the histogram. The frequencies are to be measured along Y-axis. On the Y-axis, scale must begin from '0'. On the 'X' axis, the classes are marked and it is not necessary to start the scale from '0'.

Example 1: *Construct a histogram from the following data:*

Marks	*10—19*	*20—29*	*30—39*	*40—49*	*50—59*
No. of students	*15*	*20*	*35*	*10*	*4*

Solution: The actual class limits of the given classes are 9.5–19.5, 19.5–29.5,.....

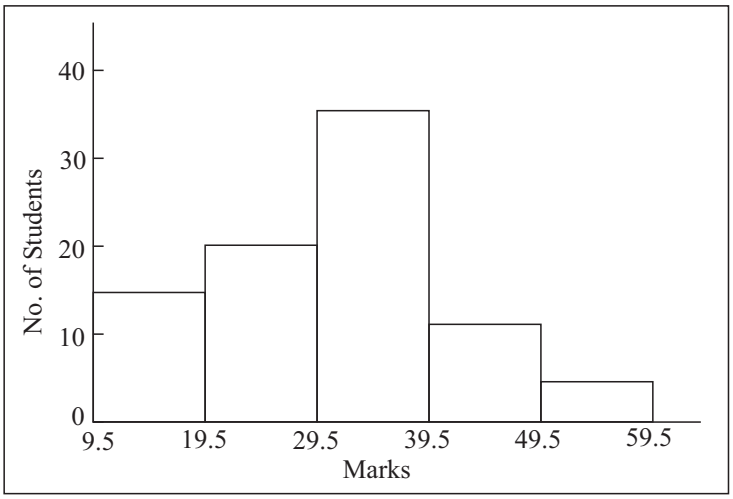


Fig. 3.1: Histogram

Example 2: *Represent the following data by means of a histogram:*

Weekly wages (in ₹)	*10—15*	*15—20*	*20—25*	*25—30*	*30—40*	*40—50*	*60—80*
No. of workers	*7*	*19*	*27*	*15*	*12*	*12*	*8*

Solution: In this distribution, the classes are not of equal width. The smallest width is 5. We shall modify the classes, whose width is greater than 5. The frequency of the class 30—40 is 12. This class would be broken into classes 30—35 and 35—40, each with frequency '6'. Similarly, the class 40—50 would be broken in 2 classes 40—45 and 45—50 with frequencies 6 each. The class 60—80 would be broken in to four classes, 60—65, 65—70, 70—75, 75—80, each with frequency 2. We rewrite the classes as:

Classes	**No. of workers**	**Classes**	**No. of workers**
10—15	7	45—50	6
15—20	19	50—55	0
20—25	27	55—60	0
25—30	15	60—65	2
30—35	6	65—70	2
35—40	6	70—75	2
40—45	6	75—80	2

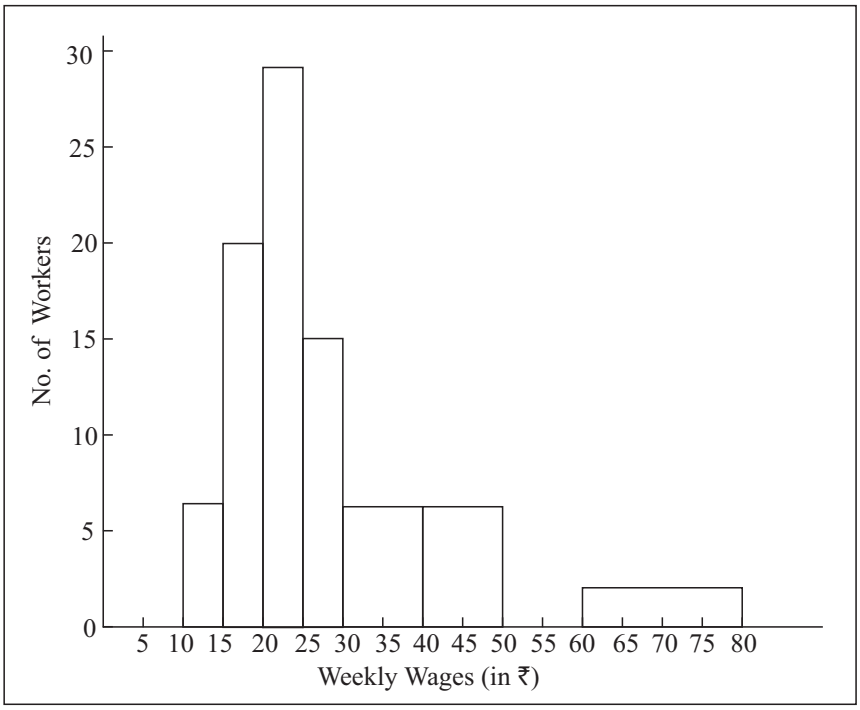


Fig. 3.2: Histogram

The vertical lines through the class limits which were not given in the original data, are not drawn.

3.5 FREQUENCY POLYGON

The **frequency polygon** of a frequency distribution is obtained by joining the mid-points of upper sides of rectangles with that for the adjacent rectangles. The polygon is closed by taking two classes on either sides with '0' frequency each. The frequency polygon can also be drawn without the help of histogram. The classes and frequencies are to be measured along X-axis and Y-axis as we did in the case of histogram. The points are taken with abscissas as mid-points of classes and ordinates as their respective frequencies. Two additional points are also taken on the X-axis corresponding to two classes on each end with '0' frequencies. It may be noted that the area between the histogram and the frequency polygon would always be exactly equal.

Example 3: *Draw a frequency polygon for the data given below:*

Income less than (in ₹)	*500*	*600*	*700*	*800*	*900*	*1000*
No. of employees	*0*	*25*	*55*	*105*	*135*	*150*

Solution: We first convert the data in the form of a frequency distribution.

Income (in ₹)	**500—600**	**600—700**	**700—800**	**800—900**	**900—1000**
No. of employees	25	30	50	30	15

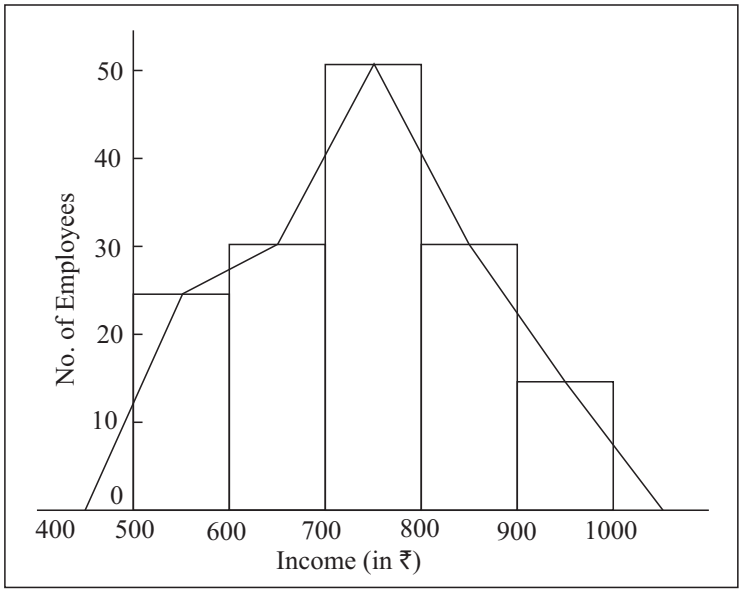


Fig. 3.3: Frequency Polygon

3.6 BAR CHART

A bar chart is a graphical representation of the frequency distribution in which the bars are centered at the mid-points of the cells. The heights of the bars are proportional to the respective class frequencies.

If a single attribute is presented then it is called simple bar chart. When more than one attribute is presented then it is called multiple bar chart.

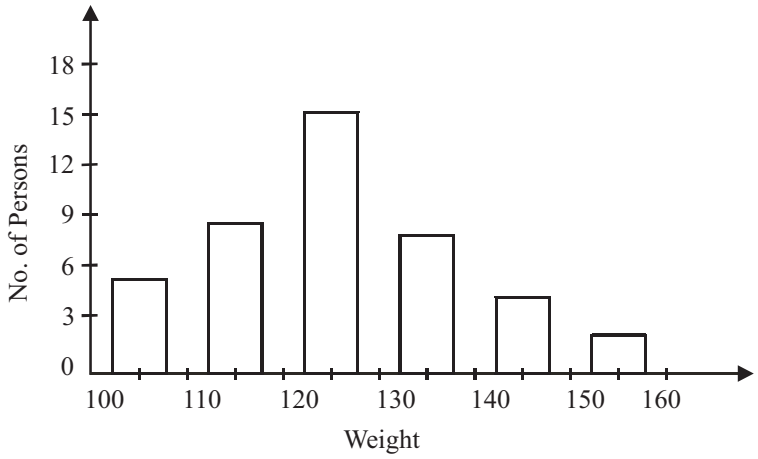


Fig. 3.4

NOTES

3.7 OGIVES

Ogives are also called cumulative frequency curves. Ogives are of two types: (*i*) 'Less than ogive' (*ii*) 'More than Ogive'. The '**less than ogive**' of a frequency distribution is obtained by joining the points with abscissas as upper limit of classes and ordinate as corresponding number of items less than that upper limit. The point (lower limit of first class, 0) is also taken. Let us take an example.

Income (in ₹)	100—200	200—300	300—400	400—500	500—600
No. of employees	15	20	40	20	5

We rewrite this data as follows:

Income less than	100	200	300	400	500	600
No. of employees	0	15	35	75	95	100

The '*less than ogive*' would be obtained by joining the points (100, 0), (200, 15), (300, 35), (400, 75), (500, 95), (600, 100) on the graph paper.

WORKING RULES FOR DRAWING OGIVES

Step I. *Arrange the classes in ascending order of magnitude. The classes must be in 'exclusive form'. The widths of classes may not be equal.*

Step II. *On the vertical axis, cumulative frequencies would be marked.*

Step III. (*i*) *For the 'less than' ogive, take the points (x, y) where :*

x = upper limit of a class

y = number of items less than x

The point (lower limit of first-class, 0) is also taken.

(*ii*) *For the 'more than' ogive, take the points (x, y) where :*

x = lower limit of a class

y = number of items more than x.

The point (upper limit of last class, 0) is also taken.

Step IV. *The points obtained in **step III** are joined by straight lines. The 'less than' ogive is left to right rising while the 'more than' ogive is left to right falling.*

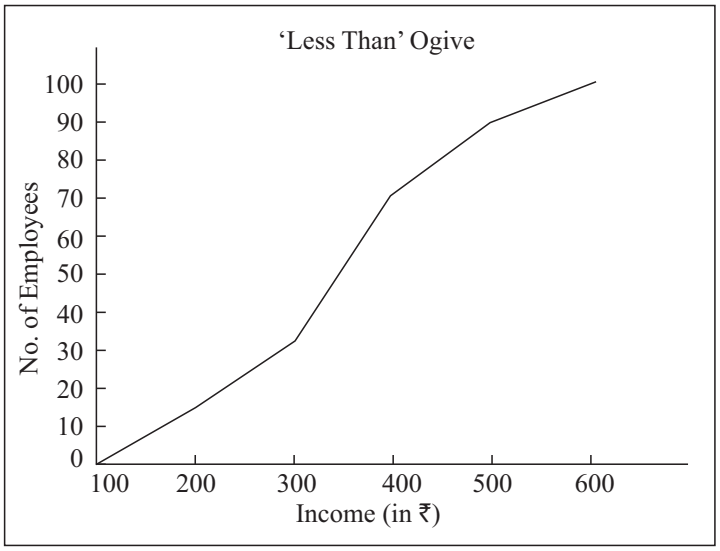


Fig. 3.5

The '**more than Ogive**' of a frequency distribution is obtained by joining the points with abscissas as lower limits of classes and ordinate as corresponding no. of items more than that lower limit. The point (upper limit of last class, 0) is also taken. The above-mentioned frequency distribution can also be expressed as :

Income more than	100	200	300	400	500	600
No. of employees	100	85	65	25	5	0

The *more than* Ogive would be obtained by joining the points (100, 100), (200, 85), (300, 65) etc. on the graph paper.

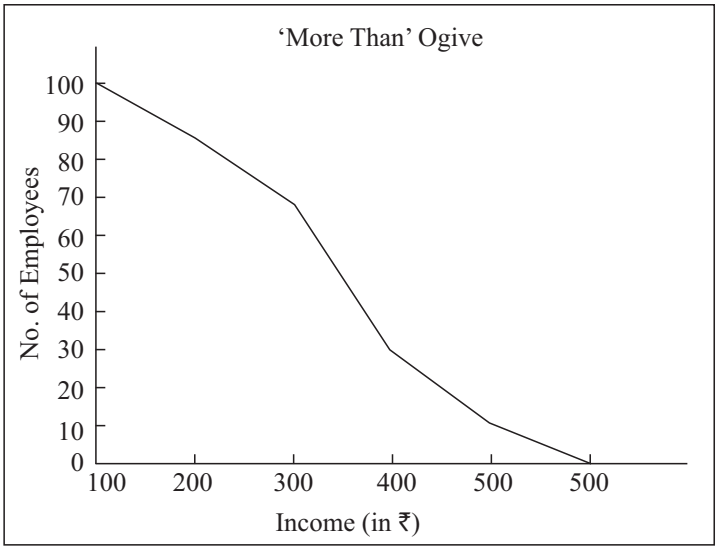


Fig. 3.6

Both '*less than*' and '*more than*' Ogives can also be drawn on the same graph paper.

Check Your Progress

State whether the following statements are True or False:

5. The frequency polygon can also be drawn without the help of histogram.
6. In a bar chart, bars are centered at end-points of the cells.
7. Area between histogram and frequency polygon is exactly equal.
8. Both 'less than' and 'more than' Ogives are drawn on separate graph paper.
9. In multiple bar charts, single attribute is presented.

3.8 SUMMARY

- A classification showing different values of a variable and their respective frequencies (*i.e.,* number of times eaech value occurs) side by side is called a frequency distribution.
- Tally mark is the upward slant stroke (/) which is put against a value when it occurs once in the raw data.
- A tabular arrangement of raw data by putting the whole range of observations into a number of smaller groups or classes, showing the respective class-frequencies against the class interval is known as a grouped frequency distribution.
- The two end values of a class-interval are called class limits.
- The mid value of a class-interval is the value exactly at the middle of the class interval and is given by

$$\text{Mid value} = \frac{\text{lower class limit} + \text{upper class limit}}{2}$$

- Percentage frequency of a C.I. is the frequency of the class interval expressed as a percentage of the total frequency of the distribution.
- A **histogram** is a set of rectangles with class intervals as basis and corresponding frequencies as heights, provided the classes in the distribution are of equal widths and are in ascending order of magnitude.
- The **frequency polygon** of a frequency distribution is obtained by joining the mid-points of upper sides of rectangles with that for the adjacent rectangles.
- A bar chart is a graphical representation of the frequency distribution in which the bars are centered at the mid-points of the cells.
- Ogives are called cumulative frequency curves. Ogives are of two types: (*i*) 'Less than Ogives' (*ii*) 'More than Ogives'.

3.9 GLOSSARY

- ***Frequency distribution:*** A classification showing different values of a variable and their respective frequencies (*i.e.,*number of times each value occurs) side by side is called a frequency distribution.
- ***Frequency:*** The number of times a different value of the variable is repeated, is called its frequency.
- ***Simple bar chart:*** If a single attribute is presented then it is called simple bar chart.

3.10 ANSWERS TO CHECK YOUR PROGRESS

1. class limits
2. mid-value
3. crosstally-mark
4. cummulative frequency
5. True
6. False
7. True
8. False
9. False

3.11 TERMINAL AND MODEL QUESTIONS

1. Following is the data regarding the number of children in 40 families. Represent it in the form of a frequency distribution.

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

2. Monthly wages of 30 labourers are as follows:

   Represent the data in the form of a frequency distribution with class interval equal to ₹ 10/.

510	590	535	530	535	536
512	585	540	535	536	578
508	504	555	545	533	532
540	506	590	590	540	568
560	520	590	598	545	569

3. Represent the following data in the form of a frequency distribution:

Less than	0	10	15	25	40	55	70
No. of items	0	10	15	25	40	55	70

4. Convert the following frequency distribution into an individual series:

X	2	3	4	5
f	1	2	3	1

5. The following table gives the wages and number of workers receiving the wages in a company. Draw a histogram to this data:

Daily wages (in ₹)	10—15	15—20	20—25	25—30	30—35
No. of workers	40	70	60	80	60

6. Draw a histogram to the following data regarding the percentage of marks obtained by 800 students in a college:

Marks	11—20	21—30	31—40	41—50	51—60
No. of students	151	330	161	108	50

7. Draw a frequency polygon to the data given below:

Height (in inches)	60—62	63—65	66—68	69—71	72—74
No. of students	15	21	35	19	10

8. Draw a histogram and frequency polygon to the data given below:

Profit (,000 ₹)	500—550	550—600	600—650	650—750	750—900
No. of companies	20	25	20	20	15

9. Draw both Ogives to the following data:

Weight (in kg)	30—34	35—39	40—44	45—49	50—54	55—59	60—64
Frequencies	3	5	12	18	14	6	2

10. Draw 'less than' as well as 'more than' Ogives to the following data:

Classes	15—19	20—24	25—29	30—34	35—39	40—44	45—49
Frequencies	150	169	300	78	49	42	12

11. From the following data, construct: (1) Bar chart, (2) Histogram.

Wage groups (in ₹)	0—10	10—20	20—30	30—40	40—50
No. of workers	2	4	11	15	25
Wage groups (in ₹)	50—60	60—70	70—80	80—90	
No. of workers	18	15	4	2	

12. Draw a frequency polygon for the following data:

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

13. Draw a bar chart of the following data:

Marks	10—20	20—30	30—40	40—50	50—60
No. of students	15	30	15	10	20

14.

No. of children (X)	0	1	2	3	4	5	6
No. of families (f)	4	7	12	5	6	3	3

15.

Monthly wages (in ₹)	**No. of Labourers**	**Monthly wages (in ₹)**	**No. of Labourers**
500—510	3	550—560	1
510—520	2	560—570	3
520—530	1	570—580	1
530—540	8	580—590	1
540—550	5	590—600	5

3.12 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 4: MEASURES OF CENTRAL TENDENCY

NOTES

Structure

4.0 INTRODUCTION

In the case of many biological characteristics, the values of the extent of the observations are not equal, but we notice a general tendency of such observations to cluster around a particular level. In this situation it may be preferable to characterise each group of observations by such a level, which is called the central tendency of that group. This single value for each group of observations serves as a representative of that group. This level around which the observations tend to cluster may vary from group to group. One of the most important objectives of this biostatistical analysis is to get a single value that describes the characteristic of the entire mass of data. Such a value is called an "average". For example, the average incubation period of one infectious illness may be 7 days and of another 11 days. Though individual values may overlap, the two distributions have different central positions and therefore differ in the characteristic of location. In practice, it is constantly necessary to discuss and compare such measures. A simple instance would be the observation that persons following one occupation lose, on the average, 5 days a year per person from illness while in another occupation they lose 10 days. The two distributions differ in their

position and we are led to seek the reasons for such a difference and to see whether it is remediable.

For quantitative data it is observed that there is a tendency of the data to be distributed about a central value which is a typical value and is called a measure of central tendency. It is also called a measure of location because it gives the position of the distribution on the axis of the variable.

There are three commonly used measures of central tendency, *viz*., Mean, Median and Mode. The mean again may be of three types, *viz*. Arithmetic Mean (A.M.), Geometric Mean (G.M.) and Harmonic Mean (H.M.).

Requirements of a Good Measure of Central Tendency

Some desirable requirements of a good measure of central tendency are as follows:

(i) It should be rigidly defined.

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

(iii) It should be easily comprehensible and easy to calculate.

(iv) It should be capable of further mathematical treatment.

(v) It should not be affected by fluctuations of sampling.

(vi) It should not be unduly affected by extreme observations.

(vii) It should have sampling stability.

(viii) It should be easy to understand.

4.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define central tendency and requirements of a good measure of central tendency
- Define and explain arithmetic mean
- Define and explain geometric mean
- Define and explain harmonic mean
- Define median and its calculation
- Define mode and its calculation
- Explain relationship between mean, median and mode
- Define quartile, decile and percentile

NOTES

4.2 ARITHMETIC MEAN (A.M.)

The arithmetic mean is simply called 'Average'. For the observations $x_1, x_2, ..., x_n$ the A.M. is defined as

$$\bar{x} = \text{A.M.} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For simple frequency distribution,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\text{N}}, \quad \text{where N} = \Sigma f_i$$

For the grouped data (frequency distribution), the arithmetic mean is given by

$$\bar{x} = \frac{1}{\text{N}} \Sigma f x,$$

where f is the frequency, x the mid-point of the class-interval and N the total number of observations.

Properties of Arithmetic Mean

The arithmetic mean has the following properties:

(i) "Algebraic sum of the deviations of a set of values from their arithmetic mean is zero." If x_i / f_i, $i = 1, 2, ..., n$ is the frequency distribution, then

$$\boxed{\Sigma f(x - \bar{x}) = 0}$$

(ii) The sum of the squares of the deviations of a set of values is minimum when taken about mean.

$$\boxed{\Sigma f(x - \bar{x})^2 \rightarrow \text{minimum}}$$

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

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + ... + n_k\bar{x}_k}{n_1 + n_2 + ... + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

(iv) If each value of the variable x is increased or decreased by a constant value, the arithmetic mean so obtained also increases or decreases by the same constant value.

(v) If the values of the variable are multiplied or divided by a constant value, the arithmetic mean so obtained is same as the initial arithmetic mean is multiplied or divided by the constant value.

NOTES

Merits and Demerits of Arithmetic Mean

The following are the merits and demerits of arithmetic mean:

(a) ***Merits of Arithmetic Mean***

(i) It is commonly understood and most widely used.

(ii) It is simple and easy to calculate.

(iii) It is based on all the observations.

(iv) It is a good measure for comparison.

(v) It is adaptable to arithmetic and algebraic treatment.

(b) ***Demerits of Arithmetic Mean***

(i) The value of mean is highly affected by abnormal and extreme values.

(ii) It may not be actually present in the series. For example, the average of 2, 3 and 10 is 5, which is not an observation of the series.

(iii) It can be calculated if certain item is missing. Further, in case of open-end interval, it is calculated on certain assumption.

(iv) It cannot be located by mere observation.

Calculation of Arithmetic Mean

Mainly three forms of data are available, which are given below:

(i) Individual series or ungrouped data

(ii) Discrete series

(iii) Continuous series

(i) **Calculation of Arithmetic Mean in Individual Series**

In individual series, arithmetic mean may be computed by applying (a) Direct method (b) Short-cut method (or deviation method).

(a) ***Direct method:*** The arithmetic mean of a set of n observations $x_1, x_2, ..., x_n$ is denoted by $\bar{x}$ and is defined as

$$\bar{x} = \frac{1}{n} \sum x$$

(*b*) ***Short-cut method:*** If the observations and magnitude of the observations is large, short-cut method is used to reduce the arithmetic calculations. The formula is

$$\bar{x} = A + \frac{1}{n} \sum d,$$

where A $\rightarrow$ assumed mean

$d \rightarrow$ deviations of the observations from assumed mean $(x - A)$

and $n \rightarrow$ total number of observations.

NOTES

Example 1: *Calculate the arithmetic mean for a series of Serum Albumin Levels (g%) of 15 Pre-school children:*

2.90	*3.75*	*3.66*
3.57	*3.30*	*3.76*
3.72	*3.62*	*3.69*
2.98	*3.76*	*3.43*
3.61	*3.38*	*3.76*

Solution: The total of all these values, *i.e.,* $\sum x = 52.89$

Total number of observations $(n) = 15$

Therefore, the arithmetic mean, $\bar{x} = \frac{1}{n}\sum x = \frac{52.89}{15}$

$= 3.53$ g%

Example 2: *Calculate the arithmetic mean from the number of the spikelets per spike in wheat:*

Number of spikelets per spike: 18, 20, 21, 19, 28, 22, 29, 30, 31, 35.

Solution: Assumed mean = 24

	Spikelets per spike '*x*'	**Deviations from the assumed mean *d* = *x* – *A***
	18	18 – 24 = – 6
	20	20 – 24 = – 4
	21	21 – 24 = – 3
	19	19 – 24 = – 5
	28	28 – 24 = 4
	22	22 – 24 = – 2
	29	29 – 24 = 5
	30	30 – 24 = 6
	31	31 – 24 = 7
	35	35 – 24 = 11
Total	$\Sigma x = 253$	$\Sigma d = 13$

(a) ***Direct method:*** Here $\Sigma x = 253$, $n = 10$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{253}{10} = 25.3$$

(b) ***Short-cut method:*** Here $\Sigma d = 13$

$$\bar{x} = A + \frac{1}{n}\Sigma d$$

$$= 24 + \frac{13}{10} = 24 + 1.3 = 25.3$$

(ii) ***Calculation of Arithmetic Mean in a Discrete Series***

In discrete series, arithmetic mean may be computed by applying (a) Direct method (b) Short-cut method (or deviation method).

(a) ***Direct method:*** Let a variable take n values x_1, x_2, ..., x_n having corresponding frequencies f_1, f_2, ..., f_n, then the arithmetic mean is obtained by the formula:

$$\bar{x} = \frac{1}{N}\Sigma fx,$$

where $N = \Sigma f$.

(b) ***Short-cut method (or Deviation method):*** According to this method,

$$\bar{x} = A + \frac{1}{N}\Sigma fd$$

where A is assumed mean

$d = (x - A)$ is the deviation of the observations from assumed mean,

and $N = \sum f$ is the total number of observations.

Example 3: *Calculate the arithmetic mean of Haemoglobin values (g%) of 26 normal children:*

Haemoglobin value (g%)	**No. of children**
10.4	*1*
11.2	*3*
11.8	*4*
12.9	*7*
13.5	*5*
13.8	*4*
14.2	*2*
Total	*26*

Solution:

Haemoglobin value 'x' (g%)	**No. of children 'f'**	***fx***
10.4	1	10.4
11.2	3	33.6
11.8	4	47.2
12.9	7	90.3
13.5	5	67.5
13.8	4	55.2
14.2	2	28.4
Total	$N = 26$	$\Sigma fx = 332.6$

$$\bar{x} = \frac{1}{N}\Sigma fx = \frac{332.6}{26} = 12.79 \text{ (g\%)}$$

(iii) ***Calculation of Arithmetic Mean in Continuous Series***

In continuous series, arithmetic mean may be computed by applying any of the following methods:

(a) Direct method (b) Short-cut method (or deviation method) and (c) Step-deviation method.

(a) **Direct method.** The following formula can be used for calculation of arithmetic mean in a continuous series:

$$\bar{x} = \frac{1}{N}\Sigma fx,$$

where x is the mid-point of various classes.

$$\text{Mid-point} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

Example 4: *Calculate the arithmetic mean of protein intake of 400 families:*

Protein intake/ consumption unit/day (g)	*15–25*	*25–35*	*35–45*	*45–55*	*55–65*	*65–75*	*75–85*
No. of families	*30*	*40*	*100*	*110*	*80*	*30*	*10*

Solution:

Protein intake/ consumption unit/day (*g*) 'class interval'	**No. of families '*f*'**	**Mid-point of the class-interval '*x*'**	**Multiply *f* and *x* '*fx*'**
15–25	30	20	600
25–35	40	30	1200
35–45	100	40	4000
45–55	110	50	5500
55–65	80	60	4800
65–75	30	70	2100
75–85	10	80	800
Total	$N = 400$		$\Sigma fx = 19000$

By direct method, $\bar{x} = \frac{1}{N}\Sigma fx = \frac{19000}{400} = 47.50$ g.

(b) ***Short-cut method:*** According to this method,

$$\bar{x} = A + \frac{1}{N}\Sigma fd \ ;$$

where A = assumed mean

$d = x - A$, x is the mid-point,

$N = \Sigma f$

(*c*) **Step-deviation method:** According to this method

$$\bar{x} = A + \frac{h}{N}\Sigma fd,$$

where A = assumed mean

$$d = \frac{x - A}{h}$$

and h = magnitude of the class-interval.

Arithmetic Mean in Case of Unequal Interval

If unequal intervals are given in a continuous frequency distribution, then it is not necessary to convert them in equal class-interval. Mid-points are calculated for every class- interval and after that question may be solved either by direct or by short-cut method. Step-deviation method is not used because class-interval is not uniform.

Example 5: *Find the arithmetic mean from the following data:*

Classes	*10–20*	*20–40*	*40–70*	*70–120*	*120–200*
Frequency	*4*	*10*	*26*	*8*	*2*

Solution:

Class	**Mid-point 'x'**	**'f'**	**Direct method**	**Short-cut method**	
			fx	***d = x – 55***	***fd***
10–20	15	4	60	– 40	– 160
20–40	30	10	300	– 25	– 250
40–70	55 = A	26	1430	0	0
70–120	95	8	760	40	320
120–200	160	2	320	105	210
		$N = 50$	$\Sigma fx = 2870$		$\Sigma fd = 120$

Direct Method: $\bar{x} = \frac{1}{N}\Sigma fx = \frac{2870}{50} = 57.4$

$\bar{x} = 57.4$

Short-cut Method: $\bar{x} = A + \frac{1}{N}\Sigma fd$

$= 55 + \frac{1}{50} \times 120 = 55 + 2.4 = 57.4$

$\bar{x} = 57.4.$

Arithmetic Mean in Case of Inclusive Series

If the data are given in the form of an inclusive series, there is no need to change it into an exclusive series, because mid-points are unaffected in this case. Moreover, in the calculation of mean it is also not necessary to rearrange a series in ascending or descending order.

Example 6: *Find the mean for the following distribution:*

Classes	*0–9*	*10–19*	*20–29*	*30–39*	*40–49*	*50–59*	*60–69*	*70–79*
Frequency	*21*	*74*	*100*	*120*	*110*	*84*	*30*	*11*

Solution:

Class	**Mid-point** 'x'	'f'	$d = \frac{x - 44.5}{10}$	fd
0–9	4.5	21	– 4	– 84
10–19	14.5	74	– 3	– 222
20–29	24.5	100	– 2	– 200
30–39	34.5	120	– 1	– 120
40–49	44.5 = A	110	0	0
50–59	54.5	84	1	84
60–69	64.5	30	2	60
70–79	74.5	11	3	33
		$N = 550$		$\Sigma fd = -449$

Step-deviation method,

$$\bar{x} = A + \frac{h}{N}\Sigma fd,$$

here $A = 44.5, h = 10, \Sigma fd = -449, N = 550$

$$\bar{x} = 44.5 + \frac{10}{550}(-449) = 44.5 - 8.16 = 36.34$$

$$\bar{x} = 36.34.$$

Arithmetic Mean in Case of Open-End Frequency Distribution

It may be possible that the first and last end of the distribution are open and class-interval is same for other classes. In such situation the interval of first and last class will also be assumed equal to same magnitude.

Example 7: *Find the arithmetic mean for the following frequency distribution:*

Marks	*below 10*	*10–20*	*20–30*	*30–40*	*40–50*	*50–60*	*60–70*	*70 and above*
No. of persons	*5*	*10*	*20*	*40*	*30*	*20*	*10*	*4*

Solution: Here the lower limit of first class and upper limit of last class are not given but the size of the other classes is 10. Thus the class-interval for first and last class will be taken to 10 to maintain uniformity.

Marks	No. of persons '*f*'	Mid-point '*x*'	$d = \frac{x - A}{10}$, where $A = 35$	*fd*
0–10	5	5	– 3	– 15
10–20	10	15	–2	– 20
20–30	20	25	– 1	– 20
30–40	40	35 = A	0	0
40–50	30	45	1	30
50–60	20	55	2	40
60–70	10	65	3	30
70–80	4	75	4	16
	$N = 139$			$\Sigma fd = 61$

Step-deviation method, $\bar{x} = A + \frac{h}{N} \Sigma fd$

$$\bar{x} = 35 + \frac{10}{139} \times 61 = 35 + 4.39 = 39.39$$

$$\bar{x} = 39.39.$$

Correcting Incorrect Mean (Misread Items)

It sometimes happens that due to an oversight or mistake in copying, certain wrong items are taken while calculating mean. The problem is how to find out the correct mean. The process is very simple. From incorrect Σx deduct wrong items and add correct items and then divide the correct Σx by the number of observations. The result so obtained will give the value of correct mean.

Example 8: *The mean marks of 100 students were found to be 40. Later on it was discovered that a score of 53 was misread as 83. Find the correct mean corresponding to the correct score:*

Solution: We are given that $n = 100$, $\bar{x} = 40$

We know that $\bar{x} = \frac{1}{n} \Sigma x$

or $\Sigma x = n \times \bar{x} = 40 \times 100 = 4000$

But this is not correct Σx

Correct Σx = Incorrect Σx – wrong items + correct items

= 4000 – 83 + 53 = 3970

Correct $\bar{x} = \frac{1}{n}(\text{correct } \Sigma x) = \frac{1}{100} \times 3970 = 39.70$

Hence the correct mean = 39.70.

Weighted Arithmetic Mean

One of the limitation of the arithmetic mean is that it gives equal importance to all the items. But there are cases where the relative importance of the different items is not the same. When this is so, we compute weighted arithmetic mean. The term 'weight' stands for the relative importance of the different items. The formula for computing weighted arithmetic mean is

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

where $\bar{x}_w$ represents the weighted mean, w_i represents the weights attached to the items x_i ; $i = 1, 2, ..., n$.

Example 9: *A contractor employs three types of workers – male, female and children. To a male worker he pays ₹ 20 per day, to a female worker ₹ 16 per day and to a child worker ₹ 6 per day. What is the average wage per day paid by the contractor.*

Solution: The average wage would be the weighted mean calculated as follows :

Wages per day (₹) '*x*'	**No. of workers '*w*'**	***wx***
20	20	400
16	15	240
6	5	30
$\Sigma x = 42$	$\Sigma w = 40$	$\Sigma wx = 670$

$$\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{670}{40} = 16.75.$$

However, the number of male, female and child worker employed is generally different. In the absence of this, we take assume weights. Let us assume that the number of male, female and child workers employed is 20, 15 and 5 respectively.

By simple arithmetic mean,

$$\bar{x} = \frac{20 \times 10 + 16 \times 10 + 6 \times 10}{10 + 10 + 10} = ₹\ 14$$

or

$$\bar{x} = \frac{20 + 16 + 6}{3} = ₹\ 14 \text{ per day.}$$

Uses of Arithmetic Mean

(i) It is used in practical statistics.

(ii) Estimates are always obtained by mean.

(iii) Common people uses mean for calculating average marks obtained by students.

4.3 GEOMETRIC MEAN (G.M.)

The geometric mean of the observations $x_1, x_2, ...,x_n$ is defined as

$$\text{G.M.} = (x_1 . x_2 ..... x_n)^{1/n}$$

For simple frequency distribution,

$$\text{G.M.} = \left(x_1^{f_1} . x_2^{f_2} ..... x_n^{f_n}\right)^{1/\text{N}}, \quad \text{N} = \sum_{i=1}^{n} f_i$$

For grouped frequency distribution, x_i is taken as class mark.

Note: 1. The logarithm of the G.M. of a variate is the A.M. of its logarithm.

2. G.M. = 0 iff a single variate value is zero.

3. G.M. is not used if any variate value is negative.

Example 10: *Find the G.M. of the following distribution:*

Humidity reading	No. of days
60	3
62	2
64	4
68	2
70	4

Solution: Here N = No. of days = 15.

x	f	log x	f log x
60	3	1.77815	5.33445
62	2	1.79239	3.58478
64	4	1.80618	7.22472
68	2	1.83251	3.66502
70	4	1.84510	7.38040
Σ	15	–	27.18937

Then, $\log \text{G.M.} = \frac{1}{\text{N}} \Sigma f_i \log x_i = \frac{1}{15}(27.18937) = 1.81262$

$\Rightarrow$ $\text{G.M.} = 64.956\,1 \approx 64.96$.

4.4 HARMONIC MEAN (H.M.)

The reciprocal of the H.M. of a variate is the A.M. of its reciprocal.

For the observations $x_1, x_2, \ldots\ldots, x_n$

$$\text{H.M.} = \frac{n}{\Sigma(1/x_i)}$$

For simple frequency distribution,

$$\text{H.M.} = \frac{\text{N}}{\Sigma(f_i / x_i)}, \quad \text{N} = \Sigma f_i$$

For grouped frequency distribution x_i is taken as class mark.

Note: A.M. $\geq$ G.M. $\geq$ H.M.

Example 11: *Suppose a train moves 100 km with a speed of 40 km/hr, then 150 km with a speed of 50 km/hr and next 135 km with a speed of 45 km/hr. Calculate the average speed:*

Solution: To get average speed we require harmonic mean of 40,50 and 45 with 100, 150 and 135 as the respective frequency or weights.

$$\text{H.M.} = \frac{100 + 150 + 135}{100 \times \frac{1}{40} + 150 \times \frac{1}{50} + 135 \times \frac{1}{45}} = \frac{385}{8.5} = 45.29$$

Hence the average speed per hour is 45.29 km.

Note: In the case of grouped frequency distributions with open end class at one extremity or at both the extremities, the A.M., G.M. and H.M. cannot be computed unless we make some plausible assumptions.

Check Your Progress

Fill in the blanks:

1. Measure of central tendency is also called .....
2. Arithmetic mean is good measure for .....
3. The term 'weight' stands for the ..... of different items.
4. Step-deviation is not used in case of ..... in a continuous frequency distribution.
5. Value of mean is highly affected by ..... and ..... values.

NOTES

4.5 MEDIAN

The 'median' is another important and widely used measure of central tendency. Median of a distribution is the value of the variable which divides it into two equal parts, *i.e.,* median is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a positional average.

In case of ungrouped data, if the number of observations is odd, then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observation, there are two middle terms and median is obtained by taking the arithmetic mean of the two middle terms.

For example, the median of the values 25, 20, 15, 35, 18, *i.e.,* 15, 18, 20, 25, 35 is 20 and median of the values 8, 20, 50, 25, 15, 30, *i.e.,* 8, 15, 20, 25, 30, 50 is $\frac{1}{2}(20+25) = 22.5$.

Calculation of Median

The data are arranged in ascending order of magnitude to find out the value of the median. If the number of observations is odd, then the middle value is the median. If the number of observations is even, then median is the average of the two middle terms.

(i) ***Calculation of Median in Individual Series:*** For calculating median in a series of individual observations, the following steps are follow:

Step I. Arrange the data in ascending or descending order of magnitude.

Step II. If the number of observations is odd, then median is the middle value or

$$\text{Median} = \text{size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item}$$

If the number of observations is even, then median is obtained by taking the arithmetic mean of the two middle terms or

$$\text{Median} = \text{size of } \left[\frac{\left(\frac{n}{2}\right)^{\text{th}} \text{ item} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{ item}}{2}\right]$$

Example 12: *Calculate the median from the data recorded on the number of clusters per plant in a pulse crop:*

Number of clusters = 10, 18, 17, 19, 10, 15, 11, 17, 12.

Solution: Arrange the data in ascending order, *i.e.,*

10, 10, 11, 12, 15, 17, 17, 18, 19, $n = 9$, *i.e.,* odd

$$\text{Median} = \text{size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{size of } \left(\frac{9+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{size of } 5^{\text{th}} \text{ item}$$

Median = 15.

(ii) **Calculation of Median in Discrete Series:** In case of discrete frequency distribution, median is obtained by considering the cumulative frequency (c.f.). The steps for calculating median are given below:

Step I. Arrange the data in ascending or descending order of magnitude.

Step II. Find out the cumulative frequencies.

Step III. Find $\frac{N}{2}$, where N = Σf.

Step IV. See the c.f. just greater than N/2.

Step V. The corresponding value of *x* is median.

Example 12: *Calculate the median from the following frequency distribution:*

Soil pH	*5*	*5.5*	*6*	*6.5*	*7*	*7.5*	*8*	*8.5*
No. of plants germinated	*2*	*4*	*8*	*10*	*15*	*25*	*22*	*14*

Solution:

Soil *pH* **'*x*'**	**No. of Plants germinated** **'*f*'**	**Cumulative frequency** **'*c.f.*'**
5	2	2
5.5	4	6
6	8	14
6.5	10	24
7	15	39
7.5	25	64
8	22	86
8.5	14	100
	N = 100	

NOTES

Here $\frac{N}{2} = \frac{100}{2} = 50$

c.f. just greater than $\frac{N}{2}$ is 64 and the value of *x* corresponding to 64 is 7.5. Therefore, the median is 7.5.

(iii) **Calculation of Median in Continuous Series.** In case of continuous frequency distribution, the class corresponding to the c.f. just greater than $\frac{N}{2}$ is called the median class and the value of median is obtained by the following formula:

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

where l is the lower limit of the median class,

f is the frequency of the median class,

h is the magnitude of the median class,

C is the c.f. of class preceeding the median class, and

$N = \Sigma f$ is the total number of frequency.

Example 14: *Calculate the median from the following frequency distribution:*

No. of branches	*0–3*	*3–6*	*6–9*	*9–12*	*12–15*
No. of plants	*4*	*8*	*22*	*10*	*4*

Solution:

No. of branches 'classes'	**No. of plants frequency '*f*'**	**Cumulative frequency '*c.f*'**
0–3	4	4
3–6	8	12
6–9	22	34
9–12	10	44
12–15	4	48
	48	

Here $\frac{N}{2} = \frac{48}{2} = 24,$

c.f. just greater than 24 is 34 and corresponding class is 6–9. Therefore, the median class is 6–9.

Hence,

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

where $l = 6, h = 3, f = 22, C = 12$

$$\text{Median} = 6 + \frac{3}{22}(24 - 12) = 6 + \frac{3}{22} \times 12$$
$$= 6 + 1.64 = 7.64$$

Thus Median is 7.64.

Calculation of Median in Unequal Class-Intervals

In unequal class-intervals, frequencies need not be adjusted to make the class-intervals equal. The formula given in continuous series can be used here.

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

Example 15: *Calculate the median from the following frequency distribution:*

Class-intervals	*0–10*	*10–30*	*30–60*	*60–80*	*80–90*	*90–100*
Frequency	*5*	*16*	*30*	*12*	*6*	*1*

Solution:

Class-intervals	**Frequency *'f'***	**Cumulative frequency *c.f.***
0–10	5	5
10–30	16	21
30–60	30	51
60–80	12	63
80–90	6	69
90–100	1	70
	$N = 70$	

Here $\frac{N}{2} = \frac{70}{2} = 35$

c.f. just greater than 35 is 51 and corresponding class is 30–60. Therefore, the median class is 30–60.

$l = 30, h = 30, f = 30, C = 21.$

$$\text{Median} = 30 + \frac{30}{30}(35 - 21) = 30 + 14 = 44$$

Hence the median is 44.

Note: If we make the class-intervals equal, the same answer should be obtained.

NOTES

Calculation of Median in Open-End Classes

Since the median is not affected by the values of extreme ends, we are not concerned with extreme values for calculation of median in open-end classes.

Example 16: *Calculate the median for the following data:*

Class-intervals	*Less than 10*	*10–20*	*20–30*	*30–40*	*40 and above*
Frequency	*4*	*8*	*14*	*6*	*4*

Solution:

Class-intervals	**Frequency** ***'f'***	**Cumulative frequency** ***c.f.***
Less than 10	4	4
10–20	8	12
20–30	14	26
30–40	6	32
40 and above	4	36
	$N = 36$	

Here $\frac{N}{2} = \frac{36}{2} = 18,$

c.f. just greater than $\frac{N}{2}$ is 26 and the corresponding class is 20–30. Therefore, the median class is 20–30.

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

$$l = 20,\ h = 10,\ f = 14,\ C = 12$$

$$\text{Median} = 20 + \frac{10}{14}(18 - 12) = 20 + 4.29 = 24.29$$

Median = 24.29.

4.6 MODE

The 'Mode' is another measure of central tendency which is conceptually very useful. It is derived from the French word "La mode" which means fashion. "Mode is the value which occurs most frequently in a set of observations and around which the other items

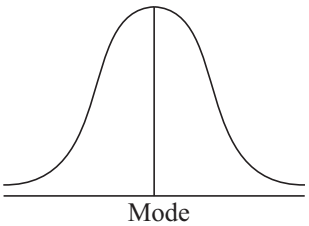


Fig. 4.1

of the set cluster densely." In other words, mode is the value of the variable which is predominant in the series. For example, the mode of a series 3, 5, 8, 5, 4, 5, 9, 3 would be 5, since the value 5 occurs most frequently than any of the others. "The value of the variable at which the curve reaches a maximum is called the mode".

There are many situations in which arithmetic mean and median fail to reveal the true characteristic of data. For example, when we talk of most common wage, most common income, most common height, most common size of sole or ready-made garments, we have in mind mode and not the arithmetic mean and median discussed earlier.

Calculation of Mode

Mode is calculated by different methods, depending upon the nature of the series.

(i) ***Calculation of Mode in Individual Observations:*** For determining the mode, count the number of items, the various values repeat themselves and the value occuring the maximum number of times is the modal value.

Example 17: *Calculate the mode of the following data relating to the weights of a sample of 10 experimental animals:*

S. No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Weight (kg.)	*10*	*11*	*10*	*12*	*12*	*11*	*9*	*8*	*11*	*11*

Solution:

Weight (*kg.*)	**No. of animals**
8	1
9	1
10	2
11	4
12	1
13	1

Since the item 11 occurs the maximum number of times, *i.e.,* 4, hence the modal value is 11.

(ii) ***Calculation of Mode in Discrete Series:*** In case of discrete frequency distribution, mode is the value of x corresponding the maximum frequency. For example, in the following frequency distribution:

Size of garment	28	29	30	31	32	33
No. of persons wearing	10	20	40	65	50	15

NOTES

From the above data, we can clearly say that the modal size is 31, because the value 31 has occured the maximum number of times, *i.e.,* 65.

Another method for discrete frequency distribution is "method of grouping".

In any one (or more) of the following cases:

I. If the maximum frequency is repeated,

II. If the maximum frequency occurs in the very beginning or at the end of the distribution, and

III. If there are irregularities in the distribution, the mode is determined by the ''method of grouping".

A 'grouping table' has six columns:

Column I. Maximum frequency (given frequency) is marked or put in a circle.

Column II. Frequencies are grouped in two's and maximum frequency is marked.

Column III. Leave the first frequency and then group the remaining in two's and maximum is marked.

Column IV. Group the frequencies in three's and maximum is marked.

Column V. Leave the first frequency and then group the remaining in three's and maximum is marked.

Column VI. Leave the first two frequencies and then group the remaining in three's and maximum frequency in marked.

After preparing the grouping table, prepare an "Analysis Table". While preparing analysis table, put column number on the left hand side and various probable values of mode on the right hand side. The values against which frequencies are the highest are marked in grouping table.

Example 18: *From the following data of the height of 100 persons in a commercial concern, determine the modal height:*

Height (in inches)	*58*	*60*	*61*	*62*	*63*	*64*	*65*	*66*	*68*	*70*
No. of persons	*4*	*6*	*5*	*10*	*20*	*22*	*24*	*6*	*2*	*1*

Solution:

Grouping Table

Height (in inches)	Col. *I*	Col. *II*	Col. *III*	Col. *IV*	Col. *V*	Col. *VI*
58	4	10	11	15	21	35
60	6					
61	5	15	30	52	66	
62	10					
63	20	42	46			52
64	22					
65	24	30	8	32	9	
66	6					
68	2	3				
70	1					

Analysis Table

Col. No.	Height (in inches)									
	58	*60*	*61*	*62*	*63*	*64*	*65*	*66*	*68*	*70*
I							1			
II					1	1				
III						1	1			
IV				1	1	1				
V					1	1	1			
VI						1	1	1		
Total				1	3	5	4	1		

Since the value 64 has occured the maximum number of times, *i.e.,* 5, the modal height is 64 inches.

(iii) ***Calculation of Mode in Continuous Series:*** In case of continuous frequency distribution, mode is given by the formula:

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(2f_1 - f_0 - f_2)},$$

where l is the lower limit of the modal class,

h is the magnitude of the modal class,

f_1 is the frequency of the modal class,

f_0 is the frequency of the class preceding the modal class,

f_2 is the frequency of the class succeeding the modal class.

Example 19: *Determine the modal size for the following frequency distribution:*

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

Solution: Here the maximum frequency is 28. Thus the class 40–50 is the modal class

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(2f_1 - f_0 - f_2)}$$

$$l = 40,\ h = 10,\ f_1 = 28,\ f_0 = 12,\ f_2 = 20.$$

$$\text{Mode} = 40 + \frac{10(28 - 12)}{(2 \times 28 - 12 - 20)} = 46.67$$

Hence the modal size is 46.67.

Calculation of Mode in Unequal Class-Intervals

The formula can be applied only in problems where the class-intervals are equal. Before, we compute the mode in unequal class-intervals, the class-intervals should be made equal and frequencies should be adjusted accordingly.

Example 20: *Calculate the mode of the following frequency distribution:*

Class-intervals	*100–110*	*110–130*	*130–140*	*140–160*	*160–170*	*170–180*
Frequency	*11*	*40*	*27*	*34*	*12*	*6*

Solution: In this problem, the class-intervals are unequal, therefore, we must adjust the frequencies and make the class-intervals equal.

Class-interval	**Frequency**
100–110	11
110–120	20
120–130	20
130–140	27
140–150	17
150–160	17
160–170	12
170–180	6

Here the maximum frequency is 27 and the class 130–140 is the modal class.

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(2f_1 - f_0 - f_2)}$$

$$l = 130, h = 10, f_1 = 27, f_0 = 20, f_2 = 17$$

$$\text{Mode} = 130 + \frac{10\,(27 - 20)}{(2 \times 27 - 20 - 17)} = 130 + 4.12$$

$$\text{Mode} = 134.12.$$

Merits and Demerits of Mode

Merits

(i) It is easy to calculate, sometimes it is found only by inspection.

(ii) It is not affected by extreme values.

(iii) It can be calculated from open end classes.

(iv) It is simple and precise.

(v) Mode is that point where there is more concentration of frequencies. Hence, it is the best representative of the data.

Demerits

(i) It is not based on all the items of the distribution.

(ii) It cannot be treated algebraically.

(iii) Equal intervals are needed for the calculation of mode, which is a draw-back.

(iv) In certain situations, it is not clearly defined. Also in case of bi-modal or multimodal distribution, it is not defined.

Uses of Mode

Mode is the average to be used to find the ideal size, *e.g.,* in business forecasting, in the manufacture of ready-made garments, shoes, etc. Mode helps the manufacturer in deciding the modals. It is useful in industry and business. Weather forecasts are also based on mode. It is very useful to agriculturists, businessmen, etc. Mode is also used in socio-economic surveys.

4.7 RELATIONSHIP AMONG MEAN, MEDIAN AND MODE

In a symmetrical distribution, mean, median and mode will coincide, *i.e.,* Mean = Median = Mode. In an asymmetrical distribution, these values will be different.

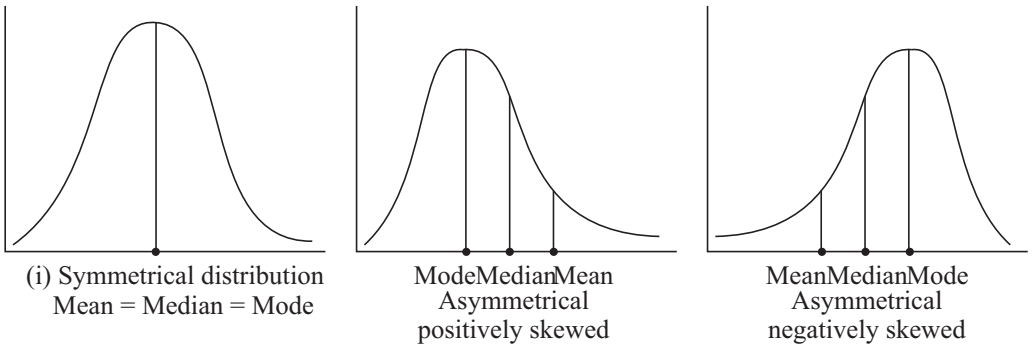


Fig. 4.2

In negatively skewed distribution, mean and median are less than mode, *i.e.,* mode is highest. In positively skewed distribution, mean and median will be more than mode, *i.e.,* mode is lowest.

For moderately asymmetrical distribution, the difference between mean and mode is three times of the differences between mean and median. Symbolically

$$\text{Mean} - \text{mode} = 3\ (\text{mean} - \text{median})$$

The above relation can be written in different ways are given below:

(i) $\text{Mean} - \text{median} = \frac{1}{3}(\text{mean} - \text{mode})$

(ii) $\boxed{\text{Mode} = 3\ \text{median} - 2\ \text{mean}}$

(iii) $\text{Median} = \text{mode} + \frac{2}{3}(\text{mean} - \text{mode})$

(iv) $\text{Mean} = \frac{1}{2}(3\ \text{median} - \text{mode})$.

Example 21: *If in a moderately asymmetrical frequency distribution, the values of median and mean are 72 and 78 respectively. Find out the value of mode.*

Solution: Mode = 3 median – 2 mean (Median = 72, Mean = 78)

= 3 × 72 – 2 × 78

= 216 – 156 = 60

Mode = 60.

Example 22: *Find out the value of median if mean = 16 and mode = 21.*

Solution: Mean = 16, mode = 21

Mode = 3 median – 2 mean

21 = 3 median – 2 × 16

3 median = 21 + 32 = 53

Median = 17.67.

4.8 QUARTILES, PERCENTILE AND DECILE

The quartiles divides the series in four equal parts. There are three quartiles name Q_1, Q_2 and Q_3, Q_1 is called the first (lower) quartile and Q_3 is called the third (upper) quartile Q_2 is called the second quartile which divides the series into two equal parts. Second quartile Q_2 coincides with median *i.e.*, the value of Q_2 and median is same. Twenty five per cent value are less than Q_1 and twenty five percent values are greater than Q_3 and the rest fifty per cent values lie between Q_1 and Q_3. Quartiles are widely used in economics and business.

$$Q_1 = \text{Size of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item for ungrouped data}$$

$$Q_2 = \text{Size of } \left(\frac{N}{4}\right)^{\text{th}} \text{ item for grouped data.}$$

For calculating the i^{th} quartile (i = 1, 2, 3) for a continuous frequency distribution following formula is used

$$Q_i = L_1 + \frac{i\left(\frac{N}{4}\right) - c.f.}{f} \times h$$

where L_1 = the lower limit of the class in which particular quartile lies

c.f. = cumulative frequency of class preceding to the particular quartile class

f = frequency of the particular quartile class

h = size of the class interval in the quartile class.

Note: $Q_1 < Q_2 < Q_3$.

Deciles

The deciles divides the series in ten equal parts. There are nine deciles namely D_1, D_2, ..., D_9. D_1 is called the first decile, D_2 is called the second decile etc. Ten per cent values are less than D_1 and ten per cent values are greater than D_9. Also the values that lie between any two deciles are ten per cent. D_5 (fifth decile) coincides with median *i.e.*, the value of D_5 and median is same.

$$D_1 = \text{Size of } \left(\frac{n+1}{10}\right)^{\text{th}} \text{ item for ungrouped data.}$$

$$D_1 = \text{Size of } \left(\frac{N}{10}\right)^{\text{th}} \text{ item for grouped data.}$$

For calculating the i^{th} decile (i = 1, 2, ..., 9) for a continuous frequency distribution following formula is used.

$$D_i = L_1 + \frac{i\left(\frac{N}{10}\right) - c.f.}{f} \times h$$

NOTES

All notations have the same meanings as in the formula for quartiles but replace the word quartile by decile.

Note: $D_1 < D_2 < ... < D_9$.

Percentiles

The percentiles divides the series in hundred equal parts. There are ninety nine percentiles namely P_1, P_2, ..., P_{99}. P_1 is called the first percentile, P_2 is called the second percentile etc. One percent values are less than P_1 and one per cent values are greater than P_{99}. Also the values that lie between any percentile are one per cent. P_{50} (50^{th} percentile) coincides with median *i.e.*, the value of P_{50} and median is same.

In particular $P_{10} = D_1$, $P_{20} = D_2$, ..., $P_{90} = D_9$

$P_{25} = Q_1$, $P_{50} = D_5 = Q_2$ = Median, $P_{75} = Q_3$

P_1 = Size of $\left(\frac{n+1}{100}\right)^{th}$ item for ungrouped data

Check Your Progress

State whether the following statements are True or False:

6. Median is a positional average.
7. For calculate median on unequal class intervals frequencies need to be adjusted to make the class-interval equal.
8. Median is affected by the values of extreme ends.
9. Mode cannot be treated algebraically.
10. Weather forecasts are based on mean.

Example 23: *Find all the quartiles for the following data:*

60, 65, 58, 61, 54, 58, 59, 68, 62, 63, 64.

Solution: Arranging the data in ascending order of magnitude we have

54, 58, 58, 59, 60, 61, 62, 63, 64, 65, 68

Here n = 11 (odd)

$$Q_1 = \text{Size of } \left(\frac{n+1}{4}\right)^{th} \text{ item} = \frac{11+1}{4} = 3^{rd} \text{ item} = 58$$

$$Q_2 = \text{Median} = \text{Size of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \frac{11+1}{2} = 6^{th} \text{ item} = 61$$

$$Q_3 = \text{Size of } 3\left(\frac{n+1}{4}\right)^{th} \text{ item} = \frac{3(11+1)}{4} = 9^{th} \text{ item} = 64.$$

NOTES

Example 24: *Find the values of lower and upper quartiles, D_2 and P_{30} from the following data:*

Marks	*0—10*	*10—20*	*20—40*	*40—60*	*60—80*	*80—100*
No. of students	*3*	*5*	*20*	*22*	*7*	*3*

Solution: First we make cumulative table as:

Marks	**No. of students (*f*)**	**Cumulative frequencies (*c.f.*)**
0—10	3	3
10—20	5	8
20—40	20	28
40—60	22	50
60—80	7	57
80—100	3	60
Total	$N = \Sigma f = 60$	

For lower quartile (Q_1) $\frac{N}{4} = \frac{60}{4} = 15$

The cumulative frequency just greater than 15 is 28. So 20–40 is the lower quartile class.

$$L_1 = 20,\ c.f. = 8,\ f = 20,\ h = 20$$

$$Q_1 = L_1 + \frac{\frac{N}{4} - c.f.}{f} \times h = 20 + \frac{15-8}{20} \times 20$$

$$= 20 + 7 = 27 \text{ marks}$$

For upper quartile (Q_3)

$$\frac{3N}{4} = \frac{3 \times 60}{4} = 45$$

The cumulative frequency just greater than 45 is 50. So 40–60 is the upper quartile class.

$$L_1 = 40,\ c.f. = 28,\ f = 22,\ h = 20$$

$$Q_3 = L_1 + \frac{\frac{3N}{4} - c.f.}{f} \times h$$

$$= 40 + \frac{45 - 28}{22} \times 20 = 40 + 15.454 = 55.454 \text{ marks}$$

For second decile (D_2)

$$\frac{2N}{10} = \frac{2 \times 60}{10} = 12$$

The cumulative frequency just greater than 12 is 28. So 20–40 is the second decile class.

$$L_1 = 20, c.f. = 8, f = 20, h = 20$$

$$D_2 = L_1 + \frac{\frac{2N}{10} - c.f.}{f} \times h = 20 + \frac{12 - 8}{20} \times 20$$

$$= 20 + 4 = 24 \text{ marks}$$

For 30th percentile (P_{30})

$$\frac{30N}{100} = \frac{30 \times 60}{100} = 18$$

The cumulative frequency just greater than 18 is 28. So 20–40 is the 30th percentile class.

$$L_1 = 20, c.f. = 8, f = 20, h = 20$$

$$P_{30} = L_1 + \frac{\frac{30N}{100} - c.f.}{f} \times h$$

$$= 20 + \frac{18 - 8}{20} \times 20 = 20 + 10 = 30 \text{ marks.}$$

Example 25: *Find the values of Q_1, Q_3, D_8 and P_{55} from the following data:*

Marks less than	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*
No. of students	*4*	*16*	*40*	*76*	*96*	*112*	*120*	*125*

Solution: The given data is in cumulative frequency form. First we calculate the frequency of each class as follows:

Marks	Frequency (f)	Cumulative frequencies ($c.f.$)
0—10	4	4
10—20	16 – 4 = 12	16
20—30	40 – 16 = 24	40
30—40	76 – 40 = 36	76
40—50	96 – 76 = 20	96
50—60	112 – 96 = 16	112
60—70	120 – 112 = 8	120
70—80	125 – 120 = 5	125
Total	$N = \Sigma f = 125$	

For lower quartile (Q_1)

$$\frac{N}{4} = \frac{125}{4} = 31.25$$

The cumulative frequency just greater than 31.25 is 40. So 20–30 is the Q_1 class.

$$L_1 = 20,\ c.f. = 16,\ f = 24,\ h = 10$$

$$Q_1 = L_1 + \frac{\frac{N}{4} - c.f.}{f} \times h$$

$$= 20 + \frac{31.25 - 16}{24} \times 10 = 20 + 6.354 = 26.354 \text{ marks}$$

For upper quartile (Q_3)

$$\frac{3N}{4} = \frac{3 \times 125}{4} = 93.75$$

The cumulative frequency just greater than 93.75 is 96. So 40–50 is the Q_3 class.

$$L_1 = 40,\ c.f. = 76,\ f = 20,\ h = 10$$

$$Q_3 = L_1 + \frac{\frac{3N}{4} - c.f.}{f} \times h$$

$$= 40 + \frac{93.75 - 76}{20} \times 10 = 40 + 8.875 = 48.875 \text{ marks}$$

For 8th decile (D_8)

$$\frac{8N}{10} = \frac{8 \times 125}{10} = 100$$

The cumulative frequency just greater than 100 is 112. So 50–60 is the D_8 class.

$$L_1 = 50,\ c.f. = 96,\ f = 16,\ h = 10$$

$$D_8 = L_1 + \frac{\frac{8N}{10} - c.f.}{f} \times h$$

$$= 50 + \frac{100 - 96}{16} \times 10 = 50 + 2.5 = 52.5 \text{ marks}$$

For 55th percentile (P_{55})

$$\frac{55N}{100} = \frac{55 \times 125}{100} = 68.75$$

The cumulative frequency just greater than 68.75 is 76. So 30–40 is the P_{55} class.

$$L_1 = 30,\ c.f. = 40,\ f = 36,\ h = 10$$

$$P_{55} = L_1 + \frac{\frac{55N}{100} - c.f.}{f} \times h$$

$$= 30 + \frac{68.75 - 40}{36} \times 10 = 30 + 7.98 = 37.98 \text{ marks.}$$

4.9 SUMMARY

- For quantitative data it is observed that there is a tendency of the data to be distributed about a central value which is a typical value and is called a measure of central tendency. It is also called a measure of location because it gives the position of the distribution on the axis of the variable.
- The arithmetic mean is simply called 'Average'. For the observations x_1, x_2, ..., x_n the A.M. is defined as

$$\bar{x} = \text{A.M.} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- "Algebraic sum of the deviations of a set of values from their arithmetic mean is zero."

- If the values of the variable are multiplied or divided by a constant value, the arithmetic mean so obtained is same as the initial arithmetic mean is multiplied or divided by the constant value.
- If unequal intervals are given in a continuous frequency distribution, then it is not necessary to convert them in equal class-interval.
- If the data are given in the form of an inclusive series, there is no need to change it into an exclusive series, because mid-points are unaffected in this case.
- One of the limitation of the arithmetic mean is that it gives equal importance to all the items.
- The term 'weight' stands for the relative importance of the different items.
- The geometric mean of the observations x_1, x_2, ..., x_n is defined as

$$\text{G.M.} = (x_1 . x_2 ..... x_n)^{1/n}$$

- The reciprocal of the H.M. of a variate is the A.M. of its reciprocal.

  For the observations x_1, x_2, ..., x_n

$$\text{H.M.} = \frac{n}{\Sigma(1 / x_i)}$$

- Median of a distribution is the value of the variable which divides it into two equal parts, *i.e.,* median is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a positional average.
- In case of discrete frequency distribution, median is obtained by considering the cumulative frequency (c.f.).
- "Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely."
- In case of discrete frequency distribution, mode is the value of x corresponding the maximum frequency.
- In a symmetrical distribution, mean, median and mode will coincide, *i.e.,* Mean = Median = Mode. In an asymmetrical distribution, these values will be different.
- In negatively skewed distribution, mean and median are less than mode, *i.e.,* mode is highest. In positively skewed distribution, mean and median will be more than mode, *i.e.,* mode is lowest.

NOTES

4.10 GLOSSARY

- ***Measure of central tendency:*** For quantitative data it is observed that there is a tendency of the data to be distributed about a central value which is a typical value and is called a measure of central tendency.
- ***Weighted arithmetic mean:*** The term 'weight' stands for the relative importance of the different items. The formula for computing weighted arithmetic mean is

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad ,$$

- ***Average:*** The arithmetic mean is simply called 'Average'.

4.11 ANSWERS TO CHECK YOUR PROGRESS

1. measures of location
2. comparision
3. relative importance
4. unequal interval
5. abnormal, extreme
6. True
7. False
8. False
9. True
10. False

4.12 TERMINAL AND MODEL QUESTIONS

1. What do you mean by measures of central tendency? Name various measures of central tendency.
2. What are the characteristics of a good average?
3. Define median and discuss its properties. Also discuss its merits and demerits.

4. What do you know about mode? Discuss its merits and demerits.
5. Systolic B.P. of normal subject (n = 10) have been recorded below in mm of Hg:

   130, 134, 132, 130, 128, 142, 131, 140, 133, 140.

   Calculate the arithmetic mean by

   (i) Direct method (ii) Short-cut method
6. The data recorded on the number of chlorophyll deficient plants in a lentil population are given below. Calculate the arithmetic mean.

No. of chlorophyll deficient plants	*No. of plants*
0	34
1	14
2	20
3	24
4	25
5	33

7. Calculate arithmetic mean of protein intake of 400 families:

Protein intake/consumption unit/day (g)	15–25	25–35	35–45	45–55	55–65	65–75	75–85
No. of families	30	40	100	110	80	30	10

8. The marks obtained by 50 biological students in a statistics course are given below. Find the average marks.

Marks	30–39	40–49	50–59	60–69	70–79	80–89
Frequency	2	4	4	20	7	3

9. Following are the serum calcium levels in 100 normal adults. Find the average serum calcium level.

Class-boundaries	*Frequency*	*Class-boundaries*	*Frequency*
7.44–7.99	4	10.24–10.79	20
8.00–8.55	3	10.80–11.35	10
8.56–9.11	12	11.36–11.91	8
9.12–9.67	19	11.92–12.47	2
9.68–10.23	21	12.48–13.03	1

10. Mean of 100 items is found to be 30. If at the time of calculation two items are wrongly taken as 32 and 12 instead of 23 and 11, find the correct mean.
11. The value of mode and median for a moderately skewed distribution are 64.2 and 68.6 respectively. Find the value of the mean.
12. Three teachers of statistics reported mean marks of their classes consisting of 69, 64 and 71 students as 30, 26 and 18. Determine the mean marks for all the three classes.
13. Data recorded on the production of butter fat during 10 consecutive days in cows is presented below. Calculate the median.
    Butter fat (kg) = 4.0, 5.7, 3.9, 4.2, 6.6, 7.0, 7.9, 8.0, 9.0, 10.0.
14. Calculate the median from the following frequency distribution of marks at a test M. Sc. (Biotech) students in biostatistics:

Marks	5	10	15	20	25	30	35	40	45	50
No. of students	20	43	75	76	72	45	39	9	8	6

15. Find the median of protein intake of 200 families:

Protein intake/consumption unit/day (g)	*No. of families*
5–15	15
15–25	20
25–35	50
35–45	55
45–55	40
55–65	15
65–75	5

16. Calculate the median from the following statistical distribution:

Value	Less than 100	100–200	200–300	300–400	400 and above
Frequency	40	89	148	64	39

17. Determine the modal plant height from the following data of the height of 100 wrinkled seeded plants.

Height in inches	56	60	61	62	63	64	65	66	68	69
No. of plants	4	6	5	10	20	22	24	6	2	1

18. Following data relate to the number of patient's stay in the hospital. Find the value of mode.

No. of days admitted	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45
No. of patients	29	195	241	117	52	10	6	3	2

19. The marks obtained by 15 students in a class test are as follows:
6, 9, 10, 12, 18, 19, 23, 23, 24, 28, 37, 48, 49, 53 and 60
Find all the quartiles.

20. Find all the quartiles for the following frequency distribution

Marks	5—10	10—15	15—20	20—25	25—30	30—35	35—40	40—45
No. of students	5	6	15	10	5	4	2	2

21. Find the values of lower and upper quartiles, D_2 and P_{30} from the following data:

Marks	0—10	10—20	20—40	40—60	60—80	80—100
No. of students	8	10	22	25	10	5

22. Find the values of median, first quartile, third quartile, D_9 and P_{19} from the following data:

Class interval	11—15	16—20	21—25	26—30	31—35	36—40	41—45	46—50	51—55
Frequency	8	15	39	47	52	41	28	16	4

4.13 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 5: MEASURES OF DISPERSION

NOTES

Structure

5.0 INTRODUCTION

Averages or measures of central tendency give us an idea of the concentration of the observations about the central part of the distribution. If we know the average alone, we cannot form a complete idea about the distribution. The central value may be same in two or more distribution, but may differ in respect of dispersion. The measures of dispersion help us in studying the important characteristics of the distribution.

"Literal meaning of dispersion is 'Scatteredness'. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution."

Dispersion is important not only as merely supplementary to the average, but because of the scatter distribution. According the Spurr and Bonini, "In matters of health, variations in body temperature, pulse beat and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variation." In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programmes.

Properties of a Good Measure of Dispersion

Characteristics of an ideal measure of dispersion are the same that of average, viz. in brief :

(i) It should be simple to understand and easy to compute.
(ii) It should be rigidly defined.
(iii) It should be based on all the observations.
(iv) It should be amenable to further algebraic treatment.
(v) It must have sampling stability.
(vi) It should not be affected by extreme observations.

5.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define dispersion and properties of good measures of dispersion
- Name various measures of dispersion
- Define range, its merits and demerits
- Define quartile deviation and interquartile range
- Define mean deviation and its merits and demerits
- Define standard deviation and its merits and demerits

5.2 MEASURES OF DISPERSION

A measure of dispersion describes the degree of scatter shown by the observations and is usually measured as an average deviation about some central value. Measures of dispersion gives us additional information that enables us to judge the reliability of our measure of central value. It makes possible to compare two series of data in respect of their variability.

Following are the most commonly used measures of dispersion:

1. Range
2. Interquartile range and quartile deviation
3. Mean deviation
4. Standard deviation
5. Coefficient of variation.

Absolute and Relative Measure of Dispersion

Measures of dispersion may be either absolute or relative. Absolute measure of dispersion cannot be used for comparison purposes if expressed in different units. The absolute measure of dispersion can be compared with another, only if the two belong to the same population. For instance, when we measure the height and weight of the students they may be in metres and kilograms respectively–two different units. These different units cannot be measured through absolute method, to know the variability. Therefore, two series cannot be compared if the absolute measure of dispersion of each series is expressed as a ratio or percentage of the average. For comparing the variability, even if the distributions are in the same units, the relative measure of dispersion is computed. In other words, the coefficient of dispersion of each group should be calculated in order to compare two series.

$$\text{Relative measure of dispersion} = \frac{\text{Absolute measure of dispersion}}{\text{Mean}} \times 100$$

The relative measures of dispersion are:

(i) $\text{Coefficient of range} = \dfrac{L - S}{L + S}$

(ii) $\text{Coefficient of quartile deviation} = \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

(iii) $\text{Coefficient of mean deviation} = \dfrac{\text{Mean Deviation}}{\text{Mean}}$

5.3 RANGE

The difference between the largest and smallest value of the variates is called its range.

$$\text{Range} = L - S,$$

where L = Largest value

S = Smallest value

$$\text{Coefficient of range} = \frac{L - S}{L + S}.$$

Advantages of Range

1, It is simple to understand and easy to calculate.

2. It is used to study variations in the prices of commodities and movement in the prices of securities.

3. It is used in weather forcasting *e.g.*, It gives an idea of the variation between maximum and minimum levels of temperature.
4. It is used in quality control for drawing R-charts.

Disadvantages of Range

1. It depends only on two values (largest and smallest) and ignores all other values, it is highly misleading.
2. It is not useful for frequency distribution.
3. It is affected by sampling fluctuation.
4. It cannot be computed if the distribution is open-ended.
5. It is not suitable for further mathematical treatment.

Uses of Range

(i) Range is used in industries for the statistical quality control of manufactured product by the construction of control chart.

(ii) The meteorological department uses the range for weather forecasts since public is interested to know the limits within which the temperature is likely to vary on a particular day.

(iii) Range is useful in studying the variations in the prices of stock, shares, and other commodities that are sensitive to price changes from one period to another period.

Check Your Progress

Fill in the blanks:

1. Literal meaning of dispersion is .....
2. Measures of dispersion enable us to judge the reliability of our measure of .....
3. Measures of dispersion may be either ..... or .....
4. The difference between the largest and smallest value of variates is called its .....
5. Range is used in quality control for drawing .....

Example 1: *Following are the systolic blood pressures of 8 persons recorded in mm of Hg:*

S.No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
B.P.	*130*	*120*	*134*	*126*	*143*	*137*	*132*	*148*

Find the range and coefficient of range.

NOTES

Solution: The lowest and highest blood pressures are 120 and 148 respectively. Thus,

$$\text{Range} = L - S$$

$$= 148 - 120 = 28$$

$$\text{Range} = 28$$

$$\text{Coefficient of range} = \frac{L-S}{L+S} = \frac{148-120}{148+120} = \frac{28}{268} = 0.1044$$

5.4 QUARTILE DEVIATION OR SEMI-INTERQUARTILE RANGE

Quartile deviation or semi-interquartile range (Q.D.) is given by

$$Q.D. = \frac{Q_3 - Q_1}{2},$$

where Q_3 is the third quartile and

Q_1 is the first quartile.

Quartile deviation is defined as half the distance between the third and the first quartile.

Quartile deviation is an absolute measure of dispersion. The relative measure of dispersion, known as coefficient of quartile deviation, is calculated as follows :

$$\text{Coefficient of } Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

The quartile deviation is an improved measure over the range, as it is not calculated from extreme items, but on quartiles.

The quartile deviation gives an idea of the distribution of the middle half of the items around the median.

Interquartile Range

The difference between the third quartile and first quartile is known as interquartile range. It is given by

$$\text{Interquartile range} = Q_3 - Q_1,$$

where Q_3 is the third quartile and Q_1 is the first quartile.

Merits and Demerits of Quartile Deviation

Merits

(i) It is simple to understand and easy to compute.

(ii) It is not influenced by the extreme values.

(iii) It can be found out with open end distribution.

(iv) It is not affected by the presence of extreme values.

Demerits

(i) It ignores the first 25% of the items and the last 25% of the items.

(ii) It is a positional average; hence not amenable to further mathematical treatment.

(iii) Its value is affected by sampling fluctuations.

(iv) It gives only a rough measure.

Example 2: *From the following data, calculate interquartile range, quartile deviation and coefficient of quartile deviation:*

48, 45, 54, 43, 51, 49, 38, 41, 37, 42, 46.

Solution: Arrange the observations in ascending order.

37, 38, 41, 42, 43, 45, 46, 48, 49, 51, 54.

$$Q_1 = \text{size of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}, \quad n = 11$$

$$= \text{size of } \left(\frac{11+1}{4}\right)^{\text{th}} \text{ item}$$

$$= \text{size of } 3^{\text{th}} \text{ item}$$

$$Q_1 = 41$$

5.5 MEAN DEVIATION

Mean deviation (M.D.) is defined as the average of the absolute deviations taken from an average usually, the mean, median or mode. Mean deviation is a measure of dispersion which is based on all values of a set of data.

Mean Deviation from Ungrouped Data

The formula for calculating mean deviation (M.D.) of n values $x_1, x_2, \ldots\ldots, x_n$ *is*

$$\text{M.D.} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \text{A}|,$$

where A = Any constant out of mean, median and mode.

NOTES

Mean Deviation from Grouped Data

1. ***Direct method:*** For a frequency distribution in which the variate value x_i occurs f_i times ($i = 1, 2, ..., k$), the formula is

$$\text{M.D.} = \frac{1}{\text{N}} \sum_{i=1}^{k} f_i \, | \, x_i - \text{A} \, |,$$

where $\text{N} = \sum_{i=1}^{k} f_i$ and A as defined above.

Merits and Demerits of Mean Deviation

Merits

(i) It is simple to understand and easy to compute.

(ii) It is based on all the observations.

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

(iv) It is less affected by the extreme values.

(v) It is rigidly defined.

(vi) It is better measure for comparison.

(vii) It is flexible, because it can be calculated from any measure of central tendency.

Demerits

(i) It is a non-algebraic treatment.

(ii) Algebraic positive and negative signs are ignored. In mean deviation + 5 and – 5 have the same meaning. It is mathematically unsound and illogical.

(iii) It is not a very accurate measure of dispersion.

(iv) It is not suitable for further mathematical calculation.

(v) It is rarely used. It is not as popular as standard deviation.

Uses of Mean Deviation

(i) It will help to understand the standard deviation.

(ii) It is useful in marketing problems.

(iii) It is useful while using small samples.

(iv) It is used in statistical analysis of economic business and social phenomena.

(v) It is useful in calculating the distribution of wealth in a community or a nation.

(vi) It is useful in forecasting business cycles.

Example 3: *Duration (in days) of sickness of 10 patients is given:*

9, 7, 8, 10, 70, 5, 6, 8, 9, 8.

Find the mean deviation from mean and median and also their coefficients:

Solution:

Calculation of Mean, Median, and M.D.

Duration (in days) 'x'	**Absolute deviation from mean $\lvert x - \bar{x} \rvert$**	**Absolute deviation from median $\lvert x - M_d \rvert$**
9	$\lvert 9 - 14 \rvert = 5$	1
7	7	1
8	6	0
10	4	2
70	56	62
5	9	3
6	8	2
8	6	0
9	5	1
8	6	0
$\Sigma x = 140$	$\Sigma\lvert x - \bar{x} \rvert = 112$	$\Sigma\lvert x - M_d \rvert = 72$

Mean, $\bar{x} = \frac{1}{n}\Sigma x = \frac{1}{10} \times 140 = 14.$

For median, observations are arranged in ascending order, *i.e.,*

5, 6, 7, 8, 8, 8, 9, 9, 10, 70

$$\text{Median} = \frac{\text{Size of } \left\{\left(\frac{n}{2}\right)^{\text{th}} + \left(\frac{n}{2}+1\right)\right\}^{\text{th}} \text{ items}}{2}$$

$$= \frac{\text{Size of } \{5^{\text{th}} + 6^{\text{th}}\} \text{ items}}{2} = \frac{8+8}{2} = 8.$$

Median = 8.

$$\text{Mean deviation about mean} = \frac{1}{n}\Sigma\lvert x - \bar{x} \rvert$$

$$= \frac{1}{10} \times 112 = 11.2$$

$$\text{Coefficient of M.D. about mean} = \frac{\text{M.D.}}{\text{mean}} = \frac{11.2}{14} = 0.77$$

$$\text{Mean deviation about median} = \frac{1}{n}\Sigma\lvert x - M_d \rvert$$

$$= \frac{1}{10} \times 72 = 7.2$$

$$\text{Coefficient of M.D. about median} = \frac{\text{M.D.}}{\text{Median}} = \frac{7.2}{8} = 0.9.$$

NOTES

5.6 STANDARD DEVIATION

Standard deviation is the most commonly used absolute measure of dispersion. The concept of standard deviation was first introduced by Karl Pearson in 1893. The 'Standard' is assigned to this measure of variation is probably because it is the most commonly used and is the most flexible in terms of variety of applications of all the measures of dispersions. It is clear that standard deviation is a measure of the spread in a set of observations. In this method, the drawback of ignoring the algebraic sign as in mean deviation is overcome by taking the square of deviations, thereby, making all the deviations positive.

"Standard deviation is positive square root of the arithmetic mean of the squares of the deviations taken from arithmetic mean". It is usually denoted by small Greek letter σ (sigma).

The formula for the standard deviation for an "ungrouped data" is

$$\text{Standard deviation } \boxed{(\sigma) = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}} \quad \text{or} \quad \sqrt{\frac{1}{n}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]}$$

For the grouped data (frequency distribution)

$$\text{Standard deviation } \boxed{(\sigma) = \sqrt{\frac{1}{N}\Sigma f(x-\bar{x})^2}} \quad \text{or} \quad \sqrt{\frac{1}{N}\left[\Sigma fx^2 - \frac{(\Sigma fx)^2}{N}\right]}$$

As it measures the dispersion or variability of a distribution, the larger the standard deviation, the greater is the value of variability. On the other hand, there will be homogeneity in a series when the standard deviation is small.

$$\boxed{\text{Coefficient of standard deviation} = \frac{\text{Standard deviation}}{\text{Mean}} = \frac{\sigma}{\bar{x}}}.$$

Mathematical Properties of Standard Deviation

Standard deviation has the following mathematical properties:

(i) Standard deviation of the combined series : If n_1 and n_2 are the sizes, $\bar{x}_1$ and $\bar{x}_2$ the means and σ_1 and σ_2 the standard deviations of the two series, then the standard deviation (σ) of the combined series is given by:

$$\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}},$$

where $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$ and $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ is the mean of combined series.

(ii) Standard deviation of *n* natural numbers : The standard deviation of the first *n* natural numbers can be obtained by the following formula :

$$\sigma = \sqrt{\frac{1}{12}(n^2 - 1)}$$

For example, the standard deviation of natural numbers 1 to 10 will be

$$\sigma = \sqrt{\frac{1}{12}(10^2 - 1)} = \sqrt{\frac{99}{12}} = 2.87.$$

(iii) The sum of the squares of deviations of items in the series from their arithmetic mean in minimum.

(iv) The standard deviation of a series remains unchanged if each observation of the series is increased or decreased by the same constant value.

(v) If each observation of a series is multiplied or divided by the same constant value, the standard deviation can also be obtained by dividing or multiplying by the same constant value.

(vi) Standard deviation is the most extensively used measure in statistical analysis of the agricultural as well as the biological and the medical science. In normal distribution, the standard deviation help us in finding the number of items that fall within the specific ranges, *i.e.,*

$\bar{x} \pm 1\sigma$ covers 68.27% of the items,

$\bar{x} \pm 2\sigma$ covers 95.45% of the items and

$\bar{x} \pm 3\sigma$ covers 99.73% of the items.

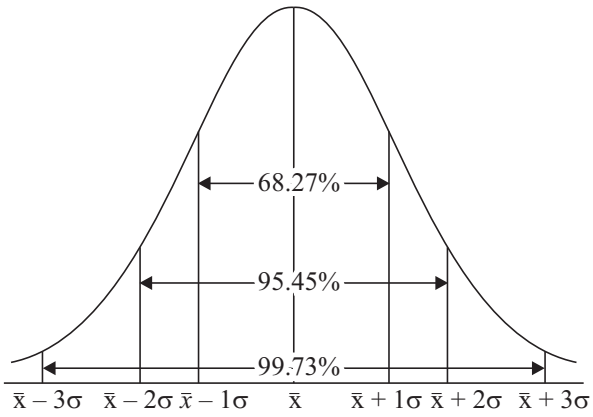


Fig. 5.1: Proportion of Items Included within + σ, + 2σ and + 3σ of the Arithmetic Mean in a Normal Curve

Merits and Demerits of Standard Deviation

Merits

(i) It is rigidly defined and its value is always definite.

(ii) It is based on all the observations and the actual signs of deviations are used.

(iii) It is less affected by sampling fluctuations.

(iv) It is possible for further algebraic treatment.

Demerits

(i) It is not easy to understand and to calculate.

(ii) It gives more weight to extreme values, because the values are squared up.

(iii) It is affected by the value of every item in the series.

(iv) As it is an absolute measure of variability, it cannot be used for the purpose of comparison.

(v) It has not found favour with the economists and businessmen.

Uses of Standard Deviation

(i) Standard deviation is the best measure of dispersion.

(ii) It is widely used in statistics because it possesses most of the characteristics of an ideal measure of dispersion.

(iii) It is widely used in sampling theory and by biologists.

(iv) It is used in coefficient of correlation and in the study of symmetrical frequency distribution.

Calculation of Standard Deviation

(i) ***Calculation of Standard Deviation in Individual Series:*** Standard deviation can be calculated by applying any of the following methods. (a) Direct Method (b) Short-cut method.

(a) ***Direct method:*** *The* formula for calculating standard deviation is given by

$$\sigma = \sqrt{\frac{1}{n}\Sigma(x - \bar{x})^2}$$

Example 4: *The following are the patients per day attended by 10 famous doctors of Lucknow city :*

25, 34, 48, 36, 42, 70, 30, 60, 45, 50

Find the standard deviation:

Solution:

Calculation of Standard Deviation

x	$(x - \bar{x})$	$(x - \bar{x})^2$
25	– 19	361
34	– 10	100
48	4	16
36	– 8	64
42	– 2	4
70	26	676
30	– 14	196
60	16	256
45	1	1
50	6	36
$\Sigma x = 440$	$\Sigma(x - \bar{x}) = 0$	$\Sigma(x - \bar{x})^2 = 1710$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{1}{10} \times 440 = 44$$

$$\sigma = \sqrt{\frac{1}{n}\Sigma (x - \bar{x})^2} = \sqrt{\frac{1}{10} \times 1710} = \sqrt{171}$$

$$\sigma = 13.076$$

Note: It is to be noted that $\Sigma (x - \bar{x})$ is always equal to zero.

(b) ***Short-cut Method:*** This method is applied when arithmetic mean is not a whole number but it is a fraction. In this method, deviations are taken from a suitable chosen assumed mean A in place of $\bar{x}$.

The following formula is used :

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2},$$

where $d = (x - A)$.

Example 5: *Find the arithmetic mean and standard deviation of the height of 10 persons given below:*

Height (in cm) : 160, 160, 161, 162, 163, 163, 163, 164, 164, 170.

Solution:

Calculation of Arithmetic Mean and Standard Deviation

Height (cm) 'x'	**$d = x - A$, where $A = 162$**	**$d^2 = (x - A)^2$**
160	– 2	4
160	– 2	4
161	– 1	1
162	0	0
163	1	1
163	1	1
163	1	1
164	2	4
164	2	4
170	8	64
	$\Sigma d = 10$	$\Sigma d^2 = 84$

Mean, $$\bar{x} = A + \frac{1}{n}\Sigma d = 162 + \frac{10}{10} = 162 + 1 = 163$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{84}{10} - \left(\frac{10}{10}\right)^2} = \sqrt{8.4 - 1} = \sqrt{7.4}$$

$= 2.72.$

(ii) ***Calculation of standard deviation in discrete series:*** In this case, standard deviation can be calculated after applying any one of the two methods :

(a) Direct Method (b) Short-cut Method

(a) ***Direct method:*** The formula for calculating standard deviation by direct method is given by :

$$\sigma = \sqrt{\frac{1}{N}\Sigma f(x - \bar{x})^2}\ .$$

Example 6: *Find the standard deviation from the following frequency distribution:*

Marks	*10*	*20*	*30*	*40*	*50*	*60*
No. of students	*8*	*12*	*20*	*10*	*7*	*3*

Solution:

Calculation of Standard Deviation

Marks 'x'	No. of student 'f'	fx	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
10	8	80	– 20.8	432.64	3461.12
20	12	240	– 10.8	116.64	1399.68
30	20	600	– 0.8	0.64	12.80
40	10	400	9.2	84.64	846.40
50	7	350	19.2	368.64	2580.48
60	3	180	29.2	852.64	2557.92
	$N = 60$	$\Sigma fx = 1850$			$\Sigma f(x - \bar{x})^2$ = 10,858.40

Mean, $$\bar{x} = \frac{1}{N}\Sigma fx = \frac{1}{60} \times 1850 = 30.8$$

and $$\sigma = \sqrt{\frac{1}{N}\Sigma f(x - \bar{x})^2} = \sqrt{\frac{1}{60} \times 10{,}858.40}$$

$$\sigma = 13.45.$$

(b) ***Short-cut Method (assumed mean):*** This method is applied when arithmetic mean is not a whole number but it is a fraction. In this method deviation taken from a suitable chosen assumed mean A in place of $\bar{x}$, *i.e.*, $d = x - A$.

The following formula is used to calculate standard deviation:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

Check Your Progress

State whether the following statements are True or False:

6. Quartile deviation is affected by the presence of extreme values.
7. Quartile deviation is an absolute measure of deviation.
8. Standard deviation is arithmetic mean of the squares of the deviations taken from arithmetic mean.
9. Mean deviation is based on all the observations.
10. Standard deviation is possible for further algebraic treatment.

NOTES

5.7 COMPARISON BETWEEN MEAN DEVIATION AND STANDARD DEVIATION

Mean deviation	Standard deviation
1. Deviations are calculated from mean, median or mode.	1. Deviations are calculated only from mean.
2. Algebraic signs are ignored while calculating mean deviation.	2. Algebraic signs are taken into account.
3. It is simple to calculate.	3. It is difficult to calculate.
4. It lacks mathematical properties, because algebraic signs are ignored.	4. It is mathematically sound, because algebraic signs are taken into account.

5.8 VARIANCE AND COEFFICIENT OF VARIATION

The square of standard deviation is called variance and is denoted by σ^2.

Symbolically:

$$\text{Variance} = \sigma^2$$

$$\sigma = \sqrt{\text{variance}}$$

The formula for calculating variance is given by :

$$\sigma^2 = \frac{1}{n}\Sigma(x - \bar{x})^2 \qquad \text{or} \qquad \sigma^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

(for individual observations)

$$\sigma^2 = \frac{1}{N}\Sigma f(x - \bar{x})^2 \qquad \text{or} \qquad \sigma^2 = \frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2$$

(for discrete and continuous series)

Example 7: *Calculate the mean, standard deviation and variance of the following frequency distribution:*

Height in inches	*95–105*	*105–115*	*115–125*	*125–135*	*135–145*
No. of children	*19*	*23*	*36*	*70*	*52*

Solution: Let assumed mean = 130.

Height in inches class-interval	No. of children '*f* '	Mid-point '*x*'	$d = \frac{x - A}{h}$	*fd*	fd^2
95–105	19	100	$\frac{100 - 130}{10} = -3$	– 57	171
105–115	23	110	$\frac{110 - 130}{10} = -2$	– 46	92
115–125	36	120	$\frac{120 - 130}{10} = -1$	– 36	36
125–135	70	130	$\frac{130 - 130}{10} = 0$	0	0
135–145	52	140	$\frac{140 - 130}{10} = 1$	52	52
	$N = 200$			$\Sigma fd = -87$	$\Sigma fd^2 = 351$

$$\text{Mean} = A + \frac{h}{N}\Sigma fd = 130 + \frac{10}{200} \times (-87) = 130 - 4.35 = 125.65$$

Mean = 125.65.

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{351}{200} - \left(\frac{-87}{200}\right)^2} \times 10$$

$$= \sqrt{1.75 - 0.189} \times 10$$

$$= \sqrt{1.56} \times 10$$

$$= 1.2489 \times 10$$

$$\sigma = 12.489$$

$$\text{variance } (\sigma^2) = (12.489)^2 = 155.97$$

variance = 156.

Coefficient of Variation

The standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original data are collected. The standard deviation of the height of plants cannot be compared with the standard deviation of the weight of plants because they are expressed in different units, *i.e.,* height (cm) and weight (gm). Therefore, the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure of dispersion is known as the coefficient of variation.

NOTES

Coefficient of standard deviation will be in fraction and as such not very good for comparison. Therefore, the coefficient of standard deviation is multiplied by 100 gives the coefficient of variation.

Whenever we want to compare the variability of the two series which differ widely in their averages or which are measured in different units, we do not calculate the measures of dispersion but we are calculate the coefficient of dispersion.

"100 times the coefficient of dispersion based upon standard deviation is called coefficient of variation (C.V.)", or coefficient of variation (C.V.) is the ratio of the standard deviation and mean. The formula for coefficient of variation (C.V.) is given by

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100 \quad .$$

For comparing the variability of two series, we calculate the coefficient of variation for each series. The series having greater C.V. is said to be more variables or less consistent, less uniform, less stable, or less homogeneous than the other and the series having lesser C.V. is said to be less variable or more consistent, more uniform, more stable or more homogeneous.

Example 8: *An analysis of monthly wages paid to the workers in a firm A and B belonging to same industry gives the following results:*

	Firm *A*	**Firm *B***
No. of workers	*500*	*600*
Average monthly wages	*₹ 186.00*	*₹ 175.00*
Variance of distribution of wages	*81*	*100*

(i) *Which firm, A or B, has a larger wage bill ?*

(ii) *In which firm, A or B, is there greater variability in individual wages ?*

(iii) *Calculate* (a) *the average monthly wage, and* (b) *the variance of the distribution of wages, of all the workers in the firm A and B taken together:*

Solution: (i) Firm A,

$$n_1 = 500, \ \bar{x}_1 = 186.00$$

$$\text{Average monthly wage} = \frac{\text{Total wage paid}}{\text{No. of workers}}$$

Hence, the total wage paid to the workers

$$= \text{Average monthly wage} \times \text{No. of workers}$$

$$= n_1 \bar{x}_1 = 500 \times 186 = ₹\ 93{,}000$$

Firm B,

$$n_2 = 600, \ \bar{x}_2 = 175.00$$

Total wage paid to the workers = $n_2 \bar{x}_2 = 600 \times 175 = ₹\ 10{,}500$

Hence we see that the firm B has larger wage bill.

(ii) Firm A,

$$\sigma_1^2 = 81,\ \bar{x}_1 = 186$$

$$\sigma_1 = 9$$

$$\text{C.V. for } A = \frac{\sigma_1}{\bar{x}_1} \times 100 = \frac{9}{186} \times 100 = 4.84$$

Firm B,

$$\sigma_2^2 = 100,\ \sigma_2 = 10,\ \bar{x}_2 = 175$$

$$\text{C.V. for } B = \frac{\sigma_2}{\bar{x}_2} \times 100 = \frac{10}{175} \times 100 = 5.71$$

Since C.V. for firm B is greater than C.V. for firm A, firm B has greater variability in individual wages.

(iii) (a) Given that

$$n_1 = 500,\ n_2 = 600,\ \bar{x}_1 = 186,\ \bar{x}_2 = 175$$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{500 \times 186 + 600 \times 175}{500 + 600}$$

$$\bar{x} = ₹\ 180.$$

(b) The combined variance σ^2 is given by the formula

$$\sigma^2 = \frac{1}{n_1 + n_2} [n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$$

Given that $n_1 = 500,\ n_2 = 600,\ \bar{x}_1 = 186,\ \bar{x}_2 = 175$

$$\sigma_1^2 = 81,\ \sigma_2^2 = 100,\ \bar{x} = 180$$

$$d_1 = \bar{x}_1 - \bar{x} = 186 - 180 = 6,\ d_2 = \bar{x}_2 - \bar{x} = 175 - 180 = -5$$

$$\sigma^2 = \frac{1}{500 + 600} [500(81 + 36) + 600(100 + 25)]$$

$$\sigma^2 = 121.36.$$

Example 9: *Calculate the mean, standard deviation and coefficient of variation of the following distribution of the body weights (grams) of a sample of animals:*

Class-interval	*101–105*	*106–110*	*111–115*	*116–120*	*121–125*
Frequency	*6*	*22*	*40*	*25*	*7*

NOTES

Solution:

Calculation of Mean, S.D. and C.V.

Class-interval	Mid-point 'x'	Frequency 'f'	$d = \frac{x - A}{h}$	fd	fd^2
101–105	103	6	$\frac{103 - 113}{5} = -2$	– 12	24
106–110	108	22	$\frac{108 - 113}{5} = -1$	– 22	22
111–115	113	40	$\frac{113 - 113}{5} = 0$	0	0
116–120	118	25	$\frac{118 - 113}{5} = 1$	25	25
121–125	123	7	$\frac{123 - 113}{5} = 2$	14	28
		$N = 100$		$\Sigma fd = 5$	$\Sigma fd^2 = 99$

Let an assumed mean (A) = 113, $h = 5$

Mean $$(\bar{x}) = A + \frac{h}{N}\Sigma fd = 113 + \frac{5}{100}(5)$$

$$\bar{x} = 113 + 0.25 = 113.25$$

Standard deviation $$(\sigma) = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{99}{100} - \left(\frac{5}{100}\right)^2} \times 5$$

$$= \sqrt{0.9875} \times 5 = 0.99 \times 5$$

$$\sigma = 4.95$$

Coefficient of variation (C.V.) $= \frac{\sigma}{\bar{x}} \times 100$

$$= \frac{4.95}{113.25} \times 100$$

$$= 4.37$$

C.V. = 4.37.

5.9 SUMMARY

- "Literal meaning of dispersion is 'Scatteredness'.
- A measure of dispersion describes the degree of scatter shown by the observations and is usually measured as an average deviation about some central value.
- Measures of dispersion may be either absolute or relative. Absolute measure of dispersion cannot be used for comparison purposes if expressed in different units.
- The difference between the largest and smallest value of the variates is called its range.

  Range = L – S.
- Quartile deviation or semi-interquartile range (Q.D.) is given by

$$Q.D. = \frac{Q_3 - Q_1}{2}.$$

- The relative measure of dispersion, known as coefficient of quartile deviation, is calculated as follows:

$$\text{Coefficient of } Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

- The difference between the third quartile and first quartile is known as interquartile range. It is given by

  Interquartile range = $Q_3 - Q_1$.
- Mean deviation (M.D.) is defined as the average of the absolute deviations taken from an average usually, the mean, median or mode. Mean deviation is a measure of dispersion which is based on all values of a set of data.
- "Standard deviation is positive square root of the arithmetic mean of the squares of the deviations taken from arithmetic mean".
- The formula for the standard deviation for an "ungrouped data" is

$$\sqrt{\frac{1}{n}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]}.$$

- The square of standard deviation is called variance and is denoted by σ^2.

  Variance = σ^2

  $\sigma = \sqrt{\text{variance}}$.

NOTES

- "100 times the coefficient of dispersion based upon standard deviation is called coefficient of variation (C.V.)". The formula for coefficient of variation (C.V.) is given by

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100.$$

5.10 GLOSSARY

- ***Quartile deviation:*** Quartile deviation is defined as half the distance between the third and the first quartile.
- ***Mean deviation:*** (M.D.) is defined as the average of the absolute deviations taken from an average usually, the mean, median or mode.
- ***Coefficient of variation:*** "100 times the coefficient of dispersion based upon standard deviation is called coefficient of variation (C.V.)".

5.11 ANSWERS TO CHECK YOUR PROGRESS

1. scatteredness
2. central value
3. absolute, relative
4. range
5. R-charts
1. False
2. True
3. False
4. True
5. True

5.12 TERMINAL AND MODEL QUESTIONS

1. From the following data, calculate interquartile range, quartile deviation, and coefficient of Q.D.

   48, 45, 54, 43, 51, 49, 38, 41, 37, 42, 46.

2. Calculate Quartile deviation and its coefficient from the given data:

Height (cm)	135	136	137	138	139	140	141	142	143	144
No. of students	15	20	32	35	33	22	20	10	8	4

3. The marks obtained by 109 biotechnology students in a biostatistics course are given below. Find out quartile deviation and its coefficient.

Marks	4–8	8–12	12–16	16–20	20–24	24–28	28–32	32–36	36–40
Frequency	6	10	18	30	15	12	10	6	2

4. Calculate the mean deviation and the coefficient of mean deviation from the following data:

   Beetroot weight (gm) = 100, 120, 150, 130, 140, 160, 200.

5. The following are the data pertaining to the number of flower per twig. Calculate the mean deviation from mean and its coefficient for the number of flowers:

No. of flowers per twig (x)	11–15	16–20	21–25	26–30	31–35	36–40	41–45
No. of twig (f)	3	4	11	12	9	7	4

6. Calculate the standard deviation from the data recorded on length of leaves.

   Length of leaves (cm) : 6.5, 6.6, 6.7, 7.0, 7.5, 7.6, 8.0, 9.0, 9.5, 10.0.

7. Calculate the mean, the variance, the standard deviation and the coefficient of variation from the data recorded on respiration rate per minute of 10 persons.

   Respiration/minute : 22, 22, 20, 24, 16, 17, 18, 19, 21, 21.

8. The incubation period of smallpox recorded on 10 patients is given below. Calculate the variance, the standard deviation and the coefficient of variation.

Patients No.	1	2	3	4	5	6	7	8	9	10
Incubation period (x)	10	14	13	11	15	10	9	12	10	16

9. A student obtained the mean and standard deviation of 100 observations as 40 and 5.1 respectively. It was later discovered that he had wrongly copied down an observation as 50 instead of 40. Calculate the correct mean and standard deviation.

10. Ten observations have mean 20 and standard deviation 5. What will be the standard deviations of new series if each item is doubled ?

11. What do you mean by dispersion? What should be the quantities of good measure of dispersion?

NOTES

12. Define standard deviation and its mathematical properties.
13. Differentiate between standard deviation and mean deviation.
14. Define variance and coefficient of variation. State the uses of coefficient of variation in biostatical analysis.

5.13 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

MS 104

Quantitative Techniques in Management

Block - II
Block Title : Measurement of Variation, Correlation and Regression

UTTARAKHAND OPEN UNIVERSITY
SCHOOL OF MANAGEMENT STUDIES AND COMMERCE
University Road, Teenpani By pass, Behind Transport Nagar, Haldwani- 263 139
Phone No: (05946)-261122, 261123, 286055
Toll Free No.: 1800 180 4025
Fax No.: (05946)-264232, e-mail: info@uou.ac.in, som@uou.ac.in
http://www.uou.ac.in
www.blogsomcuou.wordpress.com

Board of Studies

Programme Coordinator

Dr. Manjari Agarwal
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

Units Written By		Unit No.
Text material developed by	Devashish Dutta	
Typeset by	Goswami Associates, Delhi	

Editor(s)

ISBN : 978-93-85740-10-7

Edition : 2016 (Restricted Circulation)
Published by : Uttarakhand Open University, Haldwani, Nainital - 263 139
Printed at : Laxmi Publications (P) Ltd., New Delhi
DUO-8157-87.32-QUAN TECH MGMT B-II

CONTENTS

UNIT 6: MEASURES OF SKEWNESS, KURTOSIS AND MOMENTS

Structure

6.0 INTRODUCTION

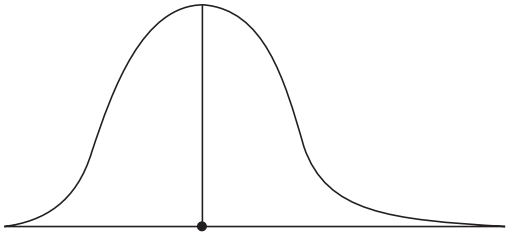
Skewness

"Skewness means lack of symmetry or lopsidedness in a frequency distribution". The object of measuring skewness is to estimate the extent to which a distribution is distorted from a perfectly symmetrical distribution. Skewness indicates whether the curve is turned more to one side than to other, *i.e.*, whether the curve has a longer tail on-one side.

Skewness can be positive as well as negative. **Skewness** is **positive** if the longer tail of the distribution lies towards the right and **negative** if it lies towards the left.

Symmetrical Distribution

A frequency distribution is called symmetric if the frequencies are symmetrically distributed on both sides of the centre point of the frequency curve. *or* A frequency distribution, in which the values of mean, median and mode are equal, is called symmetrical distribution.



Symmetrical distribution
(mean = median = mode)

Fig. 6.1

Skewed Distribution

A frequency distribution which is not symmetrical is called skewed distribution. It is of two types:

(a) Positively skewed

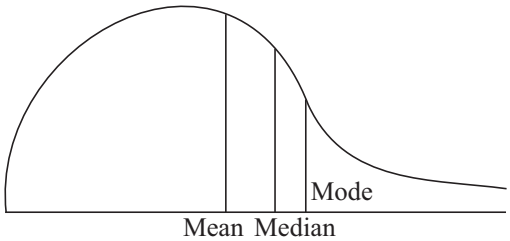
(b) Negatively skewed

Positively Skewed Distribution: A frequency distribution is said to be +ve skewed if the frequency curve, gives a longer tail to the right hand side.

or

In a +ve skewed distribution, the value of mean is maximum and that of mode is least, the median lies between mean and mode, *i.e.,*

Mean > Median > Mode.



+ve skewed distribution

Fig. 6.2

Negatively Skewed Distribution: A frequency distribution is said to be negatively skewed if the frequency curve gives a longer tail on the left hand side. *or* In a –ve skewed distribution, the value of the mode is maximum and that of mean is least, the median lies between mode and mean, *i.e.*,

Mode > Median > Mean

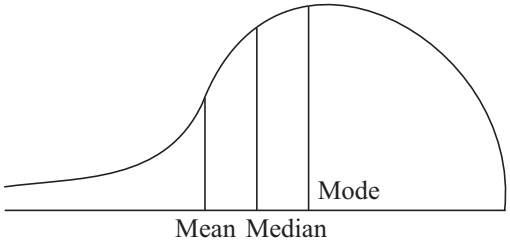


Fig. 6.3

Moderately Symmetrical Distribution

If in a frequency distribution, the interval between the mean and median is approximately one-third of the interval between the mean and mode, then the distribution is called moderately symmetrical distribution. Thus, for a moderate symmetrical distribution,

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

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

or $\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}.$

6.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define the terms skewness and terms related to it
- Explain various tests of skewness
- Explain various types of measures of skewness
- Explain relative measures of skewness
- Define 'Kurtosis'
- Explain measures of kurtosis
- Define 'Moments'
- Explain moments about an arbitrary point
- Explain moments about the origin
- Define central moments
- Derive relation between moments about mean in terms of moments about any point and vice versa

6.2 TESTS OF SKEWNESS

To check, whether a distribution is skewed or not, the following tests can be applied.

A skewed distribution satisfies:

1. The values of mean, median and mode do not coincide.
2. Quartiles are not equidistant from the median.

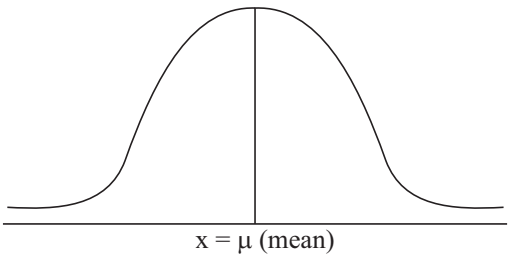


Fig. 6.4

NOTES

3. Frequencies are not symmetrically distributed on both sides of the centre point of the frequency curve.
4. The graph of the given distribution is bell shaped. (see figure)

Example 1: *The following table gives the distribution of income of household.*

Income ₹	% of households
Below 100	*7.2*
100–199	*11.7*
200–299	*12.1*
300–399	*14.8*
400–499	*15.9*
500–599	*14.9*
600–699	*10.4*
700–799	*9*
1000 and above	*4*

(i) *What are the problems in computing standard deviation for the given data ?*

(ii) *Find a suitable measure of dispersion.*

(iii) *If the income of everyone was increased by a certain proportion, will the skewness of above distribution be affected ?*

Solution: (i) Given distribution is an open-end distribution. It means, we need to make an assumption about the lower limit of the first class and the upper limit of the last class.

(ii) The suitable measure of dispersion is the quartile deviation *i.e.,*

$$\text{Q.D} = \frac{Q_3 - Q_1}{2}$$

Calculation of Q_3 and Q_1

(Income ₹)	% of household *f*	Cumulative frequency *c.f.*
Below 100	7.2	7.2
100–199	11.7	18.9
200–299	12.1	31
300–399	14.8	45.8
400–499	15.9	61.7
500–599	14.9	76.6
600–699	10.4	87
700–799	9	96
1000 and above	4	100
	$\Sigma f = 100$	

To find Q_3. Here $\frac{3N}{4} = \frac{3 \times 100}{4} = 75$

i.e., Q_3 lies in 500–599. But the actual limit should be 499.5–599.5

$l = 499.5$, C = 61.7,

$f = 14.9$, $h = 100$

$$\therefore \quad Q_3 = l + \frac{\frac{3N}{4} - C}{f} \times h = 499.5 + \frac{75 - 61.7}{14.9} \times 100$$

$$= 499.5 + 89.26 = 588.76$$

To find Q_1 Here $\frac{N}{4} = \frac{100}{4} = 25$

i.e., Q_1 lies in 200–299. But the actual limit should be 199.5–299.5

$l = 199.5$, C = 18.9,

$f = 12.1$, $h = 100$

$$\therefore \quad Q_1 = l + \frac{\frac{N}{4} - C}{f} \times h$$

$$= 199.5 + \frac{25 - 18.9}{12.1} \times 100$$

$$= 199.5 + 50.41 = 249.91$$

Hence $$Q.D = \frac{Q_3 - Q_1}{2} = \frac{588.76 - 249.91}{2} = 169.425$$

(iii) If the income of every one is increased by a certain proportion, the skewness would not be affected, since it is independent of the charge of origin.

6.3 MEASURES OF SKEWNESS

There are various methods of measuring skewness of a distribution. These measures tell the direction and extent of asymmetry in a distribution. The following diagram classifies the different measures of skewness:

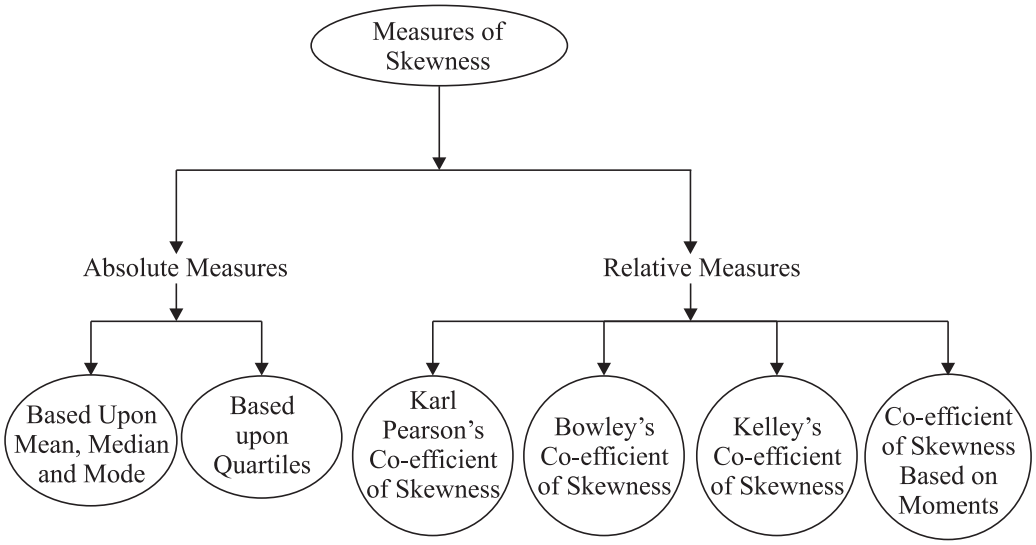


Fig. 6.5

We now discuss in detail the various measures of skewness.

(a) ***Absolute measures of skewness***

The absolute measures of skewness (S_k)

Based upon mean, median and mode are

(a) S_k = Mean – Mode

(b) S_k = 3 (Mean – Median)

(b) ***Based upon quartiles are***

(a) $S_k = Q_3 + Q_1 - 2Q_2$

(b) $S_k = Q_3 + Q_1 - 2$ Median

Limitations of the Absolute Measures of Skewness

The skewness computed by using any one of the above measures is expressed in the unit of value of the distribution, and hence we cannot compare the skewness of another distribution expressed in different units.

6.4 RELATIVE MEASURES OF SKEWNESS

To compare two or more distributions, the absolute measure of skewness cannot be used. To overcome this, we compute **relative measures of skewness,** which are known as "coefficients of skewness". The following are the coefficients of skewness :

(a) Karl Pearson's coefficient of skewness.

(b) Bowley's coefficient of skewness.

(c) Kelley's coefficient of skewness.

(d) Co-efficient of skewness based on moments.

NOTES

Karl Pearson's Coefficient of Skewness (S_{k_p})

The Karl Pearson's coefficient of skewness S_{k_P} is given by happy diwali

$$S_{k_P} = \frac{\text{Mean} - \text{Mode}}{\text{S. D.}}, \text{ where } -1 \le S_{k_P} \le 1.$$

Particular cases:

(i) If $S_{k_P} = 0$, then

$$\frac{\text{Mean} - \text{Mode}}{\text{S. D.}} = 0$$

$\Leftrightarrow$ Mean – Mode = 0

$\Leftrightarrow$ Mean = Mode.

Thus, a distribution is symmetrical iff $S_{k_P} = 0$

(ii) If $S_{k_P} > 0$, then $\frac{\text{Mean} - \text{Mode}}{\text{S. D.}} > 0$

$\Leftrightarrow$ Mean – Mode > 0

$\Leftrightarrow$ Mean > Mode.

Thus, a distribution is positively skewed iff $S_{k_P} > 0$

(iii) Similarly, a distribution is negatively skewed iff $S_{k_P} < 0$.

Limitations of Karl Pearson's Coefficient of Skewness

The Karl Pearson's coefficient of skewness (S_{k_P}) cannot be used when the mode is ill-defined. But for a moderately skewed distribution, mean, median and mode are connected by the following relation :

Mean – Mode = 3 (Mean – Median)

Therefore, for a moderately skewed-distribution,

$$S_{k_P} = \frac{\text{Mean} - \text{Mode}}{\text{S. D.}} = \frac{3\,(\text{Mean} - \text{Median})}{\text{S. D.}},$$

where $-3 \le S_{k_P} \le 3$, for a moderately skewed distribution.

Check Your Progress

Fill in the blanks:

1. A frequency distribution is said to be positively skewed if the frequency curve gives a longer tail to .....
2. If in a frequency distribution, the interval between the mean and median is approximately ..... of the interval between the mean and mode, then the distribution is called moderately symmetrical distribution.
3. In a negatively skewed distribution, the value of mode is ..... and that of mean is .....
4. To check, whether a distribution is skewed or not, the graph of the distribution should be .....
5. To compare two or more distributions, the ..... of skewness cannot be used.

Example 2: *Define skewness and hence compute coefficient of skewness from the values of the variables given below:*

25, 15, 23, 40, 27, 25, 23, 25, 20.

Solution: **Skewness.** Skewness means lack of symmetry or lopsidedness in a frequency distribution.

The coefficent of skewness is given by

$$S_{k_P} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}.$$

$$\text{Mean} = \bar{x}$$

$$= \frac{25 + 15 + 23 + 40 + 27 + 25 + 23 + 25 + 20}{9}$$

$$= \frac{223}{9} = 24.77$$

To find mode, arranging the given data in ascending order, we have

15, 20, 23, 23, 25, 25, 25, 27, 40.

Here maximum frequency is 25. Therefore, Mode = 25.

Also to find S.D. consider the table.

x_i	x_i^2
25	625
15	225
23	529
40	1600
27	729
25	625
23	529
25	625
20	400
223	5887

The standard deviation of the given values is given by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} = \sqrt{\frac{5887}{9} - \left(\frac{223}{9}\right)^2}$$

$$= \sqrt{654.11 - 613.55} = \sqrt{40.56} = 6.36$$

Hence, the required coefficient of skewness (Karl Pearson's coefficient) is given by

$$S_{k_P} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{24.77 - 25}{6.36}$$

$$= -\frac{0.23}{6.36} = -0.036.$$

Example 3: *For a moderate skewed distribution, the A.M. is 200, the coefficient of variation is 8 and Karl's Pearson coefficient of skewness is 0.3. Find the the median and mode.*

Solution: Given mean

$$\bar{x} = 200, \text{ C.V.} = 8, S_{k_P} = 0.3$$

Using $\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100$, we have $8 = \frac{\sigma}{200} \times 100 = \frac{\sigma}{2}$

$\Rightarrow$ $\sigma = 16$

Also $S_{k_P} = \frac{\text{Mean} - \text{Mode}}{\sigma}$ gives $0.3 = \frac{200 - \text{Mode}}{16}$

$\Rightarrow$ $200 - \text{Mode} = 4.8 \Rightarrow \text{Mode} = 200 - 4.8 = 195.2$

Also $\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$, gives 195.2

$= 3 \text{ Median} - 2 \times 200$

NOTES

$$\Rightarrow \quad 3 \text{ Median} = 195.2 + 400 = 595.2$$

$$\Rightarrow \quad \text{Median} = \frac{595.2}{3} = 198.4$$

Example 4: *In a moderate skewed distribution, the mean is 20 and median is 18.5. If the coefficient of variation is 30%, find Karl Pearson's coefficient of skewness of the distribution.*

Solution: Given, Mean = 20, Median = 18.5,

Also C. V. = 30%

$$\Rightarrow \quad \frac{\sigma}{\bar{x}} \times 100 = 30$$

$$\Rightarrow \quad \frac{\sigma}{20} \times 100 = 30$$

$$\Rightarrow \quad \sigma = 6.$$

$$\Rightarrow \quad \text{S.D.} = 6$$

Now, for a moderate skewed distribution

$$S_{k_P} = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}} = \frac{3\,(20 - 18.5)}{6} = 0.75.$$

Example 5: *For a group of 20 items,* $\sum x_i = 200$, $\sum x_i^2 = 5000$ *and median = 15. Find Karl Pearson's coefficient of skewness.*

Solution: Given $n = 20$, $\sum x_i = 200$, $\sum x_i^2 = 5000$

$$\therefore \quad \text{Mean} = \frac{1}{n} \sum x_i = \frac{200}{20} = 10$$

$$\text{S.D.} = \sqrt{\frac{1}{n} \sum x_i^2 - \left(\frac{1}{n} \sum x_i\right)^2} = \sqrt{\frac{5000}{20} - 10^2}$$

$$= \sqrt{250 - 100} = \sqrt{150} = 12.24$$

Therefore, the required Karl's Pearson coefficient of skewness is given by

$$S_{k_P} = \frac{3\,(\text{Mean} - \text{Median})}{\text{S.D.}}$$

$$= \frac{3(10 - 15)}{12.24} = \frac{-15}{12.24} = -1.22$$

$$\Rightarrow \quad S_{k_P} = \mathbf{-1.22.}$$

Example 6: *Calculate Karl Pearson's coefficient of skewness from the data given below.*

x_i :	1	2	3	4	5	6	7
f_i :	10	18	30	25	12	3	2

Solution: The Karl Pearson's coefficient of skewness is given by

$$S_{k_P} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} \qquad ...(1)$$

We first calculate Mean, Mode and S.D. of the given distribution.

Calculation of Mean

Take the assumed mean A = 4, and consider the following table:

x_i	f_i	$u_i = x_i - A$	$f_i u_i$	$f_i u_i^2$
1	10	– 3	– 30	90
2	18	– 2	– 36	72
3	30	– 1	– 30	30
4	25	0	0	0
5	12	1	12	12
6	3	2	6	12
7	2	3	6	18
	$\Sigma f_i = 100$		$\Sigma f_i u_i = -72$	$\Sigma f_i u_i^2 = 234$

The mean $\bar{x}$ is given by $\bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_i u_i = 4 + \frac{1}{100}(-72) = \frac{328}{100} = 3.28$

Calculation of Mode

We know that Mode is that value of the variable which has the maximum frequency. Here the maximum frequency (which is 30) occurs corresponding to the value 3 of the variable x_i. Therefore, Mode = 3.

Calculation of Standard Deviation

We know that the S.D. is given by

$$\sigma = h\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i u_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right)^2}$$

$$= 1 \times \sqrt{\frac{1}{100} \times 234 - \left(\frac{1}{100} \times (-72)^2\right)} = \sqrt{\frac{234}{100} - \frac{5184}{10000}}$$

$$= \sqrt{\frac{23400 - 5184}{10000}} = \sqrt{\frac{18216}{10000}} = \sqrt{1.8216} = 1.3496.$$

NOTES

$\therefore$ From (1), $S_{k_P} = \dfrac{3.28 - 3}{1.3496} = 0.2074$

Example 7: *Compute Karl Pearson's coefficient of skewness from the following data :*

Marks	*0—10*	*10—20*	*20—30*	*30—40*	*40—50*
Frequency	*8*	*11*	*26*	*9*	*6*

Solution: The Karl Pearson's coefficient of skewness is given by

$$S_{k_P} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} \qquad ...(1)$$

We calculate mean, mode and S.D. of the given distribution.

Consider the following table. Take assumed Mean = 25 and $h = 10$.

Marks	**Mid-Values** (x_i)	f_i	$u_i = \dfrac{x_i - A}{h}$	$f_i u_i$	$f_i u_i^2$
0—10	5	8	– 2	– 16	32
10—20	15	11	– 1	– 11	11
20—30	25	26	0	0	0
30—40	35	9	1	9	9
40—50	45	6	2	2	24
		$\Sigma f_i = 60$		$\Sigma f_i u_i = -6$	$\Sigma f_i u_i^2 = 76$

Calculation of Mean

The Mean $\bar{x}$ is given by

$$\bar{x} = A + h\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right) = 25 + 10\left(-\frac{6}{60}\right) = 25 - 1 = 24$$

Calculation of Mode

Here, the maximum frequency is 26, corresponding to the modal class 20–30.

Therefore, $l = 20,\ h = 10, f_m = 26, f_1 = 11$

Hence $$\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

$$= 20 + \frac{26 - 11}{52 - 11 - 9} \times 10 = 20 + \frac{15}{32} \times 10 = 24.69$$

Calculation of Standard Deviation

The standard deviation is given by

$$\sigma = h\left[\sqrt{\frac{1}{N}\sum f_i u_i^2 - \left(\frac{1}{N}\sum f_i u_i\right)^2}\right]$$

$$= 10\left[\sqrt{\frac{76}{60} - \left(\frac{-6}{60}\right)^2} = 10 \times \sqrt{\frac{76}{60} - \frac{36}{3600}}\right.$$

$$= 10 \times \sqrt{\frac{76 \times 60 - 36}{3600}} = 10 \times \sqrt{\frac{4524}{3600}}$$

$$= 10 \times \frac{67.26}{60} = \frac{67.26}{6} = 11.21$$

Hence from (1) the Karl Pearson's coefficient of correlation is given by

$$S_{k_P} = \frac{24 - 24.69}{11.21} = \frac{-0.69}{11.21} = -\ 0.062.$$

Bowley's Coefficient of Skewness

The Bowley's coefficient of skewness (S_{k_B}) is defined by

$$S_{k_B} = \frac{Q_3 + Q_1 - 2\ \text{Median}}{Q_3 - Q_1}$$

$$= \frac{(Q_3 - \text{Median}) + (Q_1 - \text{Median})}{Q_3 - Q_1}.$$

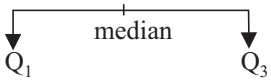
Particular cases:

(i) If the distribution is symmetrical, then Median is equidistant from Q_1 and Q_3, *i.e.*,

$$Q_1 - \text{Median} = \text{Median} - Q_3$$

$$\Leftrightarrow \quad Q_1 + Q_3 - 2\ \text{Median} = 0$$

$$\Leftrightarrow \quad S_{k_B} = 0$$



Hence the distribution is symmetrical iff $S_{k_B} = 0$

(ii) The distribution is positively skewed iff

$$Q_1 - \text{Median} > \text{Median} - Q_3$$

$$\Leftrightarrow \quad Q_3 + Q_1 - 2\ \text{Median} > 0$$

$$\Leftrightarrow \quad S_{k_B} > 0$$

Similarly, the distribution is negatively skewed iff $S_{k_B} < 0$.

NOTES

(iii) The Bowley's coefficient of skewness is also called "quartile measure of skewness" and its values lies between – 1 and 1, *i.e.*, $-1 \le S_{k_B} \le 1$.

Remark: In case of open-end distributions and where the extreme values are present, S_{k_B} is most effective tool to measure skewness.

Example 8: *In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of the upper and lower quartiles is 100 and the median is 38. Find the value of the upper quartile.*

Solution: Given $S_{k_B} = 0.6$

Sum of upper and lower quartile = 100

$$\Rightarrow \quad Q_3 + Q_1 = 100 \qquad ...(1)$$

Also Median = 38

Using $$S_{k_B} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$, we get

$$0.6 = \frac{100 - 76}{Q_3 - Q_1} = \frac{24}{Q_3 - Q_1}$$

$$\Rightarrow \quad Q_3 - Q_1 = \frac{24}{0.6} = 40 \qquad ...(2)$$

Adding (1) and (2), $2Q_3 = 140$

$$\Rightarrow \quad Q_3 = 70.$$

Example 9: (a) *In a distribution, the difference between two quartiles is 15, their sum is 35 and Median is 20. Find the coefficient of skewness.*

(b) *In a frequency distribution, the median is 13, $Q_1 + Q_3 = 25$ and $Q_3 - Q_1 = 10$. Find Bowley's coefficient of skewness.*

(c) *In a frequency distribution, the median is 38, Bowley's coefficient is 0.6 and $Q_1 + Q_3 = 100$. Find Q_1 and Q_3.*

Solution: (a) Given $Q_3 - Q_1 = 15$ ...(1)

$$Q_3 + Q_1 = 35 \qquad ...(2)$$

Median = 20

By def., $$S_{k_B} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

$$= \frac{35 - 70}{15} = -\frac{1}{3} = -0.333.$$

(b) Given $Q_1 + Q_3 = 25$

$Q_3 - Q_1 = 10$

Median = 13

$\therefore$ Bowley's coefficient of skewness is given by

$$S_{k_B} = \frac{Q_3 + Q_1 - 2 \text{ median}}{Q_3 - Q_1} = \frac{25 - 26}{10} = \frac{-1}{10}$$

(c) Given Bowley's coefficient of skewness = 0.6 *i.e.*, $S_{k_B} = 0.6$

$$\Rightarrow \quad \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} = 0.6$$

$$\Rightarrow \quad \frac{100 - 76}{Q_3 - Q_1} = 0.6$$

$$\Rightarrow \quad Q_3 - Q_1 = \frac{24 \times 10}{6} = 40 \qquad ...(1)$$

Also $Q_3 + Q_1 = 100$ ...(2)

Adding (1) and (2) $2Q_3 = 140 \Rightarrow Q_3 = 70$

From (1), $Q_1 = Q_3 - 40 = 70 - 40 = 30.$

Example 10: *Compute the Bowley's coefficient of skewness for the marks obtained by 10 students in a class ; which are given below :*

43, 12, 31, 20, 17, 26, 35, 40, 5, 37.

Solution: The Bowley's coefficient of skewness is given by

$$S_{k_B} = \frac{Q_3 + Q_1 - 2 \text{ (Median)}}{Q_3 - Q_1} \qquad ...(1)$$

We first compute Q_1, Q_3 and Median for the given data.

Computation of Q_1, Q_3

Given $n = 10$, and writing the data, in ascending order, we have

5, 12, 17, 20, 26, 31, 35, 37, 40, 43

$\therefore$ Q_1 = Value of $\left(\frac{n+1}{4}\right)$th observation

= Value of (2.75)th observation

= 2nd observation + (0.75) (3rd observation – 2nd observation)

= 12 + (0.75) (17 – 12) = 12 + 3.75 = 15.75

$$Q_3 = \text{Value of } \frac{3}{4}(n+1)\text{th observation}$$

= Value of 8.25th observation

= 8th observation + (0.25) (9th observation – 8th observation)

= 37 + (0.25) (40 – 37) = 37 + .75 = 37.75

Also Median = Mean of 5th and 6th observation

$$= \frac{26+31}{2} = 28.5 \qquad (\because \quad n = 10 \text{ is even})$$

$\therefore$ From (1), the required Bowley's coefficient of skewness is given by

$$S_{k_B} = \frac{37.75 + 15.75 - 2(28.5)}{37.75 - 15.75} = \frac{-3.5}{22} = -0.158.$$

Example 11: *Compute Bowley's coeficient of skewness from the data*

x_i :	5	15	25	35	45
f_i :	10	20	40	20	10

Solution: We know that Bowley's coeffficient of skewness is given by

$$S_{k_B} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1} \qquad ...(1)$$

We first calculate Q_1, Q_3 and Median.

Consider the following table:

x_i	f_i	c.f.
5	10	10
15	20	30
25	40	70
35	20	90
45	10	100
	$\Sigma f_i = 100$	

Computation of Q_1 (lower quartile).

Here, $$N = \sum_{i=1}^{n} f_i = 100$$

$\therefore$ $$\frac{N}{4} = \frac{100}{4} = 25$$

The cumulative frequency just greater than 25 is 30 and the corresponding value of x is 15.

$\therefore \quad Q_1 = 15$

Computation of Median

NOTES

Here, $\frac{N}{2} = \frac{100}{2} = 50$

The cumulative frequency just greater than 50 is 70 and the corresponding value of x is 25.

$\therefore \quad$ Median = 25

Computation of Q_3 (upper quartile)

Here, $\frac{3N}{4} = \frac{3 \times 100}{4} = 75$

The cumulative frequency just greater than 75 is 90 and the corresponding value of x is 35.

$\therefore \quad Q_3 = 35$

Hence, from (1), $S_{k_B} = \frac{35 + 15 - 2(25)}{35 - 15} = \frac{50 - 50}{20} = 0$

$\Rightarrow S_{k_B} = 0$, *i.e.*, the distribution is **symmetric.**

Measure of Skewness Based Upon Moments

Karl Pearson defined the following measures of skewness based upon moments, which are

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} \; ; \gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

(i) If $\gamma_1 = 0$, then the distribution is symmetrical.

(ii) If $\gamma_1 > 0$, then the distribution is positively skewed.

(iii) If $\gamma_1 < 0$, then the distribution is negatively skewed.

Example 12: *The first four moments about the mean are 0, 2, 0.7 and 18.75. Find the coefficient of skewness based upon these moments.*

Solution: Let μ_1, μ_2, μ_3, μ_4 be the moments about the mean. Then given,

$\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 0.7$

$\mu_4 = 18.75$

The required coefficient of skewness based upon these moments is given by

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{0.7}{\sqrt{8}} = \frac{0.7}{2.828} = \mathbf{0.247.}$$

NOTES

Remark: In the above example, $\gamma_1 = 0.24770 > 0$, it means the distribution is +ve skewed.

Example 13: *If the first three moments of a distribution about the value 2 are 1, 22 and 100. Find the moment measure of skewness.*

Solution: Let μ_1', μ_2', μ_3' are the moments about the value 2, then given

$$\mu_1' = 1, \mu_2' = 22, \mu_3' = 100.$$

Now, $\mu_2 = \mu_2' - \mu_1'^2 = 22 - 1 = 21$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 100 - 3 \times 22 \times 1 + 2 \times 1$$
$$= 100 - 66 + 2 = 36$$

The required moment measure of skewness is given by

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{36}{\sqrt{9261}} = \frac{36}{94.23} = 0.3820$$

$\Rightarrow$ $\gamma_1 = 0.3820.$

Remark: In the above example, $\lambda_1 = 0.3820 > 0$, it means the distribution is positively skewed.

Example 14: *The value of mean, variance and γ_1 for a distribution are 10, 16 and 1 respectively. Find the first three moments about the origin and discuss the nature of skewness.*

Solution: Let v_1, v_2, v_3 be the first three moments about the origin, then by def,

$$v_1 = \text{Mean} = 10$$
$$v_2 = \mu_2 + v_1^2 = \text{variance} + v_1^2 = 16 + 100 = 116$$

Also, $\gamma_1 = 1 \Rightarrow \sqrt{\beta_1} = 1 \Rightarrow \beta_1 = 1$

$$\Rightarrow \frac{\mu_3^2}{\mu_2^3} = 1 \Rightarrow \mu_3^2 = \mu_2^3 = 16^3 \quad | \ \mu_2 = \text{variance}$$

$$\Rightarrow \mu_3 = 16^{3/2} = 64$$

Hence, $v_3 = \mu_3 + 3v_2v_1 - 2v_1^3 = 64 + 3 \times 116 \times 10 - 2 \times 1000$

$$= 64 + 3480 - 2000 = 1544$$

As $\gamma_1 = 1 > 0 \Rightarrow \gamma_1 > 0$, it means, the distribution is positively skewed.

Example 15: *In two distributions A and B, the second central moments are 9 and 16 respectively and the third central moments are – 8.1 and – 12.8 respectively. Which distribution is less skewed to the left ?*

Solution: The extent of skewness can be well determined by their coefficient of skewness.

For the distribution A, Given

$$\mu_2 = 9, \mu_3 = -8.1$$

$$\therefore \quad \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{-8.1}{\sqrt{729}} = \frac{-8.1}{27} = -0.3$$

For the distribution B, Given $\mu_2 = 16$, $\mu_3 = -12.8$

$$\therefore \quad \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{-12.8}{\sqrt{4096}} = \frac{-12.8}{64} = -0.2$$

Here, for the distribution A, $\gamma_1 = -0.3$ and for the distribution B $\gamma_1 = -0.2$.

Clearly, the distribution B is less skewed to the left.

Kelley's Co-efficient of Skewness

Kelley's coefficient of skewness are based upon deciles and percentiles.

(i) ***Based Upon Deciles.*** The kelley's coefficient of skewness, based upon deciles is given by

$$S_{k_k} = \frac{D_1 + D_9 - 2 \text{ Median}}{D_9 - D_1}$$

(ii) ***Based Upon Percentiles.*** The Kelley's coefficient of skewness, based upon percentiles is given by

$$S_{k_k} = \frac{P_{10} + P_{90} - 2 \text{ Median}}{P_{90} - P_{10}}.$$

Example 16: *The marks obtained by 15 students of M.B.A first year are*

10, 60, 36, 20, 5, 8, 19, 28, 16, 9, 46, 53, 14, 32, 40.

Find Kelley's coefficient of skewness based upon percentiles

Solution: Arranging the marks in ascending order, we get

5, 8, 9, 10, 14, 16, 19, 20, 28, 32, 36, 40, 46, 53, 60.

Also $n = 15$

$$\therefore \quad P_{10} = \text{Value of } \frac{10}{100} (n + 1)\text{th observation}$$

= Value of 1.6th observation

= Value of first observation + (0.6) (2nd observation – first observation)

= 5 + (0.6) (9 – 8) = 5 + 0.6 = 5.6

$$P_{50} = \text{Value of } \frac{50}{100} (n + 1)\text{th observation}$$

= Value of 8th observation = 20

$$P_{90} = \text{Value of } \frac{90}{100}\ (n+1)\text{th observation}$$

$$= \text{Value of } 14.4\text{th observation}$$

$$= 14\text{th observation} + (0.4)\ (15\text{th observation} - 14\text{th observation})$$

$$= 53 + (0.4)\ (60 - 53) = 53 + 2.8 = 55.8$$

The required Kelley's coefficient of skewness based upon percentiles is given by

$$S_{k_k} = \frac{P_{10} + P_{90} - 2\ \text{Median}}{P_{90} - P_{10}}$$

$$= \frac{5.6 + 55.8 - 2\ (20)}{55.8 - 5.6} \qquad \mid P_{50} = \text{Median}$$

$$= \frac{61.4 - 40}{50.2} = \frac{21.4}{50.2} = 0.42.$$

Example 17: *Find Kelley's coefficient of skewness based upon percentiles for the given data:*

Marks	*0–5*	*5–10*	*10–15*	*15–20*	*20–25*	*25–30*	*30–35*	*35–40*	*40–45*	*45–50*	*50–55*	*55–60*
Freq.	*4*	*9*	*32*	*108*	*127*	*176*	*158*	*185*	*130*	*48*	*20*	*3*

Solution: Consider the following table:

Marks	**Frequency**	***c.f.***
0—5	4	4
5—10	9	13
10—15	32	45
15—20	108	153
20—25	127	280
25—30	176	456
30—35	158	614
35—40	185	799
40—45	130	929
45—50	48	977
50—55	20	997
55—60	3	1000

To find P_{10}. Here $N = 1000$

$$\therefore \qquad \frac{10N}{100} = \frac{10 \times 1000}{100} = 100$$

From the table, we observe that the cumulative frequency just greater than $\frac{10N}{100} =$ 100 is 153 and the corresponding percentiles class is 15—20.

$\therefore$ $l = 15, f = 108, C = 45, h = 5$

Now $$P_{10} = l + \frac{\frac{10N}{100} - C}{f} \times h = 15 + \frac{100 - 45}{108} \times 5 = 15 + \frac{275}{108}$$

$$= 15 + 2.54 = 17.54$$

To find P_{50}, *i.e.*, Median

Here $$\frac{N}{2} = \frac{1000}{2} = 500$$

$\therefore$ The cumulative frequency just greater than $\frac{N}{2} = 500$ is 614 and the corresponding median class is 30—35. Therefore, $l = 30, f = 158, C = 456, h = 5$

Hence, $$P_{50} = \text{Median} = l + \frac{\frac{N}{2} - C}{f} \times h$$

$$= 30 + \frac{500 + 456}{158} \times 5 = 30 + 1.39 = 31.39$$

To find P_{90}. Here, $\frac{90N}{100} = \frac{90}{100} \times 1000 = 900$

i.e., the cumulative frequency just greater than $\frac{90N}{100} = 900$ is 929 and hence, the 90th percentile class is 40—45.

$\therefore$ $l = 40, f = 130, C = 799, h = 5$

Hence $$P_{90} = l + \frac{\frac{90N}{100} - C}{f} \times h = 40 + \frac{900 - 929}{130} \times 5$$

$$= 40 + 3.88 = 43.88$$

Hence, the required Kelley's coefficient of skewness based upon percentiles is given by

$$S_{k_k} = \frac{P_{90} + P_{10} - 2 \text{ Median}}{P_{90} - P_{10}} = \frac{43.88 + 17.54 - 2\,(31.39)}{43.88 - 17.54}$$

$$= \frac{-1.36}{26.34} = \mathbf{-0.0516.}$$

NOTES

6.5 DEFINING THE TERM 'KURTOSIS'

Given two frequency distributions, which have the same variability as measured by the standard deviation, they may be relatively more or less flat topped than the "Normal curve". A frequency curve may be symmetrical but it may not be equally flat topped with the Normal curve. The relative flatness of the top is called **Kurtosis** or convexity of the frequency curve.

Kurtosis enables us to have an idea about the "flatness or peakedness" of the frequency curve.

Normal Curves or Mesokurtic Curves

The curves which are neither flat, nor sharply peaked, are known as Normal curves or mesokurtic curves (see curve A in figure given below).

Platykurtic Curves

The curves which are flatter than the Normal curves, are known as platykurtic curves (see curve B in figure given below).

Leptokurtic Curves

The curves which are more sharply peaked than the normal curve, are known as leptokurtic curves (see curve C in the figure given below).

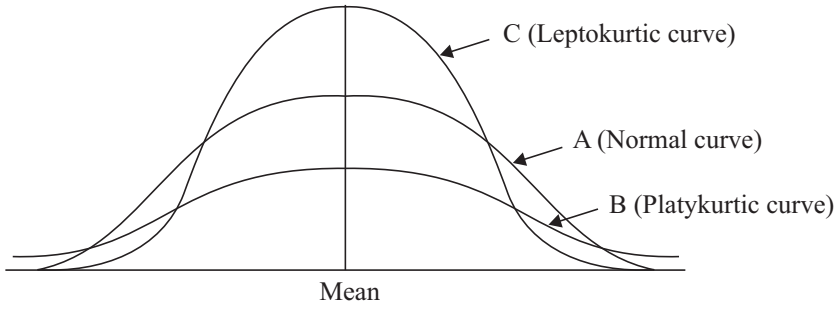


Fig. 6.6

6.6 MEASURES OF KURTOSIS

(1) The kurtosis of a frequency curve is measured by the value of the Pearsonian coefficient β_2, given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

(2) Another measure of kurtosis is γ_2, which is given by $\gamma_2 = \beta_2 - 3$

Remark:

(i) For Normal curve or mesokurtic curve, $\beta_2 = 3$, $\gamma_2 = 0$

(ii) For platykurtic curve, $\beta_2 < 3$ or $\gamma_2 < 0$

(iii) For leptokurtic curve, $\beta_2 > 3$ or $\gamma_2 > 0$

Theorem: *For a symmetrical distribution, all the moments of odd order about the mean vanish.*

Proof.: Let $\bar{x}$ denote the mean of the variate x, then

$$\mu_{2r+1} = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^{2r+1}, \; N = \Sigma f_i$$

In a symmetrical distribution, the values of the variate equidistant from the mean have equal frequencies.

$\therefore \quad f_1 (x_1 - \bar{x})^{2r+1} + f_n(x_n - \bar{x})^{2r+1} = 0$

[$\because$ $x_1 - \bar{x}$ and $x_n - \bar{x}$ are equal in magnitude but opposite in sign. Also $f_1 = f_n$]

Similarly, $f_2(x_2 - \bar{x})^{2r+1} + f_{n-1}(x_{n-1} - \bar{x})^{2r+1} = 0$ and so on.

$\therefore$ If n is even, all the terms in $\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^{2r+1}$ cancel in pairs. If n is odd, again the terms cancel in pairs and the middle term vanishes, since middle term $= \bar{x}$.

Hence $\mu_{2r+1} = 0$

In particular $\mu_3 = 0$ and hence $\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}} = 0.$

Thus, β_1 gives a measure of departure from symmetry, *i.e.*, of skewness.

Remark: The above result does not hold for asymmetrical distribution.

Check Your Progress

State whether the following statements are True or False:

6. The distribution is symmetrical iff Bowley's coefficient of skewness $(S_{k_b}) > 0$.
7. The value of Bowley's coefficient of skewness lies between – 1 and 1.
8. Kelley's coefficient of skewness are based upon quartile and decile.
9. The curves which are more sharply peaked than the normal curve are called leptokurtic curve.
10. For a symmetrical distribution, all the moments of odd order about the mean vanish.

Example 18: *Calculate the first four moments of the following distribution about the mean and hence find* β_1 *and* β_2*:*

x :	0	1	2	3	4	5	6	7	8
f :	1	8	28	56	70	56	28	8	1

Also discuss the nature of the distribution.

Solution: Let us first calculate moments about $x = 4$.

$$\mu_r' = \frac{1}{N} \Sigma f(x-4)^r = \frac{1}{N} \Sigma f d^r, \text{ where } d = x - 4$$

x	*f*	*d = x – 4*	*fd*	fd^2	fd^3	fd^4
0	1	– 4	– 4	16	– 64	256
1	8	– 3	– 24	72	– 216	648
2	28	– 2	– 56	112	– 224	448
3	56	– 1	– 56	56	– 56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
	N = 256		0	512	0	2816

$$\mu_1' = \frac{1}{N} \Sigma fd = 0 ; \qquad \mu_2' = \frac{1}{N} \Sigma fd^2 = \frac{512}{256} = 2$$

$$\mu_3' = \frac{1}{N} \Sigma fd^3 = 0 ; \qquad \mu_4' = \frac{1}{N} \Sigma fd^4 = \frac{2816}{256} = 11$$

Moments about mean are

$$\mu_1 = 0 \text{ (}always\text{)} ; \qquad \mu_2 = \mu_2' - \mu_1'^2 = 2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0 ; \quad \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 11$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 ; \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75.$$

Here, $\beta_2 = 2.75$, it means, the distribution is symmetrical.

Also $\gamma_2 = \beta_2 - 3 = 2.75 - 3 = -0.25$.

It means, the distribution is platykurtic.

Example 19: *The second, third and the fourth central moments of a distribution are 2, 0.6 and 18.25 respectively. Test the skewness and kurtosis of the distribution.*

Solution: Given $\mu_2 = 2$, $\mu_3 = 0.6$, $\mu_4 = 18.25$

For skewness, $\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{0.6}{\sqrt{8}} = 0.21 > 0$

$\Rightarrow \quad \gamma_1 > 0$. Hence, the distribution is +ve skewed.

For kurtosis, $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.25}{4} = 4.5673$

$\Rightarrow \quad \beta_2 > 3$. Hence, the distribution is leptokurtic.

Example 20: *The standard deviation of a symmetrical distribution is 3. Find the value of the fourth moment about the mean for which the distribution becomes mesokurtic, leptokurtic and platykurtic.*

Solution: Given S.D. = 3 $\Rightarrow$ variance = 9

$\Rightarrow \quad \mu_2 = 9$

(i) For the distribution to be mesokurtic, we know

$$\beta_2 = 3 \quad \Rightarrow \quad \frac{\mu_4}{\mu_2^2} = 3$$

$$\Rightarrow \quad \mu_4 = 3 \times 9^2 = 243$$

Hence, the fourth moment about the mean of the distribution must be 243.

(ii) For the distribution to be leptokurtic, we know

$$\beta_2 > 3 \quad \Rightarrow \quad \mu_4 > 243$$

(iii) For the distribution to be platykurtic, we know

$$\beta_2 < 3 \quad \Rightarrow \quad \mu_4 < 243$$

Example 21: *Find the kurtosis of the following data :*

Marks	*0–10*	*10–20*	*20–30*	*30–40*	*40–50*
Freq.	*10*	*20*	*40*	*20*	*10*

Solution: Take assumed mean A = 25, h = 10. Consider the following table.

Marks	**Mid-value x_i**	f	$u_i = \frac{x_i - A}{h}$	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
0—10	5	10	– 2	–20	40	– 80	160
10—20	15	20	– 1	– 20	20	– 20	20
20—30	25	40	0	0	0	0	0
30—40	35	20	1	20	20	20	20
40—50	45	10	2	40	40	80	160
	Total	$\Sigma f = 100$		$\Sigma f_i u_i = 0$	$\Sigma f_i u_i^2 = 120$	$\Sigma f_i u_i^3 = 0$	$\Sigma f_i u_i^4 = 360$

NOTES

NOTES

We wish to find the value of β_2, where

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} \text{ ; where } \mu_4, \mu_2 \text{ are the moments about the mean.}$$

We first find $\mu_1', \mu_2', \mu_3', \mu_4'$.

Here, $$\mu_1' = h\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right) = 10 \times \left(\frac{1}{100} \times 0\right) = 0$$

$$\mu_2' = h^2\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i^{\,2}\right) = 100\left(\frac{1}{100} \times 120\right) = 120$$

$$\mu_3' = h^3\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i^{\,3}\right) = 1000\left(\frac{1}{100} \times 0\right) = 0$$

$$\mu_4' = h^4\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i^{\,4}\right) = 10000\left(\frac{1}{100} \times 360\right) = 36000$$

$\therefore$ $$\mu_2 = \mu_2' - \mu_1'^2 = 120 - 0 = 120$$

$$\mu_4 = \mu_4' - 4\mu_3' + 6\mu_2' \mu_1'^2 - 3\mu_1'^4$$

$$= 36000 - 0 + 0 - 0 = 36000$$

Hence, $$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{36000}{120^2} = \frac{36000}{14400} = \frac{360}{144} = 2.5 < 3$$

$\Rightarrow$ The distribution is **platykurtic.**

6.7 DEFINING 'MOMENTS'

The term "moment" in statistics gives the same meaning as in statics. In statics, the 'moment' refers to the tendency of a force to rotate a body about a point.

6.8 MOMENTS ABOUT AN ARBITRARY POINT

Let x : x_1 x_2 x_3 ..... x_n

f : f_1 f_2 f_3 ..... f_n, be a frequency distribution.

Then r^{th} moment μ_r' about the point $x = A$ is defined by

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r, \text{ where } N = \sum_{i=1}^{n} f_i \text{ ; } r = 0, 1, 2, ...$$

Particular cases. Put $r = 0, 1, 2, 3, 4$, in above, we get

$$\mu_0' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = \frac{N}{N} = 1$$

$$\Rightarrow \qquad \boxed{\mu_0' = 1}$$

Also
$$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^1 = \frac{1}{N}\sum_{i=1}^{n} (f_i x_i - f_i A)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \frac{A}{N}\sum_{i=1}^{n} f_i = \overline{X} - \frac{A}{N} XN = \overline{X} - A$$

$$\Rightarrow \qquad \boxed{\mu_1' = \overline{X} - A}$$

Similarly
$$\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^2$$

$$\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^3$$

$$\mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^4$$

6.9 MOMENTS ABOUT THE ORIGIN

Let x : x_1 x_2 x_3 ..... x_n
f : f_1 f_2 f_3 ..... f_n, be a frequency distribution.
Then r^{th} moment about the origin, *i.e.,* about the point $x = 0$, is defined as

$$v_r = \mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^r \text{, where } N = \sum_{i=1}^{n} f_i, \quad r = 0, 1, 2, ...$$

Particular cases. Put $r = 0, 1, 2, 3, 4$, in above, we get

$$v_0 = \mu_0' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = \frac{N}{N} = 1$$

$$\Rightarrow \qquad v_0 = 1$$

For $r = 1$,
$$v_1 = \mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \overline{X}$$

Also
$$v_2 = \mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2$$

NOTES

$$v_3 = \mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^3$$

$$v_4 = \mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^4$$

or $$v_1 = \mu_1' + A \qquad | \ \mu_1' = \overline{X} - A$$

$$v_2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X} + \overline{X})^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})^2 + \frac{1}{N}\sum_{i=1}^{n} f_i \overline{X}^2 + \frac{2}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})\overline{X}$$

$$= \mu_2 + \overline{X}^2 \left(\frac{1}{N}\sum_{i=1}^{n} f_i\right) + 0 = \mu_2 + \overline{X}^2$$

$\Rightarrow$ $$v_2 = \mu_2 + v_1^2$$

Similarly, $$v_3 = \mu_3 + 3v_2 v_1 - 2v_1^3$$

$$v_4 = \mu_4 + 4v_3 v_1 - 6v_2 v_1^2 + 3v_1^4 \text{ and so on}$$

6.10 CENTRAL MOMENTS (MOMENTS ABOUT MEAN)

Let x : x_1 x_2 ..... x_n

f : f_1 f_2 ..... f_n, be a frequency distribution.

Then r^{th} moments μ_r about the mean $\overline{X}$, is defined by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})^r, \text{ where } N = \sum_{i=1}^{n} f_i \ ; \ r = 0, 1, 2, ...$$

and $$\overline{X} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

Particular cases. Put $r = 0, 1, 2, 3, 4$, in above we get

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - X)^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = \frac{N}{N} = 1$$

$\Rightarrow$ $$\mu_0 = 1$$

For $r = 1$, $$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})^1 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \frac{1}{N}\sum_{i=1}^{n} f_i \overline{X}$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \overline{X}\left(\frac{1}{N}\cdot\sum_{i=1}^{n} f_i\right)$$

$$= \overline{X} - \overline{X} = 0 \qquad \left|\sum_{i=1}^{n} f_i = N\right.$$

$$\Rightarrow \qquad \mu_1 = 0$$

Also, $$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_1 - \overline{X})^2$$

$$= \text{Variance (X)} = \sigma^2$$

Further, $$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})^3$$

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})^4 \text{ and so on}$$

Remarks: (i) The moments about the mean are known as central moments.

(ii) The results $\mu_0 = 1$, $\mu_1 = 0$ and $\mu_2 = \sigma^2$, are of fundamental importance and should be committed to memory.

(iii) In case of frequency distribution with class intervals, the values of X are the mid-points of the class-intervals.

6.11 RELATION BETWEEN MOMENTS ABOUT MEAN IN TERMS OF MOMENTS ABOUT ANY POINT AND VICE VERSA

By definition, $$\mu_r' = \frac{1}{N}\Sigma f(x - A)^r, \quad \text{where A is any point}$$

$$= \frac{1}{N}\Sigma f d^r, \quad \text{where } d = x - A \qquad ...(i)$$

Putting $r = 1$, $$\mu_1' = \frac{1}{N}\Sigma f d$$

$\therefore$ A.M. $$M = A + \frac{1}{N}\Sigma f d = A + \mu_1'$$

NOTES

or $\mu_1' = M - A$ ...(*ii*)

Now $\mu_r' = \frac{1}{N} \Sigma f(x - M)^r$

$= \frac{1}{N} \Sigma f(x - A + A - M)^r = \frac{1}{N} \Sigma f(d - \mu_1')^r$ | Using (*ii*)

$= \frac{1}{N} \Sigma f [d^r - {}^rC_1 d^{r-1}\mu_1' + {}^rC_2 d^{r-2}\mu_1'^2 - {}^rC_3 d^{r-3} \mu_1'^3 + ...$

$+ (-1)^r . \mu_1'^r]$

$= \frac{1}{N} \Sigma f d^r - {}^rC_1\mu_1' . \frac{1}{N} \Sigma f d^{r-1} + {}^rC_2\mu_1'^2 \frac{1}{N} \Sigma f d^{r-2}$

$- {}^rC_3\mu'^3 \frac{1}{N} \Sigma f d^{r-3} + ... + (-1)^r\mu_1'^r . \frac{1}{N} \Sigma f$

$= \mu_r' - {}^rC_1\mu'_{r-1} \mu'_1 + {}^rC_2\mu'_{r-2} \mu_1'^2 - {}^rC_3\mu'_{r-3} \mu'^3_1 + ...$

$+ (-1)^r\mu_1'^r$

| Using (*i*)

In particular, putting $r = 2, 3, 4$, we get

$\mu_2 = \mu_2' - 2\mu_1'^2 + \mu_0' \mu_1'^2 = \mu_2' - \mu_1'^2$ | $\because \quad \mu_0' = 1$

$\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 3\mu_2'^3 - \mu_0' \mu_1'^3 = \mu_3' - 3\mu_2' \mu_1' + 2\mu_1'^3$

$\mu_4 = \mu_4' - 4\mu_3' \mu_1' + 6\mu_2' \mu_1'^2 - 4\mu_1' \mu_1'^3 + \mu_0' \mu_1'^4$

$= \mu_4' - 4\mu_3' \mu_1' + 6\mu_2' \mu_1'^2 - 3\mu_1'^4$

Hence $\mu_1 = 0$

$\mu_2 = \mu_2' - \mu_1'^2$

$\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 2\mu_3'^3$ $(\mu_1' = M - A)$

$\mu_4 = \mu_4' - 4\mu_3' \mu_1' + 6\mu_2' \mu_1'^2 - 3\mu_1'^4$

By using these formulae, we can find moments about mean, when moments about any point are given.

Conversely, $\mu_r = \frac{1}{N} \Sigma f(x - M)^r = \frac{1}{N} \Sigma f d^r$, where $d = x - M$ ...(*iii*)

Now $\mu_r' = \frac{1}{N} \Sigma f(x - A)^r = \frac{1}{N} \Sigma f(x - M + M - A)^r = \frac{1}{N} \Sigma f(d + \mu_1')^r$

| Using (*ii*)

$= \frac{1}{N} \Sigma f(d^r + {}^rC_1 d^{r-1} \mu_1' + {}^rC_2 d^{r-2} \mu_1'^2 + {}^rC_3 d^{r-3} \mu_1'^3 + ...$

$+ \mu_1'^r)$

$$= \frac{1}{N}\Sigma fd^{r} + {}^{r}C_1\mu_1' \cdot \frac{1}{N}\Sigma fd^{r-1} + {}^{r}C_2\mu_1'^{2}\frac{1}{N}\Sigma fd^{r-2} + ...$$

$$+ \mu_1'^{r} \cdot \frac{1}{N}\Sigma f$$

NOTES

$$= \mu_r + {}^{r}C_1\mu_{r-1}\mu_1' + {}^{r}C_2\mu_{r-2}\mu_1'^{2} + ... + \mu_1'^{r} \quad | \text{ Using } (iii)$$

In particular, putting $r = 2, 3, 4$ and noting that $\mu_1 = 0$, $\mu_0 = 1$, we get

$$\mu_2' = \mu_2 + 2\mu_1\mu_1' + \mu_0\mu_1'^{2} = \mu_2 + \mu_1'^{2}$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + 3\mu_1\mu_1'^{2} + \mu_0\mu_1'^{3} = \mu_3 + 3\mu_2\mu_1' + \mu_1'^{3}$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^{2} + 4\mu_1\mu_1'^{3} + \mu_0\mu_1'^{4}$$

$$= \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^{2} + \mu_1'^{4}.$$

By using these formulae, we can find moments about any point, when moments about mean are given.

Effect of Change of Origin and Scale on Moments

Let $u = \dfrac{x - A}{h}$, *i.e.*, $x = A + hu$

$\therefore$ $\bar{x} = A + h\bar{u}$, where bar denotes the mean of the respective variable.

$\therefore$ $x - \bar{x} = h(u - \bar{u})$

$$\mu_r' = \frac{1}{N}\Sigma f(x - A)^r = \frac{1}{N}\Sigma fh^r u^r = h^r \cdot \frac{1}{N}\Sigma fu^r$$

Also $$\mu_r = \frac{1}{N}\Sigma f(x - \bar{x})^r = \frac{1}{N}\Sigma fh^r(u - \bar{u})^r = h^r \cdot \frac{1}{N}\Sigma f(u - \bar{u})^r$$

Hence, the rth moment of the variable x is h^r times the corresponding moment of the variable u.

Sheppard's Corrections for Moments

In the case of class intervals we assume that the frequencies are concentrated at mid-points of class intervals. Since this assumption is not true in general, some error is likely to creep into the calculation of moments. W.F. Sheppard gave the following formulae by which these errors may be corrected :

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{1}{12}h^2 ; \quad \mu_3 \text{ (corrected)} = \mu_3$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h^4, \text{ where } h \text{ is the width of class intervals.}$$

Charlier's Check

To check the accuracy in the calculation of first four moments, we often use the following identities known as *Charlier checks* :

$$\Sigma f(x+1) = \Sigma fx + \Sigma f = \Sigma fx + N$$

$$\Sigma f(x+1)^2 = \Sigma fx^2 + 2\Sigma fx + N$$

$$\Sigma f(x+1)^3 = \Sigma fx^3 + 3\Sigma fx^2 + 3\Sigma fx + N$$

$$\Sigma f(x+1)^4 = \Sigma fx^4 + 4\Sigma fx^3 + 6\Sigma fx^2 + 4\Sigma fx + N.$$

Pearson's β and γ Co-efficients

Karl Pearson defined the following four co-efficients based upon the first four moments about mean.

$$\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}}, \gamma_1 = +\sqrt{\beta_1} \;; \qquad \beta_2 = \frac{\mu_4}{\mu_2^{\,2}}, \gamma_2 = \beta_2 - 3$$

These co-efficients are independent of units of measurement and therefore are pure numbers.

Based upon moments, co-efficient of skewness is $S_k = \dfrac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$.

Example 22: *The first three moments of a distribution about the value 2 of the variable are 1, 16 and – 40. Show that mean = 3, variance = 15 and $\mu_3 = -86$.*

Solution: Let x_i/f_i ; $i = 1, 2, 3, ..., n$, be the given frequency distribution. Let μ_1', μ_2', μ_3' be the given moments about the point $x = 2$.

Then, as per given $\mu_1' = 1, \mu_2' = 16$

$$\mu_3' = -40$$

But $\mu_1' = \overline{X} - A$. Here, $A = 2$

$$\therefore \qquad 1 = \overline{X} - 2 \quad \Rightarrow \quad \overline{X} = 3$$

Also, to find variance, we find μ_2

Now $\mu_2 = \mu_2' - \mu_1'^2 = 16 - 1 = 15$

$\Rightarrow$ variance = 15

Lastly $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -40 - 48 + 2 = -86.$

Example 23: *The first four moments of a distribution about x = 2 are 1, 2.5, 5.5 and 16. Calculate the first four moments about the mean and about zero.*

Solution: Let x_i/f_i ; $i = 1, 2, ..., n$ be the given frequency distribution. Let μ_1', μ_2', μ_3', μ_4' be the first four moments about $x = 2$,

Then, as per given, $\mu_1' = 1, \mu_2' = 2.5$

$$\mu_3' = 5.5, \mu_4' = 16, A = 2$$

To find the first four moments about mean, *i.e.*, To find $\mu_1, \mu_2, \mu_3, \mu_4$.

By def. $\mu_1 = 0$

$$\mu_2 = \mu_2' - \mu_1'^2 = 2.5 - 1 = 1.5$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 5.5 - 3\,(2.5) + 2\,(1) = 5.5 - 7.5 + 2 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\,\mu_1' + 6\mu_2'\,\mu_1'^2 - 3\mu_1'^4$$

$$= 16 - 4 \times 5.5 \times 1 + 6 \times 2.5 \times 1 - 3 \times 1$$

$$= 16 - 22 + 15 - 3 = 6$$

Hence, $\mu_1 = 0, \mu_2 = 1.5, \mu_3 = 0, \mu_4 = 6$

Now to find the moments about $x = 0$. Let v_1, v_2, v_3, v_4 be the required moments about zero.

Then

$$v_1 = \mu_1' + A = 1 + 2 = 3$$

$$v_2 = \mu_2 + v_1^2 = 1.5 + 9 = 10.5$$

$$v_3 = \mu_3 + 3v_2v_1 - 2v_1^3 = 0 + 3 \times 10.5 \times 3 - 2 \times 3^3 = 40.5$$

$$v_4 = \mu_4 + 4v_3\,v_1 - 6v_2v_1^2 + 3v_1^4$$

$$= 6 + 4 \times 40.5 \times 3 - 6 \times 10.5 \times 9 + 3 \times 81 = 168.$$

Example 24: *The first four moments of a distribution about the value 4 of the variable are – 1.5, 17, – 30 and 108. Find the moments about mean,* β_1 *and* β_2.

Solution: Let $\mu_1', \mu_2', \mu_3', \mu_4'$ be the first four moments of the given distribution, then

$$A = 4, \mu_1' = -1.5, \mu_2' = 17, \mu_3' = -30, \mu_4' = 1.8$$

Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the first four moments about the mean, then by def., $\mu_1 = 0$ (True for every distribution).

$$\mu_2 = \mu_2' - \mu_1'^2 = 17 - (-1.5)^2$$

$$= 17 - 2.25 = 14.75$$

$$\mu_3 = \mu_3' - 3\mu_2'\,\mu_1' + 2\mu_1'^3$$

$$= -30 - 3 \times 17 \times 1.5 + 2 \times (1.5)^3 = -30 - 76.5 - 6.75$$

$$= -113.25$$

Also

$$\mu_4 = \mu_4' - 4\mu_3'\,\mu_1' + 6\mu_2'\,\mu_1'^2 - 3\mu_1'^4$$

$$= 108 - 4 \times (-30) \times (-1.5) + 6 \times 17 \times (-1.5)^2 - 3 \times (-1.5)^4$$

$$= 108 - 180 + 229.5 - 15.1875 = 142.3125$$

Lastly,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-113.25)^2}{(14.75)^3} = \frac{12825.5625}{3209.04687} = 3.996$$

Also

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{(-113.25)^2}{(14.75)^3} = \frac{142.3125}{(14.75)^2} = 0.654$$

NOTES

Example 25: *Consider the following distribution :*

Variable :	*0—10*	*10—20*	*20—30*	*30—40*
Frequency :	*1*	*3*	*4*	*2*

Find (i) *moments about assumed mean*

(ii) *moments about actual mean*

(iii) *moments about zero.*

Solution: Let μ_1', μ_2', μ_3', μ_4' be the moments about any point, *i.e.*, assumed mean (Take A = 25). Consider the following table :

Variable	**Mid-points** x_i	**Frequency** f_i	$u_i = \frac{x_i - A}{h}$ **(A = 25, h = 10)**	$f_i u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
0—10	5	1	– 2	– 2	4	– 8	16
10—20	15	3	–1	– 3	3	– 3	3
20—30	25	4	0	0	0	0	0
30—40	35	2	1	2	2	2	2
		$\Sigma f_i = 10$		$\Sigma f_i u_i = -3$	$\Sigma f_i u_i^2 = 9$	$\Sigma f_i u_i^3 = -9$	$\Sigma f_i u_i^4 = 21$

(i) Now $$\mu_1' = h\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i\right) = 10\left(\frac{1}{10}\times(0-3)\right) = -3$$

$$\mu_2' = h^2\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i^2\right) = 100\left(\frac{1}{10}\times 9\right) = 90$$

$$\mu_3' = h^3\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i^3\right) = 1000\left(\frac{1}{10}\times 9\right) = 900$$

$$\mu_4' = h^4\left(\frac{1}{N}\sum_{i=1}^{n} f_i u_i^4\right) = 1000\times\left(\frac{1}{10}\times 21\right) = 21000$$

(ii) Let $\mu_1, \mu_2, \mu_3, \mu_4$ are the moments about the actual mean, then by definition,

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 90 - (-3)^2 = 81$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= -900 - 3\times(-3)\times(90) + 2(-3)^3$$

$$= -900 - 810 - 54 = -144$$

$$\mu_4 = \mu_4' - 4\mu_3'\ \mu_1' + 6\mu_2'\ \mu_1'^2 - 3\mu_1'^4$$
$$= 21000 - 4 \times (-900) \times (-3) + 6 \times (90) \times (-3)^2 - 3 \times 81$$
$$= 21000 - 10800 + 4860 - 243 = 14817$$

(iii) Let v_1, v_2, v_3, v_4 be the moments about zero, then by definition,

$$v_1 = \mu_1' + A = -3 + 25 = 22$$
$$v_2 = \mu_2 + v_1^2 = 81 + (22)^2 = 565$$
$$v_3 = \mu_3 + 3v_2v_1 - 2v_1^3 = -144 + 3 \times 565 \times 22 - 2 \times (22)^3$$
$$= -144 + 37220 - 21296 = 15850$$
$$v_4 = \mu_4 + 4v_2v_1 - 6v_2v_1^2 + 3v_1^4$$
$$= 14817 + 4 \times 22 \times 15850 - 6 \times (22)^2 \times 565 + 3 \times (22)^4$$
$$= 14817 + 1394800 - 1640760 + 702768 = 471625.$$

6.12 SUMMARY

- *"Skewness means lack of symmetry or lopsidedness in a frequency distribution"*. The object of measuring skewness is to estimate the extent to which a distribution is distorted from a perfectly symmetrical distribution.
- A frequency distribution is called symmetric if the frequencies are symmetrically distributed on both sides of the centre point of the frequency curve.
- **Positively skewed distribution:** A frequency distribution is said to be +ve skewed if the frequency curve, gives a longer tail to the right hand side.
- A frequency distribution is said to be negatively skewed if the frequency curve gives a longer tail on the left hand side.
- If in a frequency distribution, the interval between the mean and median is approximately one-third of the interval between the mean and mode, then the distribution is called moderately symmetrical distribution.
- To compare two or more distributions, the absolute measure of skewness cannot be used. To overcome this, we compute **relative measures of skewness,** which are known as "coefficients of skewness".
- The Karl Pearson's coefficient of skewness (S_{k_P}) cannot be used when the mode is ill-defined.
- The Bowley's coefficient of skewness (S_{k_B}) is defined by

$$S_{k_B} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$
$$= \frac{(Q_3 - \text{Median}) + (Q_1 - \text{Median})}{Q_3 - Q_1}.$$

- The Bowley's coefficient of skewness is also called "quartile measure of skewness" and its values lies between – 1 and 1, *i.e.*, $-1 \le S_{k_B} \le 1$.
- Kurtosis enables us to have an idea about the "flatness or peakedness" of the frequency curve.
- The curves which are neither flat, nor sharply peaked, are known as Normal curves or mesokurtic curves.
- The curves which are flatter than the Normal curves, are known as platykurtic curves.
- The curves which are more sharply peaked than the normal curve, are known as leptokurtic curves.
- In statics, the 'moment' refers to the tendency of a force to rotate a body about a point. In statistics, we define the 'moment' as below.
- Let $x: \quad x_1 \quad x_2 \quad x_3 \quad \ldots\ldots \quad x_n$
  $f: \quad f_1 \quad f_2 \quad f_3 \quad \ldots\ldots \quad f_n$, be a frequency distribution.
  Then r^{th} moment μ_r' about the point $x = A$ is defined by

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r, \text{ where } N = \sum_{i=1}^{n} f_i ; r = 0, 1, 2, \ldots\ldots$$

- Let $x: \quad x_1 \quad x_2 \quad x_3 \quad \ldots\ldots \quad x_n$
  $f: \quad f_1 \quad f_2 \quad f_3 \quad \ldots\ldots \quad f_n$, be a frequency distribution.
  Then r^{th} moment about the origin, *i.e.,* about the point $x = 0$, is defined as

$$v_r = \mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^r, \text{ where } N = \sum_{i=1}^{n} f_i, \quad r = 0, 1, 2, \ldots\ldots$$

- Let $x: \quad x_1 \quad x_2 \quad \ldots\ldots \quad x_n$
  $f: \quad f_1 \quad f_2 \quad \ldots\ldots \quad f_n$, be a frequency distribution.
  Then r^{th} moments μ_r about the mean $\overline{X}$, is defined by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})^r, \text{ where } N = \sum_{i=1}^{n} f_i ; r = 0, 1, 2, \ldots.$$

  and $$\overline{X} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

- Then r^{th} moments μ_r about the mean $\overline{X}$, is defined by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{X})^r, \text{ where } N = \sum_{i=1}^{n} f_i ; r = 0, 1, 2, \ldots.$$

6.13 GLOSSARY

- ***Symmetrical distribution:*** A frequency distribution, in which the values of mean, median and mode are equal, is called symmetrical distribution
- ***Kurtosis:*** A frequency curve may be symmetrical but it may not be equally flat topped with the Normal curve. The relative flatness of the top is called **Kurtosis** or convexity of the frequency curve.
- ***Moment:*** The term "moment" in statistics gives the same meaning as in statics. In statics, the 'moment' refers to the tendency of a force to rotate a body about a point.

6.14 ANSWERS TO CHECK YOUR PROGRESS

1. right-hand side
2. one-third
3. maximum, least
4. bell-shaped
5. absolute measure
6. False
7. True
8. False
9. True
10. True

6.15 TERMINAL AND MODEL QUESTIONS

1. Define skewness and mention absolute and relative measures of skewness.
2. Explain the main types of skewness. What is positive and negative skewness?
3. What are the tests of skewness?
4. Calculate the Karl Pearson's coefficient of skewness from the data given below

x_i :	15	20	25	30	35	40
f_i :	12	28	25	24	20	21

5. Calculate the Karl Pearson's coefficient of skewness from the following data:

Marks :	0—10	10—20	20—30	30—40	40—50	50—60	60—70
No. of students :	10	12	24	32	28	11	3

6. Calculate the quartile coefficient of skewness of the following distribution:

Variate (x):	1—5	6—10	11—15	16—20	21—25	26—30	31—35
Frequency (f) :	3	4	68	30	10	6	2

7. Compute quartile coefficient of dispersion and skewness of the following data:

Central zise :	1	2	3	4	5	6	7	8	9	10
Frequency :	2	9	11	14	20	24	20	16	5	2

8. In a frequency distribution, the coefficient of skewness based upon the quartiles is 0.6. If the sum of the upper and lower quartile is 100 and median is 38, find the value of the upper and lower quartile.
9. For a distribution, the Bowley's coefficient of skewness is 0.36, $Q_1 = 8.6$ and Median = 12.3. Find the quartile coefficient of dispersion.
10. If the first quartile is 142 and semi-inter quartile range is 18. Find the Median, given that the distribution is symmetrical.
11. Compute Bowely's coefficient of skewness from the following data:
    25, 15, 23, 40, 27, 25, 23, 25, 20.
12. Find the measure of skewness based on quartiles and median from the following data.

x_i :	10—20	20—30	30—40	40—50	50—60	60—70	70—80
f_i :	358	2417	976	129	62	18	10

13. What do you understand by kurtosis ? How it is measured?
14. The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Comment upon the nature of the distribution.
15. The first four central moments of a frequency distribution are 0, 60, – 50 and 80, 20 respectively. Discuss the Kurtosis of the distribution.
16. Compute the coefficients of Skewness and Kurtosis based on moments for the following distribution.

x :	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
f :	1	5	12	22	17	9	4	3	1	1

17. Compute μ_2, μ_3 and μ_4 from the following data and determine whether the distribution is platykurtic, leptokurtic or mesokurtic.

Wages :	0—10	10—20	20—30	30—40	40—50	50—6060—70
No. of workers :	8	12	20	30	15	10	5

18. Compute β_1 and β_2 for the following distribution.

Marks :	25—30	30—35	35–40	40—45	45—50	50—55	55—60	60—65
f :	2	8	18	27	25	16	7	2

19. Prove that the frequency distribution curve of the following frequency distribution is leptokurtic.

Marks :	10—15	15—20	20—25	25—30	30—35	35—40	40—45	45—50	50—55
f :	1	4	8	19	35	20	7	5	1

20. Distinguish between skewness and kurtsis.
21. For a distribution the mean is 10, variance is 16, γ_1 is 1 and β_2 is 4. Find the first four moments about the mean.
22. What is the effect of change of origin and scale on moments about any point and central moments of a frequency distribution.
23. The first three moments about the point $x = 7$ are 3, 11, 47 respectively. Find mean, variance and β_1.
24. Calculate the first four moments about the mean for the following data:

Variate :	1	2	3	4	5	6	7	8	9
Frequency :	1	6	13	25	30	22	9	5	2

25. Following table gives the distributions of populations in towns A and B according to different age group.

Age groups	**Population in (,000)**	
	Town *A*	**Town *B***
0—10	18	10
10—20	16	12
20—30	15	24
30—40	12	32
40—50	10	29
50—60	5	11
60—70	2	3
> 70	1	1

(i) Which town has more variation in age distribution?

(ii) Which town is more skewed with regards to age distribution?

6.16 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi

2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi

3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 7: CORRELATION

NOTES

Structure

7.0 INTRODUCTION

In this chapter we shall briefly discuss a method of investigating the relationship between two characteristics, both of which are quantitative in nature. In such investigations we shall have two observations—one for each characteristic made for each individual. For example, age and blood pressure of a group of individuals may have been recorded in the course of an investigation. Are these two characteristics related? If so, what is the nature of this relationship? Does the knowledge of this relationship help us predict more accurately the value of the characteristic of an individual, when the value of the other characteristic of that individual is known?

7.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define correlation and its usefulness
- Explain various types of correlation
- Explain various methods of studying correlation
- Calculate coefficient of correlation by various methods
- Define multiple and partial correlation and properties of their coefficients

7.2 DEFINITION OF CORRELATION

According to Ya Lun Chou, "correlation analysis attempts to determine the degree of relationship between variables".

According to W.I. King, "correlation means that between two series or group of data there exists some casual connection".

According to Croxton and Cowden, "the relationship of quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation".

According to A.M. Tuttle, "correlation is an analysis of the covariation between two or more variables".

Thus, the association of any two variables is known as correlation. In other words, we say that corresponding to a change in one variable there is a change in another variable, they are said to be correlated. This change may be in either direction. If one variable increases (or decreases) the other may also increases (or decreases). The above definitions make it clear that the term "correlation refers to the study of relationship between two variables".

Usefulness

Correlation is useful in physical, biological and social sciences.

1. Correlation is very useful to economists to study the relationship between variables, like price and quantity demanded. The businessmen, it helps to estimate costs, sales, price and other related variables.
2. Some variables show some kind of relationship ; correlation analysis helps in measuring the degree of relationship between the variables like age and blood pressure, supply and demand, price and supply, income and expenditure, heights of human beings and their weights, number of bolls and yield in cotton plant, age of wife and husband etc.
3. The relation between variables can be verified and tested for significance, with the help of correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.
4. The coefficient of correlation is a relative measure and we can compare the relationship between the variables which are expressed in different units.
5. Sampling error can also be calculated.
6. Correlation is the basis for the concept of regression and ratio of variation.

Correlation and Causation

Correlation analysis deals with the association or co-variation between two or more variables and helps to determine the degree of relationship between two or more variables. But correlation does not indicate a cause and effect relationship between two variables. It explains only co-variation. The high degree of correlation between two variables may exist due to any one or a combination of the following reasons.

1. ***Correlation may be due to pure chance:*** Especially in a small sample, the correlation is due to pure chance. There may be a high degree of correlation between two variables in a sample, but in the population there may not be any relationship between the variables. For example, the production of corn, availability of dairy products, chlorophyll content and plant height. These variables have no relationship. However, if a relationship is formed, it may be only a chance or coincidence. Such types of correlation is known as spurious or nonsensical correlation.

2. ***Both variables are influenced by some other variables:*** A high degree of correlation between the variables may be due to some cause or different causes affected each of these variables. For example, a high degree of correlation may exist between the yield per acre of paddy or wheat due to the effect of rainfall and other factors like fertilizers used, favourable weather conditions etc. But none of the two variables is the cause of the other. It is difficult to explain which is the cause and which is the effect, they may not have caused each other, but there is an outside influence.

3. ***Mutual dependence:*** In this case, the variables affect each other. The subjective and relative variable are to be judged for the circumstances. For example, the production of jute and rainfall. Rainfall is the subject and jute production is relative. The effect of rainfall is directly related to the jute production.

Check Your Progress

Fill in the blanks:

1. Correlation analysis deals with ..... between two or more variables.
2. The effect of correlation is to reduce the range of ..... of our prediction.
3. Correlation is useful in ....., ... and social sciences.
4. Correlation is the basis for the concept of ..... and .....
5. Coefficient of correlation is a ..... measure.

7.3 TYPES OF CORRELATION

In a bivariate distribution, correlation is classified into many types, but the important are:

1. Positive and negative correlation
2. Simple and multiple correlation
3. Partial and total correlation
4. Linear and non-linear correlation.

Positive and Negative Correlation

Positive and negative correlation depend upon the direction of the change of the variables. If two variables tend to move together in the same direction *i.e.*, an increase or decrease in the value of one variable is accompanied by an increase or decrease in the value of other variable, then the correlation is called positive or direct correlation. Height and weight, age of wife and husband, intake of calories and proteins, rainfall and yield of crops, price and supply are examples of positive correlation.

If two variables tend to move together in opposite directions *i.e.*, an increase or decrease in the value of one variable is accompanied by a decrease or increase in the value of other variable, then the correlation is called negative or inverse correlation. Price and demand, yield of crops and price, literacy status and total fertility rates among adult female population, rise in prices and consumption of qualitative food (milk or eggs) etc. are the examples of negative correlation.

The following are the quantitative examples of positive and negative correlation:

Positive correlation				**Negative correlation**			
Increase in both the variables		**Decrease in both the variables**		**Increase**	**Decrease**	**Decrease**	**Increase**
x	***y***	***x***	***y***	***x***	***y***	***x***	***y***
7	12	65	41	40	35	50	15
10	19	52	32	45	24	45	20
16	23	43	27	50	20	40	42
20	29	31	21	62	16	25	47
27	36	27	17	67	11	20	51

Simple and Multiple Correlation

When we study only two variables, the relationship is described as simple correlation. For examples, the yield of wheat and use of fertilizers, plant yield and number of

tillers, number of pods and number of clusters, quantity of money and price level, demand and price etc. But in a multiple correlation, we study more than two variables simultaneously. However, multiple correlation consists of the measurement of the relationship between a dependable variable and two or more independent variables. For examples, when we study the relationship between plant yield with that of a number of pods and a number of clusters in pulses and if we study the relationship between agricultural production, rainfall and quantity of fertilizers used, it will be a multiple correlation.

Partial and Total Correlation

To study of two variables excluding some other variables is called partial correlation. For examples, the correlation between yield of maize and fertilizers excluding, the effect of pesticides and manures is called partial correlation, we study price and demand, eliminating the supply side is called partial correlation. In total correlation, all the facts are taken into account.

Linear and Non-linear Correlation

If the ratio of change between two variables is uniform, then there will be linear correlation between them. Consider the following:

x	*2*	*4*	*12*	*20*	*30*
y	*3*	*6*	*18*	*30*	*45*

The ratio of change between the variables is the same. If we plot these on the graph, we get a straight line.

In a curvilinear or non-linear correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variable. The graph of non-linear or curvilinear relationship will form a curve.

In majority of cases, we find curvilinear relationship, which is a complicated one, so we generally assume that the relationship between the variables under study is linear. In social sciences, linear correlation is rare, because the exactness is not so perfect as in natural sciences.

7.4 METHODS OF STUDYING CORRELATION

The different methods of finding out the relationship between two variables are:

A. **Graphic Method**

1. Scatter diagram or scattergram or dot diagram,
2. Simple graph or correlation graph,

B. **Mathematical Method**

3. Karl Pearson's coefficient of correlation,
4. Spearman's coefficient of rank correlation,
5. Coefficient of concurrent deviation,
6. Method of least squares.

Scatter Diagram Method

This is the simplest method of finding out whether there is any relationship present between two variables by plotting the values on a chart, known as scatter diagram. By this method a rough idea about the correlation of two variables can be judged. In this method, the given data are plotted on a graph paper in the form of dots. X variables are plotted on the horizontal axis and Y variables on the vertical axis. Thus, we have the dots and we can know the scatter or concentration of the various points. This will show the type of correlation.

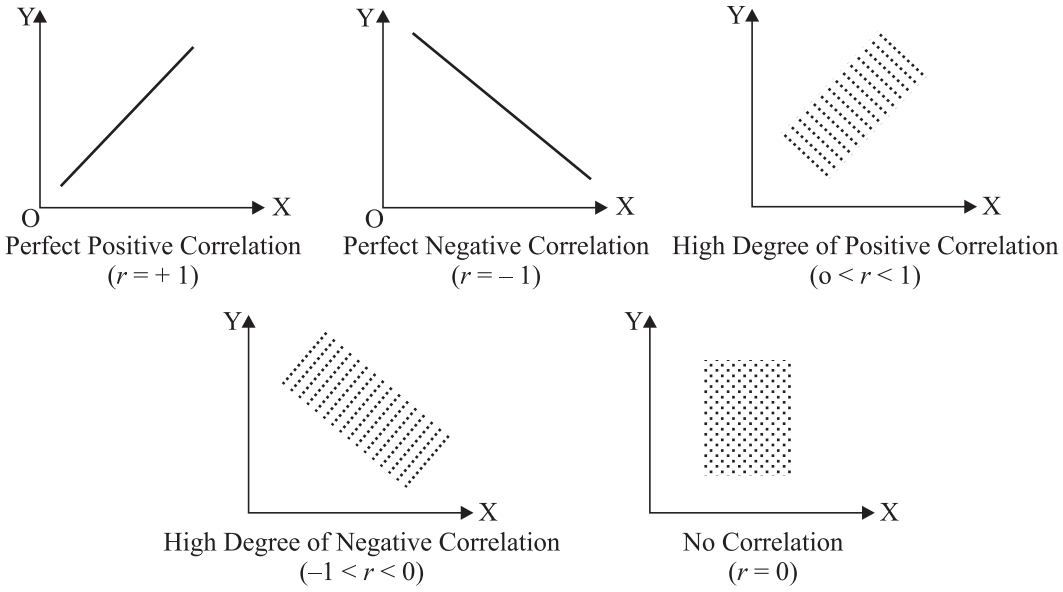


Fig. 7.1

If the plotted points form a straight line running from the lower left-hand corner to the upper righthand corner, then there is a perfect positive correlation (*i.e.*, $r = +1$). On the other hand, if the points are in a straight line, having a falling trend from the upper left-hand corner to the lower right-hand corner, it reveals that there is a perfect negative or inverse correlation (*i.e.*, $r = -1$). If the plotted points fall in a narrow band, and the points are rising from lower left-hand corner to the upper right-hand corner, there will be a high degree of positive correlation between the two variables. If the plotted points fall in a narrow band from the upper left-hand corner to the lower right-hand corner, there will be a high degree of negative

correlation. If the plotted points lie scatter all over the diagram, there is no correlation between the two variables.

Merits and Demerits of Scatter Diagram

Merits

(i) It is a simple and attractive method of finding out the nature of correlation between two variables.

(ii) It is a non-mathematical method of studying correlation. It is easy to understand.

(iii) We can get a rough idea at a glance whether it is positive or negative correlation.

(iv) It is not affected by the extreme values.

(v) The correlation between two variables can be known only on the basis of diagram.

Demerits

(i) It gives only a rough idea about the correlation.

(ii) It does not give the degree or extent of relationship between two variables.

Simple Graph

The values of two variables are plotted on a graph paper, we get two curves, one for X variable and another for Y variable. These two curves reveal the direction and closeness of the two curves and reveal whether or not the variables are related. If both the curves move in the same direction *i.e.*, parallel to each other, either upward or downward, correlation is said to be positive. On the other hand, if they move in opposite directions, then the correlation is said to be negative. Such type of graphical representation is common in ecological, environmental biology and genetical studies.

Example 1: *Draw a correlation graph from the following data:*

Period	*Jan.*	*Feb.*	*Mar.*	*April*	*May*	*June*
Variable 1	*15*	*18*	*22*	*20*	*25*	*20*
Variable 2	*30*	*35*	*43*	*41*	*51*	*40*

Solution:

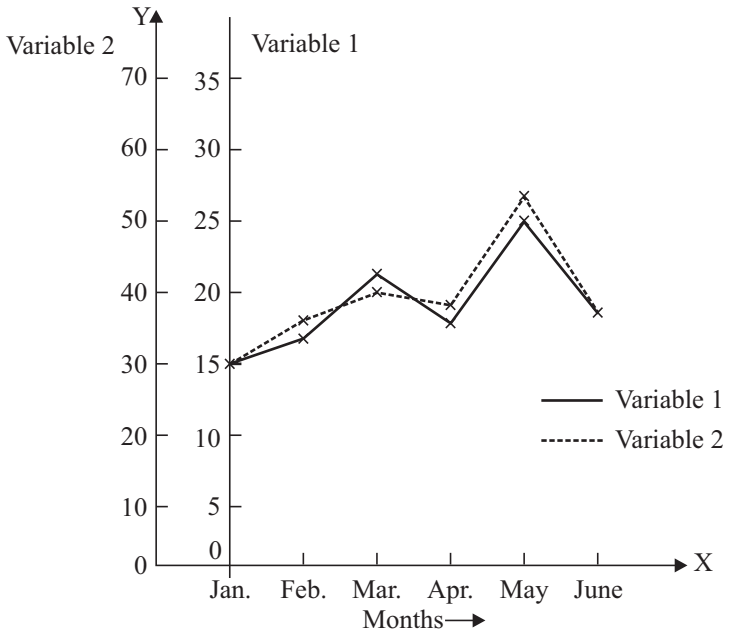


Fig. 7.2: Correlation Graph

The above method is used in the case of time series. This method also does not reveal the extent to which the variables are related.

Karl Pearson's Coefficient of Correlation

Karl Pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between two variables. Karl Pearson's method is the most widely used method in practice and is known as Pearsonian coefficient of correlation. It gives information about the direction as well as the magnitude of the relationship between two variables. It is denoted by the symbol '*r*' ; the formula for calculating Pearsonian *r* is:

$$\text{(i) } r = \frac{\text{Covariance } xy}{\sigma_x \times \sigma_y}, \quad \text{(ii) } r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n\sigma_x \times \sigma_y} \quad \text{(iii) } r = \frac{\Sigma d_x d_y}{\sqrt{\Sigma d_x^2 \times \Sigma d_y^2}}$$

where n = number of pairs of observations,

$d_x = (x - \bar{x})$ = deviation taken from $\bar{x}$ for series x,

$d_y = (y - \bar{y})$ = deviation taken from $\bar{y}$ for series y,

σ_x = Standard deviation of series x,

σ_y = Standard deviation of series y.

NOTES

Calculation of Coefficient of Correlation

The coefficient of correlation in individual or ungrouped data is calculated by three methods:

(i) Direct Method (ii) Actual Mean Method

(iii) Short-cut Method or Assumed Mean Method.

(i) ***Direct Method:*** Sometimes the mean values of the two series are in the form of fraction and the number of observations and their value is not too large, the following formula may be used for calculating the value of r:

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}]}}$$

where n = the number of paired observations,

Σx = the sum of the observations on variable x,

Σy = the sum of the observations on variable y,

Σx^2 = the sum of squares of the observations on variable x,

Σy^2 = the sum of squares of the observations on variable y,

Σxy = the sum of products of observations on variables x and y.

Example 2: *The height and body weight of ten males are given below. Calculate the coefficient of correlation:*

Height (inches)	*65*	*68*	*62*	*70*	*65*	*72*	*67*	*66*	*68*	*70*
Weight (pounds)	*128*	*140*	*120*	*152*	*138*	*160*	*135*	*130*	*125*	*165*

Solution:

Calculation of Correlation Coefficient

Height (inches) x	**Weight (pounds)** y	x^2	y^2	xy
65	128	4225	16384	83320
68	140	4624	19600	9520
62	120	3844	14400	7440
70	152	4900	23104	10640
65	138	4225	19044	8970
72	160	5184	25600	11520
67	135	4489	18225	9045
66	130	4356	16900	8580
68	125	4624	15625	8500
70	165	4900	27225	11550
$\Sigma x = 673$	$\Sigma y = 1393$	$\Sigma x^2 = 45371$	$\Sigma y^2 = 196107$	$\Sigma xy = 94085$

Coefficient of correlation:

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}]}}$$

$$= \frac{10 \times 94085 - 673 \times 1393}{\sqrt{[\{10 \times 45371 - (673)^2\}\{10 \times 196107 - (1393)^2\}]}}$$

$r = 0.837.$

Example 3: *The following data relate to the intake of proteins and gain in body weight of 10 albino rats:*

Serial No. of rat	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Protein intake (gm)/day 'x'	*10*	*11*	*12*	*14*	*16*	*13*	*15*	*12*	*13*	*17*
Gain in weight (gm) 'y'	*12*	*15*	*16*	*22*	*24*	*18*	*23*	*14*	*20*	*25*

Find the value of correlation coefficient.

Solution:

Calculation of Correlation Coefficient

Serial No. of rat	**Protein intake (*gm*)/day '*x*'**	**Gain in weight (*gm*) '*y*'**	x^2	y^2	xy
1	10	12	100	144	120
2	11	15	121	225	165
3	12	16	144	256	192
4	14	22	196	484	308
5	16	24	256	576	384
6	13	18	169	324	234
7	15	23	225	529	345
8	12	14	144	196	168
9	13	20	169	400	260
10	17	25	289	625	425
	$\Sigma x = 133$	$\Sigma y = 189$	$\Sigma x^2 = 1813$	$\Sigma y^2 = 3759$	$\Sigma xy = 2601$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}]}}$$

$$= \frac{10 \times 2601 - 133 \times 189}{\sqrt{[\{10 \times 1813 - (133)^2\}\{10 \times 3759 - (189)^2\}]}}$$

$$= \frac{26010 - 25137}{\sqrt{[(18130 - 17689)(37590 - 35721)]}} = \frac{873}{\sqrt{(441)(1869)}}$$

$$= \frac{873}{\sqrt{824229}} = \frac{873}{907.87}$$

$$r = 0.96.$$

NOTES

(ii) ***Actual Mean Method.*** If the arithmetic means of both the series are not in fraction, this method is suitable for finding cofficient of correlation. The following formula is used:

$$r = \frac{\Sigma d_x d_y}{\sqrt{\Sigma d_x^2 \, \Sigma d_y^2}}$$

where $d_x = x - \bar{x}$, and $d_y = y - \bar{y}$.

It should be noted here that in this method when deviations d_x and d_y are taken from their respective means, the sum of d_x and d_y will always be zero. $\Sigma d_x = 0$ and $\Sigma d_y = 0$.

Example 4: *Find the coefficient of correlation between the heights of fathers and sons from the following data:*

Height of fathers (x)	*64*	*65*	*66*	*67*	*68*	*69*	*70*
Height of sons (y)	*66*	*67*	*68*	*69*	*70*	*71*	*72*

Solution:

Calculation of Coefficient of Correlation

x	$d_x = x - \bar{x}$	d_x^2	y	$d_y = y - \bar{y}$	d_y^2	$d_x d_y$
64	64 – 67 = – 3	9	66	66 – 69 = – 3	9	9
65	65 – 67 = – 2	4	67	67 – 69 = – 2	4	4
66	66 – 67 = – 1	1	68	68 – 69 = – 1	1	1
67	67 – 67 = 0	0	69	69 – 69 = 0	0	0
68	68 – 67 = 1	1	70	70 – 69 = 1	1	1
69	69 – 67 = 2	4	71	71 – 69 = 2	4	4
70	70 – 67 = 3	9	72	72 – 69 = 3	9	9
$\Sigma x = 469$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 28$	$\Sigma y = 483$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 28$	$\Sigma d_x d_y = 28$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{469}{7} = 67$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{483}{7} = 69$$

$$r = \frac{\Sigma d_x \, d_x}{\sqrt{\Sigma d_x^2 \, . \, \Sigma d_y^2}} = \frac{28}{\sqrt{28 \times 28}} = \frac{28}{28} = 1$$

$$r = 1.$$

(iii) ***Short-cut Method or Assumed Mean Method:*** When the arithmetic mean of any or both of the series are in fraction or when the series is large, the calculation by direct method will involve a lot of time. To avoid such tedious calculation, we can use the assumed mean method.

The formula is:

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{[\{n\Sigma d_x^2 - (\Sigma d_x)^2\}\{n\Sigma d_y^2 - (\Sigma d_y)^2\}]}}$$

where d_x = $(x - A)$ = deviation of the observations of x-series from an assumed mean (A).

d_y = $(y - A)$ = deviation of the observations of y-series from an assumed mean (A).

n = number of observations.

Σd_x = sum of the deviations of x-series from an assumed mean,

Σd_y = sum of the deviations of y-series from an assumed mean,

Σd_x^2 = sum of the squares of deviations of x-series from an assumed mean,

Σd_y^2 = sum of the squares of deviations of y-series from an assumed mean,

$\Sigma d_x d_y$ = sum of the product of the deviations of x and y series from their assumed mean.

Example 5: *In an attempt to plot a Dose-Response Curve (DRC), a bioassay was performed with the following results:*

Dose (mg)	*Log dose*	*Response*
2	*0.30*	*32*
4	*0.60*	*58*
8	*0.90*	*94*
16	*1.20*	*120*
32	*1.50*	*150*
64	*1.80*	*174*
128	*2.10*	*213*

Calculate coefficient of correlation using the above data.

Solution: According to the distribution, taking log dose as '*x*' and response as '*y*'.

Log dose 'x'	**Response 'y'**	$d_x = x - A$ **where** $A = 1.2$	$d_y = y - A$ **where** $A = 120$	d_x^2	d_y^2	$d_x d_y$
0.30	32	– 0.90	– 88	0.81	7744	79.2
0.60	58	– 0.60	– 62	0.36	3844	37.2
0.90	94	– 0.30	– 26	0.09	676	7.8
1.20	120	0	0	0	0	0
1.50	150	0.30	30	0.09	900	9.0
1.80	174	0.60	54	0.36	2916	32.4
2.10	213	0.90	93	0.81	8649	83.7
		$\Sigma d_x = 0$	$\Sigma d_y = 1$	Σd_x^2 = 2.52	Σd_y^2 = 24729	$\Sigma d_x d_y$ = 249.3

Using the formula,

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{[\{n\Sigma d_x^2 - (\Sigma d_x)^2\}\{n\Sigma d_y^2 - (\Sigma d_y)^2\}]}}$$

$$= \frac{7 \times 249.3 - 0 \times 1}{\sqrt{[\{7 \times 2.52 - (0)^2\}\{7 \times 24729 - (1)^2\}]}}$$

$$= \frac{1745.1}{\sqrt{(17.64)(173102)}} = \frac{1745.1}{1747.43}$$

$r = 0.9986.$

Example 6: *Find out the coefficient of correlation from the following data:*

Height of father (in inches)	*65*	*66*	*67*	*67*	*68*	*69*	*71*	*73*
Height of son (in inches)	*67*	*68*	*64*	*68*	*72*	*70*	*69*	*70*

Solution:

Calculation of Coefficient of Correlation

Height of father (in inches) 'x'	**Deviations from assumed mean (67) $d_x = x - 67$**	**Square of deviations d_x^2**	**Height of son (in inches) 'y'**	**Deviations from assumed mean (68) $d_y = y - 68$**	**Square of deviations d_y^2**	**Product of deviations of x and y series $d_x d_y$**
65	– 2	4	67	– 1	1	2
66	– 1	1	68	0	0	0
67	0	0	64	– 4	16	0
67	0	0	68	0	0	0
68	1	1	72	4	16	4
69	2	4	70	2	4	4
71	4	16	69	1	1	4
73	6	36	70	2	4	12
	$\Sigma d_x = 10$	$\Sigma d_x^2 = 62$		$\Sigma d_y = 4$	$\Sigma d_y^2 = 42$	$\Sigma d_x d_y = 26$

The coefficient of correlation:

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{[\{n\Sigma d_x^2 - (\Sigma d_x)^2\}\{n\Sigma d_y^2 - (\Sigma d_y)^2\}]}}$$

$$= \frac{8 \times 26 - 10 \times 4}{\sqrt{[\{8 \times 62 - (10)^2\}\{8 \times 42 - (4)^2\}]}} = \frac{168}{\sqrt{126720}}$$

$$= \frac{168}{355.98} = 0.472.$$

Assumptions of Pearsonian Coefficient of Correlation

There are some assumptions of Karl Pearson's coefficient of correlation. They are as follows:

(i) ***Linear Relationship:*** If the two variables are plotted on a scatter diagram, it is assumed that the plotted points will form a straight line. So, there is a linear relationship between the variables.

(ii) ***Normality:*** The correlated variables are affected by a large number of independent causes, which form a normal distribution. Variables like quantity of money, age, weight, height, price, demand, etc. are affected by such forces, that normal distribution is formed.

(iii) ***Casual Relationship:*** Correlation is only meaningful, if there is a cause and effect relationship between the force affecting the distribution of items in the two series. It is meaningless, if there is no such relationship. There is no relationship between rice and weight, because the factors that affect these variables are not common.

(iv) ***Proper Grouping:*** It will be a better correlation analysis if there is an equal number of pairs.

(v) ***Error of Measurement:*** If the error of measurement is reduced to the minimum, the coefficient of correlation is more reliable.

Merits and Demerits

Merits

(i) It is the best measure as far as the algebraic point of view is concerned because it is based on all the observations of both the series.

(ii) It is the most popular mathematical method used for measuring the degree of relationship.

(iii) It is an ideal measure as far as the biostatistical point of view is concerned because it is based on arithmetic mean and standard deviation, which are the best measures of central tendency and dispersion respectively.

(iv) The main features of this coefficient is that it gives information about the direction as well as the magnitude of the relationship between the two variables.

Demerits

(i) It assumes linear relationship between two variables but in practice it is not always possible.

(ii) It lies between + 1 and –1 need a very careful interpretation, otherwise it will be misinterpreted. Careless interpretation will be fallacious.

(iii) It is affected by extreme values.

Mathematical Properties of Karl Pearson's Coefficient of Correlation

(i) Coefficient of correlation lies between + 1 and – 1. Symbolically

$$-1 \leq r \leq +1$$

That is 'r' cannot be less than – 1 and cannot exceed + 1.

(ii) Coefficient of correlation is independent of change of origin and scale of the variables x and y. By change of scale, we mean that all values of x and y series are multiplied or divided by some constant. By change of origin, we mean that a constant is subtracted from all values of x and y series.

Degree of Correlation

The degree of correlation between two variables can be ascertained by the quantitative value of coefficient of correlation which can be found out by calculation, Karl Pearson has given a formula for measuring correlation coefficient (r). However, the results of this formula varies between + 1 and – 1. In case of perfect positive correlation, the result will be $r = +1$ and in case of perfect negative correlation, the result will be $r = -1$. However, in the absence of correlation, the result will be $r = 0$. It indicates that the degree of scattering is very large. In experimental research, it is very difficult to find such values of r as + 1, – 1 and 0. The following table will show the approximate degree of correlation according to Karl Pearson's formula:

Degree of correlation	Positive	Negative
Perfect correlation	+ 1	– 1
Very high degree of correlation	+ 0.9 or more	– 0.9 or more
Sufficiently high degree of correlation	from + 0.75 to + 0.9	from – 0.75 to – 0.9
Moderate degree of correlation	from + 0.6 to + 0.75	from – 0.6 to – 0.75
Only the possibility of correlation	from + 0.3 to + 0.6	from – 0.3 to – 0.6
Possibly no correlation	less than + 0.3	less than – 0.3
Absence of correlation	0	0

Coefficient of Correlation and Probable Error

To find out the reliability or the significance of the value of Karl Pearson's coefficient of correlation, probable error is used. With the help of probable error the limits of coefficient of correlation are obtained and the reliability of the value of the coefficient is assessed. The formula for calculating probable error of Karl Pearson's coefficient of correlation is given by:

$$\text{Probable Error of } r = \frac{0.6745(1 - r^2)}{\sqrt{n}}$$

where 0.6745 is a constant number

r = Pearsonian coefficient of correlation

n = number of pairs.

The limits for population correlation coefficient are:

$$\boxed{r \pm \text{P.E. } (r)}$$

Functions of Probable Error:

(i) If the value of r is less than the probable error, the value of r is not at all significant.

(ii) If the value of r is more than six times the probable error (r = 6 P. E.), the value of r is significant.

(iii) If the probable error is less than 0.3, the correlation should not be considered at all.

(iv) If the probable error is small, the correlation is definitely existing.

Example 7: *Test the significance of correlation for the following values based on the number of observation:*

(i) 10 and

(ii) 100 and r = + 0.4 and + 0.9.

Solution: $r < 6\ P.E.\ (r)$

$$\Rightarrow \quad \frac{r}{P.E.(r)} > 6$$

No. of observations	r	$P.E. = \frac{0.6745(1 - r^2)}{\sqrt{n}}$	$\frac{r}{P.E.(r)}$	Significant/ not significant
10	0.4	$\frac{0.6745\,(1 - (0.4)^2)}{\sqrt{10}} = 0.18$	$\frac{0.4}{0.18} = 2.22$	Not significant
100	0.4	$\frac{0.6745 \times (1 - (0.4)^2)}{\sqrt{100}} = 0.06$	$\frac{0.4}{0.06} = 6.67$	Significant
10	0.9	$\frac{0.6745 \times (1 - 0.9^2)}{\sqrt{10}} = 0.04$	$\frac{0.9}{0.04} = 22.5$	Highly significant
100	0.9	$\frac{0.6745 \times (1 - 0.9^2)}{\sqrt{100}} = 0.0128$	$\frac{0.9}{0.0128} = 70.3$	Very high significant

Note: (i) It will always be good to calculate the probable error before starting the interpretation of coefficient of correlation.

(ii) ***Standard Error:*** Standard error is considered better than the probable error in modern statistics. The formula is:

$$\text{S.E. of } r = \frac{1 - r^2}{\sqrt{n}}.$$

Correlation of Grouped Bi-variate Data

When the number of observations is very large, the data is classified into two-way frequency distribution or correlation table. The class intervals for *y* are in the column headings and for *x,* in the stubs. The formula for calculating the coefficient of correlation is:

$$r = \frac{N\Sigma fd_x d_y - \Sigma fd_x \,.\, \Sigma fd_y}{\sqrt{[\{N\Sigma fd_x^2 - (\Sigma fd_x)^2\}\{N\Sigma fd_y^2 - (\Sigma fd_y)^2\}]}}$$

This formula is the same as the previous formula, which was discussed for assumed mean. The only difference is that here the deviations are also multiplied by the respective frequencies.

Example 8: *Calculate coefficient of correlation between the marks obtained by a batch of 100 students in Biostatistics and Biomathematics as given below:*

Marks in biomathematics	**Marks in biostatistics**					
	20–30	***30–40***	***40–50***	***50–60***	***60–70***	**Total**
15–25	*5*	*9*	*3*			*17*
25–35		*10*	*25*	*2*		*37*
35–45		*1*	*12*	*2*		*15*
45–55			*4*	*16*	*5*	*25*
55–65				*4*	*2*	*6*
Total	*5*	*20*	*44*	*24*	*7*	*100*

Solution: Coefficient of correlation:

$$r = \frac{N\,\Sigma fd_x\, d_y - (\Sigma fd_x)\Sigma(fd_y)}{\sqrt{[\{N\Sigma fd_x^2 - (\Sigma fd_x)^2\}\{N\Sigma fd_y^2 - (\Sigma fd_y)^2\}]}}$$

$$r = \frac{100 \times 88 - (8) \times (-34)}{\sqrt{[\{100 \times 92 - (8)^2\}\{(100 \times 154 - (-34)^2\}]}}$$

$$r = + 0.7953.$$

Example 9: *Establish the formula*

$$\sigma^2_{x-y} = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

where r is the coefficient of correlation between x and y:

Solution: We know that

$$\sigma_x^2 = \frac{1}{n}\Sigma(x - \bar{x})^2$$

$$\therefore \qquad \sigma^2_{x-y} = \frac{\Sigma[(x - y) - (\overline{x - y})]^2}{n}$$

$\overline{x - y}$ = mean of $(x - y)$ series = mean of x – mean of $y = \bar{x} - \bar{y}$

$$\sigma^2_{x-y} = \frac{\Sigma[(x - y) - (\bar{x} - \bar{y})]^2}{n} = \frac{\Sigma[(x - \bar{x}) - (y - \bar{y})]^2}{n}$$

$$= \frac{\Sigma[(x - \bar{x})^2 + (y - \bar{y})^2 - 2(x - \bar{x})(y - \bar{y})]}{n}$$

$$= \frac{\Sigma(x - \bar{x})^2}{n} + \frac{\Sigma(y - \bar{y})^2}{n} - \frac{2\Sigma(x - \bar{x})(y - \bar{y})}{n}$$

$$\sigma^2_{x-y} = \sigma_x^2 + \sigma_y^2 - \frac{2\Sigma(x - \bar{x})(y - \bar{y})}{n} \qquad ...(i)$$

We know that $r = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}$

or $$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n} = r\,\sigma_x\sigma_y \qquad ...(ii)$$

From (*i*) and (*ii*), we get

$$\sigma^2_{x-y} = \sigma_x^2 + \sigma_y^2 - 2\,r\,\sigma_x\sigma_y$$

Proved.

Solution: Let the marks in Biostatistics be denoted by x and the marks in Biomathematics by y.

y	x		*20–30*	*30–40*	*40–50*	*50–60*	*60–70*				
		Mid-point	*25*	*35*	*45 = A*	*55*	*65*				
		$d_x = \frac{(x-A)}{h}$ / $d_y = (y-A)/h$	*– 2*	*– 1*	*0*	*1*	*2*	*f*	fd_y	fd_y^2	fd_xd_y
15–25	Mid point 20	= – 2	20 5	18 9	0 3			17	– 34	68	38
25–35	30	= – 1		10 10	0 25	– 2 2		37	– 37	37	8
35–45	40 = A	= 0		0 1	0 12	0 2		15	0	0	0
45–55	50	= 1			0 4	16 16	10 5	25	25	25	26
55–65	60	= 2				8 4	8 2	6	12	24	16

Solution: Let the marks in Biostatistics be denoted by x and the marks in Biomathematics by y. (*Cont*).

		f	***5***	***20***	***44***	***24***	***7***	***N = 100***	$\Sigma fd_y = -34$	$\Sigma fd_y^2 = 154$	$\Sigma fd_x d_y = 88$
		fdx	– 10	– 20	0	24	14	$\Sigma fd_x = 8$			
		fd_x^2	20	20	0	24	28	$\Sigma fd_x^2 = 92$			
		$fdxdy$	20	28	0	22	18	$\Sigma d_x d_y = 88$			

Spearman's Coefficient of Rank Correlation

In 1904, Charles Edward Spearman, a British psychologist found out the method of ascertaining the coefficient of correlation by ranks. This method is based on rank. This measure is useful in dealing with qualitative characteristics, such as intelligence, beauty, morality, character etc. It cannot be measured quantitatively, as in the case of Karl Pearson's coefficient of correlation, but it is based on the ranks given to the observations. It can be used when the data are irregular or extreme items are erratic or inaccurate, because coefficient of rank correlation is not based on the assumption of formality of data.

For example, we want to find correlation between honesty and beauty of 10 persons, then the numerical measurements are not possible in this case, but these 10 persons can be ranked as 1, 2, 3, ... etc. On the basis of these ranks, Spearman's coefficient of rank correlation can be obtained. Thus, in this method ranks are assigned to various items in the two series and the differences of corresponding rank values are calculated. The formula for Spearman's coefficient of rank correlation which is denoted by R is:

$$R = 1 - \frac{6\,\Sigma D^2}{n^3 - n}$$

where R = The coefficient of rank correlation

D = The difference between pair ranks *i.e.*, $D = R_1 - R_2$

n = Number of paired of observation.

Like the Karl Pearson's coefficient of correlation, the value of R lies between +1 and –1.

We may come across two types of problems:

(*i*) When ranks are given (*ii*) When ranks are not given.

Calculation of Coefficient of Rank Correlation (when ranks are given)

When the actual ranks are given, the following steps are as follows:

1. Compute the difference of the two ranks (R_1 and R_2) and denoted by D.
2. Square the D and get ΣD^2.
3. Substitute the figures in the formula $R = 1 - \frac{6\,\Sigma D^2}{n^3 - n}$.

Example 10: *Following are the ranks obtained by 10 students in two subjects, statistics and mathematics. To what extent the knowledge of the students in the two subject is related?*

Students	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Statistics	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Mathematics	*2*	*4*	*1*	*5*	*3*	*9*	*7*	*10*	*6*	*8*

Solution:

Calculation of Coefficient of Rank Correlation

Students	Rank of statistics (R_1)	Rank of mathematics (R_2)	$D = R_1 - R_2$	D^2
A	1	2	– 1	1
B	2	4	– 2	4
C	3	1	2	4
D	4	5	– 1	1
E	5	3	2	4
F	6	9	– 3	9
G	7	7	0	0
H	8	10	– 2	4
I	9	6	3	9
J	10	8	2	4
				$\Sigma D^2 = 40$

Using the formula,

$$R = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 40}{10^3 - 10} = 1 - \frac{240}{990}$$

$$= 1 - 0.24 = 0.76.$$

Example 11: *Two judges in a beauty competition rank the 12 entries as follows:*

Rank by Ist judge	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Rank by IInd judge	*12*	*9*	*6*	*10*	*3*	*5*	*4*	*7*	*8*	*2*	*11*	*1*

Find the coefficient of rank correlation.

Solution:

R_1	R_2	$D = R_1 - R_2$	D^2
1	12	– 11	121
2	9	– 7	49
3	6	– 3	9
4	10	– 6	36
5	3	2	4
6	5	1	1
7	4	3	9
8	7	1	1
9	8	1	1
10	2	8	64
11	11	0	0
12	1	11	121
			$\Sigma D^2 = 416$

$$R = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

$$= 1 - \frac{6 \times 416}{12^3 - 12}$$

$$= 1 - \frac{2496}{1716}$$

$$= 1 - 1.4545$$

$$R = -0.4545.$$

When ranks are not given: When no rank is given, but actual data are given, then we must give ranks. We can give ranks by taking the highest as 1 or the lowest value as 1, next to the highest (lowest) as 2 and follow the same procedure for both the variables.

Coefficient of Rank Correlation for Equal or Repeated Ranks

When two or more items have equal values, it is difficult to give ranks to them. In such situations average ranks are given to all the items which are of equal value. For example, in (12, 15, 11, 15, 18, 15), 18 is ranked 1, now 15 occurs 3 times *i.e.*, in all we will assign equal rank by average method, this average is obtained by $\frac{2+3+4}{3}$ = 3. Thus, the rank for each 15 will be 3. Next item 12 will be ranked 5 and 11 will get rank 6th. A slightly different formula is used when there is more than one item having the same values. The formula is:

$$R = 1 - \frac{6\left\{\Sigma D^2 + \Sigma\left(\frac{m^3 - m}{12}\right)\right\}}{n^3 - n}$$

where m = the number of items whose ranks are common.

Example 12: *From the following data calculate the coefficient of rank correlation after making adjustment for tied ranks:*

X	*48*	*33*	*40*	*9*	*16*	*16*	*65*	*24*	*16*	*57*
Y	*13*	*13*	*24*	*6*	*15*	*4*	*20*	*9*	*6*	*19*

Solution:

Calculation of Coefficient of Rank Correlation

X	*Y*	Rank for *X* R_1	Rank for *Y* R_2	$D = R_1 - R_2$	D^2
48	13	8	5.5	2.5	6.25
33	13	6	5.5	0.5	0.25
40	24	7	10	– 3	9.00
9	6	1	2.5	– 1.5	2.25
16	15	3	7	4	16.00
16	4	3	1	2	4.00
65	20	10	9	1	1.00
24	9	5	4	1	1.00
16	6	3	2.5	0.5	0.25
57	19	9	8	1	1.00
					$\Sigma D^2 = 41$

16 is repeated 3 times in *X* items, hence $m_1 = 3$, 13 and 6 are repeated twice in *Y* items, hence $m_2 = 2$ and $m_3 = 2$. Therefore, the formula is:

$$R = 1 - \frac{6\left\{\Sigma D^2 + \frac{m_1^3 - m_1}{12} + \frac{m_2^3 - m_2}{12} + \frac{m_3^3 - m_3}{12}\right\}}{n^3 - n}$$

$$= 1 - \frac{6\left\{41 + \frac{3^3 - 3}{12} + \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12}\right\}}{10^3 - 10}$$

$$= 1 - \frac{6\{41 + 2 + 0.5 + 0.5\}}{990} = 1 - \frac{264}{990} = 1 - 0.267$$

$R = 0.733$.

Check Your Progress

State whether the following statements are True or False:

6. If the ratio of change between two variables is uniform, then there will be non-linear correlation between them.
7. Scatter diagram method is affected by extreme values.
8. Pearsonian coefficient of correlation gives information about the direction as well as magnitude of the relationship between two variables.
9. Correlation is meaning only if there is a cause and effect relationship between force affecting the distribution of items in the two series.
10. Karl Pearson's coefficient of correlation lies between 0 and 1.

NOTES

7.5 MULTIPLE AND PARTIAL CORRELATION

When the values of one variable are associated with or influenced by other variable, *i.e.,* the age of husband and wife, the height of father and son, the supply and demand of a commodity and so on, Karl Pearson's coefficient of correlation can be used as a measure of linear relationship between them. But sometimes there is interrelation between many variables and the value of one variable may be influenced by many others, *e.g*., the yield of crop per acre say (x_1) depends upon quality of seed (x_2), fertility of soil (x_3), fertilizer used (x_4), irrigation facilities (x_5), weather conditions (x_6) and so on. Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in that group, our study is that of multiple correlation and multiple regression.

Coefficient of Multiple Correlation

In a trivariate distribution in which each of the variables x_1, x_2 and x_3 has N observations, the coefficient of multiple correlation of x_1 on x_2 and x_3 usually denoted by $R_{1.23}$.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

Similarly, the coefficient of multiple correlation of x_2 on x_1 and x_3.

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}\, r_{23}\, r_{13}}{1 - r_{13}^2}}$$

and the coefficient of multiple correlation of x_3 on x_1 and x_2

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}}$$

where $R_{1.23}$ = coefficient of multiple correlation with x_1 as a dependent variable and x_2 and x_3 as two independent variables

$R_{2.13}$ = coefficient of multiple correlation with x_2 as a dependent variable and x_1 and x_3 as two independent variables.

$R_{3.12}$ = coefficient of multiple correlation with x_3 as a dependent variable and x_1 and x_2 as two independent

r_{12} = coefficient of correlation between x_1 and x_2

r_{23} = coefficient of correlation between x_2 and x_3

r_{13} = coefficient of correlation between x_1 and x_3.

Properties of Coefficient of Multiple correlation

1. Coefficient of multiple correlation measures the closeness of the association between the observed values and the expected values of a variable obtained from the multiple linear regression of that variable on the other variables.

2. Since $R_{1.23}$ is the simple correlation between x_1 and $e_{1.23}$ *(residuals),* it must lie between – 1 and + 1. But $R_{1.23}$ is a non-negative quantity, we conclude that $0 \leq R_{1.23} \leq 1$.
3. If $R_{1.23} = 1$, then association is perfect and all the regression residuals are zero, and a such $\sigma^2_{1.23} = 0$.
4. If $R_{1.23} = 0$, then all total and partial correlation involving x_1 are zero. So x_1 is completely uncorrelated with all the other variables.
5. $R_{1.23}$ is not less than any total correlation coefficient, *i.e.*, $R_{1.23} \geq r_{12}, r_{13}, r_{23}$.

Coefficient of Partial Correlation

Sometimes the correlation between two variables x_1 and x_2 may be partly due to the correlation of a third variable, x_3 with both x_1 and x_2. In such a situation, one may want to know what the correlation between x_1 and x_2 would be if the effect of x_3 on each of x_1 and x_2 were eliminated. This correlation is called the partial correlation and the coefficient of correlation between x_1 and x_2 after the linear effect of x_3 on them has been eliminated is called the coefficient of partial correlation.

The coefficient of partial correlation between x_1 and x_2 excluding x_3 (keeping x_3 constant) is given by

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}}$$

Similarly, the coefficient of partial correlation between x_1 and x_3 excluding x_2 (keeping x_2 constant) is given by

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1 - r_{12}^2)\,(1 - r_{23}^2)}}$$

and the coefficient of partial correlation between x_2 and x_3 excluding x_1 (keeping x_1 constant) is given by

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{(1 - r_{12}^2)\,(1 - r_{13}^2)}}$$

Remarks:

1. If $r_{12.3} = 0$, we have then $r_{12} = r_{13}\, r_{23}$, it means that r_{12} will not be zero if x_3 is correlated with both x_1 and x_2.
2. Coefficient of partial correlation helps in deciding whether to include or not an additional independent variable in regression analysis.

Example 13: *From the data relating to the yield of dry bark (x_1), height (x_2) and girth (x_3) for 18 cinchona plants the following coefficients of correlation were obtained*

$$r_{12} = 0.77,\ r_{13} = 0.72 \quad \text{and} \quad r_{23} = 0.52.$$

Find the coefficient of partial correlation $r_{12.3}$ and coefficient of multiple correlation $R_{1.23}$:

Solution: Coefficient of partial correlation:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}}$$

$$= \frac{0.77 - 0.72 \times 0.52}{\sqrt{\{1 - (0.72)^2\}\,\{1 - (0.52)^2\}}}$$

$$r_{12.3} = 0.62.$$

Coefficient of Multiple correlation:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.77)^2 + (0.72)^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - (0.52)^2}}$$

$$= \sqrt{0.7334}$$

$$R_{1.23} = + 0.8564.$$

(Since the multiple correlation coefficient is non-negative)

Example 14: *On the basis of observations made on 39 cotton plants, the total correlation of yield of cotton (x_1), number of bolls i.e., seed vessels (x_2) and height (x_3) are found to be*

$$r_{12} = 0.8,\ r_{13} = 0.65 \quad \text{and} \quad r_{23} = 0.7.$$

Calculate the coefficient of partial correlation between yields of cotton and number of bolls eliminating the effect of height:

Solution: We have to find the coefficient of partial correlation between the yield of cotton and the number of bolls eliminating the effect of height *i.e.*, $r_{12.3}$.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}}$$

$$= \frac{0.8 - (0.65)\,(0.7)}{\sqrt{\{1 - (0.65)^2\}\,\{1 - (0.7)^2\}}}$$

$$= \frac{0.8 - 0.455}{\sqrt{(1 - 0.4225)\,(1 - 0.49)}}$$

$$= \frac{0.345}{\sqrt{(0.5775)\,(0.51)}} = \frac{0.345}{0.543}$$

$$r_{12.3} = 0.635.$$

7.6 SUMMARY

- Thus, the association of any two variables is known as correlation. In other words, we say that corresponding to a change in one variable there is a change in another variable, they are said to be correlated.
- Correlation is the basis for the concept of regression and ratio of variation.
- Correlation analysis deals with the association or co-variation between two or more variables and helps to determine the degree of relationship between two or more variables.
- Positive and negative correlation depend upon the direction of the change of the variables. If two variables tend to move together in the same direction *i.e.*, an increase or decrease in the value of one variable is accompanied by an increase or decrease in the value of other variable, then the correlation is called positive or direct correlation.
- If two variables tend to move together in opposite directions *i.e.*, an increase or decrease in the value of one variable is accompanied by a decrease or increase in the value of other variable, then the correlation is called negative or inverse correlation.
- When we study only two variables, the relationship is described as simple correlation.
- But in a multiple correlation, we study more than two variables simultaneously.
- To study of two variables excluding some other variables is called partial correlation.
- In total correlation, all the facts are taken into account.
- If the ratio of change between two variables is uniform, then there will be linear correlation between them.
- In a curvilinear or non-linear correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the other variable.
- This is the simplest method of finding out whether there is any relationship present between two variables by plotting the values on a chart, known as scatter diagram.
- If the plotted points form a straight line running from the lower left-hand corner to the upper righthand corner, then there is a perfect positive correlation (*i.e.*, $r = +1$). On the other hand, if the points are in a straight line, having a falling trend from the upper left-hand corner to the lower right-hand corner, it reveals that there is a perfect negative or inverse correlation (*i.e.*, $r = -1$).

- Karl Pearson's method is the most widely used method in practice and is known as Pearsonian coefficient of correlation. It gives information about the direction as well as the magnitude of the relationship between two variables. It is denoted by the symbol '*r*'.
- Coefficient of correlation lies between + 1 and – 1.
- To find out the reliability or the significance of the value of Karl Pearson's coefficient of correlation, probable error is used. With the help of probable error the limits of coefficient of correlation are obtained and the reliability of the value of the coefficient is assessed.
- When the number of observations is very large, the data is classified into two-way frequency distribution or correlation table.

7.7 GLOSSARY

- ***Spurious Correlation:*** These variables have no relationship. However, if a relationship is formed, it may be only a chance or coincidence. Such types of correlation is known as spurious or nonsensical correlation.
- ***Direct Correlation:*** If two variables tend to move together in the same direction *i.e.*, an increase or decrease in the value of one variable is accompanied by an increase or decrease in the value of other variable, then the correlation is called positive or direct correlation.
- ***Pearson Coefficient of Correlation:*** It gives information about the direction as well as the magnitude of the relationship between two variables. It is denoted by the symbol '*r*'.
- ***Partial Correlation:*** Sometimes the correlation between two variables x_1 and x_2 may be partly due to the correlation of a third variable, x_3 with both x_1 and x_2. In such a situation, one may want to know what the correlation between x_1 and x_2 would be if the effect of x_3 on each of x_1 and x_2 were eliminated. This correlation is called the partial correlation.

7.8 ANSWERS TO CHECK YOUR PROGRESS

1. co-variation
2. uncertainty
3. physical, biological
4. regression, ratio of variation
5. relative

6. False
7. False
8. True
9. True
10. False

7.9 TERMINAL AND MODEL QUESTIONS

1. What is meant by correlation? What are the properties of the coefficient of correlation.
2. Define Karl Pearson's coefficient of correlation. What is it intended to measure?
3. Distinguish between:

   (*a*) Positive and negative correlation

   (*b*) Linear and non-linear correlation

   (*c*) Simple, partial and multiple correlation.
4. The following data were collected in an experiment on jute in a village of West Bengal, in which the length *x* (in cm) of green plant and the weight *y* (in gm) of dry fibre were recorded for 8 plants:

Plant no.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
x (in cm)	172	148	162	183	160	141	150	190
y (in gm)	6.4	2.3	3.5	4.7	4.1	2.9	2.8	6.6

   Calculate the coefficient of correlation between *x* and *y*.
5. Calculate Karl Pearson's coefficient of correlation from the data given below:

Roll no.	***1***	***2***	***3***	***4***	***5***	***6***	***7***
Marks in physics	20	35	42	37	13	39	24
Marks in statistics	32	37	50	30	25	24	40

6. Draw a scatter diagram and indicate whether the correlation is positive or negative.

x	10	20	30	40	50	60	70	80
y	32	20	24	36	40	28	48	44

7. Calculate from the following data reproduced pertaining to 66 selected villages in Meerut District, the value of coefficient of correlation between total cultivable areas and the area under wheat.

Area under wheat (in acres)	Total cultivable area (in hect.)					
	0–500	*500–1000*	*1000–1500*	*1500–2000*	*2000–2500*	**Total**
0–200	12	6	–	–	–	18
200–400	2	18	4	2	1	27
400–600	–	4	7	3	–	14
600–800	–	1	–	2	1	4
800–1000	–	–	–	1	2	3
Total	14	29	11	8	4	66

8. Determine Karl Pearson's coefficient of correlation from the following bivariate frequency distribution and also calculate probable error:

Age of husbands (years)	Age of wives (in years)					Total
	23–30	*30–37*	*37–44*	*44–51*	*51–58*	
18–25	9	3	–	–	–	12
25–32	–	20	10	4	–	34
32–39	–	–	12	5	3	20
39–46	–	–	8	7	5	20
46–53	–	–	–	10	4	14
Total	9	23	30	26	12	100

9. The ranks of 10 students in the beginning and at the end of a course are as follows. Find the coefficient of rank correlation.

Before course	1	6	3	9	5	2	7	10	8	4
After course	6	8	3	7	2	1	5	9	4	10

10. Calculate the coefficient of rank correlation for the following table of marks of students in statistics and chemistry:

Marks in statistics	80	64	54	49	48	35	32	29	20	18	15	10
Marks in chemistry	36	38	39	41	27	43	45	52	51	42	40	50

11. Ten competitors in a voice contest are ranked by three judges in the following orders:

First judge	1	6	5	10	3	2	4	9	7	8
Second judge	3	5	8	4	7	10	2	1	6	9
Third judge	6	4	9	8	1	2	3	10	5	7

Use the coefficient of rank correlation to discuss which pair of judges have the nearest approach to common liking in voice.

7.10 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi

2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi

3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 8: REGRESSION ANALYSIS AND PROPERTIES OF REGRESSION COEFFICIENTS

Structure

8.0 INTRODUCTION

Correlation gives the degree and direction relationship between two variables, it does not give the nature of relationship between two variables. In regression, we intend to describe the dependence of a variable on an independent variable. We employ regression equations to lend support to hypothesis regarding the possible causation of changes in y by changes in x, for purposes of prediction of y in terms of x, and for purposes of explaining some of the variations of y by x, by using the latter variable as a statistical control. Studies of the effects of temperature on heartbeat rate, protein intake on growth rate in a child, age of person on blood pressure, and dose of an insecticide on mortality of the insect population are all typical examples of regression.

In regression analysis, we can predict or estimate the value of one variable from the given value of the other variable. Regression explains the functional form of two variables one as dependent variable and other as independent variable. For examples, "temperature and oxygen content of water are correlated. We can find out the expected amount of the dissolved oxygen for a given temperature". "Age and blood pressure

are correlated, we may find the expected amount of systolic blood pressure for a given age. Thus the regression of the systolic blood pressure readings (y) on the age of subjects (x)". "The regression of gain in height or weight (y) on the levels of protein or calorie intake (x)". The studying the way in which the yield of wheat vary in relation to the change of the amount of fertilizer applied, yield would constitute the dependent variable and the independent variable is the fertilizer levels.

Regression means to 'return' or 'going back'. In 1877, Sir Francis Galton, first introduced the word 'regression' while studying the relationship between the heights of fathers and sons. He studies about the heights of 100 fathers and sons and gave his opinion that tall fathers were having tall sons and short fathers were having short sons. He found out that the average height of the sons of tall fathers was less than the average height of the tall fathers whereas the average height of the sons of short fathers, was more than the average height of short fathers. He referred to this tendency to return to the average height as regression. Galton studied the average relationship between these two variables graphically and called the line describing the relationship, the line of regression.

8.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define regression and regression analysis
- Distinguish between correlation and regression
- Define regression equations
- Explain mathematical properties of regression coefficients
- Explain various methods of fitting regression lines
- Define multiple regression

8.2 SIGNIFICANCE OF REGRESSION STUDY

The coefficient of correlation between the two variables gives us an abstract form– a pure number, of the amount of relationship between the two variables. It is an abstract number which measures the degree of relationship between the two variables. While dealing with biological data, we are required to make prediction or estimation. For instances, with a rise in price of wheat, the demand for the commodity goes down, with better monsoon, output of agricultural produces increases etc. The first objective of regression analysis is to provide estimates of the values of the dependent variable from values of independent variable. Prediction or estimation is one of the major problems in almost all spheres of human activity. The prediction or estimation of future activities are important. The regression analysis is one of the scientific method for making such predictions.

NOTES

The regression analysis confined to the study of only two variables at a time is termed as simple regression. The regression analysis for studying more than two variables at a time is known as multiple regression.

Uses of Regression Analysis

Regression analysis is useful in many scientific studies.

(i) Regression analysis is used in biostatistics in all those fields where two or more relative variables are having the tendency to go back to the average.

(ii) Regression analysis predicts the value of dependent variables from the values of independent variables.

(iii) Regression analysis is highly useful and the regression line or equation helps to estimate the value of dependent variable, when the values of independent variables are used in the equation.

(iv) We can calculate the coefficient of correlation (r) with the help of regression coefficients.

(v) Regression analysis in statistical estimation of demand curves, supply curves, production function, cost function, consumption function, etc. can be predicted.

8.3 DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation	Regression
1. Correlation is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction.	1. Regression means going back and it is a mathematical measure showing the average relationship between two variables.
2. Both the variables x and y are random variables.	2. Here x is a random variable and y is a fixed variable. Sometimes both the variables may be random variables.
3. It find out the degree of relationship between two variables and not the cause and effect of the variable.	3. It indicates the cause and effect relationship between the variables and establishes a function relationship.
4. It is used for testing and verifying the relation between two variables and gives limited information.	4. Besides verification, it is used for the prediction of one value, in relationship to the other given value.
5. The coefficient of correlation is a relative measure. The range of relationship lies between $\pm$ 1.	5. Regression analysis is an absolute measure. If we know the value of independent variable, we can find the value of the dependent variable.
6. There may be nonsense correlation between two variables.	6. In regression, there is no such nonsense regression.

(*Contd.*)

Correlation	Regression
7. It has limited application, because it is confined only to linear relationship between the variables.	7. It has wider application, as it studies linear and non-linear relationship between the variables.
8. If the coefficient of correlation is positive, then the two variables are positively correlated and vice versa.	8. The regression coefficient explains that the decrease in one variable is associated with the increase in the other variable.

8.4 REGRESSION EQUATIONS

If the bivariate data are plotted on a graph paper, a scatter diagram is obtained which indicates some relationship between two variables. The dots of scatter diagram tend to concentrate around a curve. This curve is known as regression curve and its functional form is called regression equation. When this curve is a straight line, it is called regression line and regression is said to be linear and if it is a curve, it is called non-linear regression.

In bivariate data we have two variables and therefore, we have two regression lines as follows:

(i) Regression line of y on x, it is denoted by

$$y = a + bx$$

where x is independent variable and y is dependent variable.

(ii) Regression line of x on y, it is denoted by

$$x = a + by$$

where y is independent variable and x is dependent variable.

When the regression lines show some trend upward or downward, we say that there is some correlation between two variables. But if both the lines of regression are perpendicular to each other, we say that both the variables are uncorrelated or $r = 0$. Further, if both the lines of regression coincide, we say that there is a perfect correlation between the variables or $r = \pm 1$.

With the help of bivariate data, the two regression lines can be fitted by the method of least squares, in which we obtain the following equations known as normal equations:

For y on x:

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

NOTES

For x on y:

$$\Sigma x = na + b\Sigma y$$
$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

Solving the above set of normal equations simultaneously for a and b, we get both the regression lines.

If in the bivariate data, the mean, the standard deviation and the coefficient of correlation of both the variables x and y are given, then the regression lines of y on x and x on y are given as:

For y on x:

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

For x on y:

$$(x - \bar{x}) = r \frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

where $r \frac{\sigma_y}{\sigma_x}$ is called regression coefficient of y on x and is denoted by b_{yx}. Similarly $r\frac{\sigma_x}{\sigma_y}$ is called regression coefficient of x on y and is denoted by b_{xy}. It is noted that the regression lines of y on x and x on y both passes through the point $(\bar{x}, \bar{y})$. That is their point of intersection is $(\bar{x}, \bar{y})$.

Check Your Progress

Fill in the blanks:

1. Regression analysis confined to the study of only two variables at a time is termed as .....
2. Regression line of x on y is denoted by .....
3. Functional form of regression curve is called .....
4. If both the lines of regression coincide, there is a ..... between the variables.
5. Regression analysis is an ..... measure.

8.5 MATHEMATICAL PROPERTIES OF THE REGRESSION COEFFICIENTS

The following are the important properties of the regression coefficients:

(i) *The geometric mean of the regression coefficients is the coefficient of correlation. Symbolically*

$$\boxed{r = \sqrt{b_{yx} \times b_{xy}}}$$

Proof.: The regression coefficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

i.e. $$\text{G.M.} = \sqrt{b_{yx} \times b_{xy}}$$

$$= \sqrt{r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}} = \sqrt{r^2} = r$$

$\therefore$ $$r = \sqrt{b_{yx} \times b_{xy}}\,.$$

Note: If b_{yx} is negative, b_{xy} will also be negative. Both regression coefficients should be same algebraic sign. If b_{yx} and b_{xy} are +ve, r will be +ve and vice versa.

(ii) *Arithmetic mean of the regression coefficients is greater than or equal to coefficient of correlation. Symbolically*

$$\boxed{\frac{b_{yx} + b_{xy}}{2} \geq r}$$

Proof.: The regression coefficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$\text{A.M.} = \frac{b_{yx} + b_{xy}}{2} \geq r \quad \text{or} \quad \frac{r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}}{2} \geq r$$

or $$\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} - 2 \geq 0 \quad \text{or} \quad \frac{1}{\sigma_x\sigma_y}[\sigma_x^2 + \sigma_y^2 - 2\sigma_x\,\sigma_y] \geq 0$$

or $$\frac{1}{\sigma_x\,\sigma_y}[\sigma_x - \sigma_y]^2 \geq 0 \text{ which is true.}$$

$\therefore$ $$\frac{b_{yx} + b_{xy}}{2} \geq r.$$

(iii) Regression coefficients are independent of change of origin but not the scale.

(iv) The signs of both the regression coefficients should be same. If they are negative, the value of r will be negative and if they are positive, the value of r will be positive.

NOTES

Example 1: *If* θ *be the acute angle between the two regression lines in the case of two variables x and y, show that*

$$tan\ \theta = \frac{1 - r^2}{r} \cdot \frac{\sigma_x\ \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

where r, σ_x *and* σ_y *have their usual meanings. Explain the significance where* $r = 0$ *and* $r = \pm 1$:

Solution: Lines of regression are

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad ...(i)$$

where the slope, $m_1 = r\frac{\sigma_y}{\sigma_x}$

and $$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad ...(ii)$$

where the slope, $m_2 = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$.

We know that $\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2} = \frac{\frac{1}{r}\frac{\sigma_y}{\sigma_x} - r\frac{\sigma_y}{\sigma_x}}{1 + \frac{1}{r}\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_y}{\sigma_x}}$

$$= \frac{\left(\frac{1}{r} - r\right)\frac{\sigma_y}{\sigma_x}}{1 + \frac{\sigma_y^2}{\sigma_x^2}} = \frac{\left(\frac{1 - r^2}{r}\right)\frac{\sigma_y}{\sigma_x}}{\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}}$$

$$\boxed{\tan\theta = \left(\frac{1 - r^2}{r}\right)\frac{\sigma_x\ \sigma_y}{\sigma_x^2 + \sigma_y^2}}. \quad ...(iii)$$

(a) If $r = 0$, then there is no relationship between the two variables and they are independent.

On putting the value of $r = 0$ in (iii), we get $\tan\theta = \infty$, $\theta = \frac{\pi}{2}$. So the lines (i) and (ii) are perpendicular.

(b) If $r = 1$ or -1, then there is a perfect positive or negative correlation between the two variables.

On putting these values of r in (iii), we get $\tan\theta = 0$ or $\theta = 0$. So lines (i) and (ii) are coincide.

8.6 METHODS OF FITTING REGRESSION LINES

There are mainly three methods for fitting of regression lines:

(i) Method of least squares (normal equations)

(ii) Deviation from actual mean method.

(iii) Deviation from assumed mean method.

(i) ***Method of Least Squares:*** If the set of paired data gives the indication that regression is linear, then we can fit two regression lines (a) y on x (b) x on y. The regression lines are as follows:

For y on x: $y = a + bx$

Normal equations $\begin{cases} \Sigma y = na + b\,\Sigma x \\ \Sigma xy = a\,\Sigma x + b\,\Sigma x^2 \end{cases}$

For x on y: $x = a + by$

Normal equations $\begin{cases} \Sigma x = na + b\,\Sigma y \\ \Sigma xy = a\,\Sigma y + b\,\Sigma y^2 \end{cases}$

Example 2: *Determine the equation of a straight line which best fits the data:*

x	*10*	*12*	*13*	*16*	*17*	*20*	*25*
y	*10*	*22*	*24*	*27*	*29*	*33*	*37*

Solution: Straight line $y = a + bx$

The two normal equations are:

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\,\Sigma x + b\Sigma x^2$$

x	***y***	x^2	***xy***
10	10	100	100
12	22	144	264
13	24	169	312
16	27	256	432
17	29	289	493
20	33	400	660
25	37	625	925
$\Sigma x = 113$	$\Sigma y = 182$	$\Sigma x^2 = 1983$	$\Sigma xy = 3186$

NOTES

Substituting the values of $\Sigma x = 113$, $\Sigma y = 182$, $\Sigma x^2 = 1983$, $\Sigma xy = 3186$ and $n = 7$ in the normal equations:

$$182 = 7 \times a + b \times 113$$

or $$7a + 113b = 182 \quad ...(i)$$

$$3182 = a \times 113 + 1983 \times b$$

or $$113a + 1983b = 3182 \quad ...(ii)$$

Multiplying (i) by 113,

$$791a + 12769b = 20566 \quad ...(iii)$$

Multiplying (ii) by 7,

$$791a + 13881b = 22302 \quad ...(iv)$$

Subtracting (iv) from (iii),

$$-1112b = -1736$$

$$b = \frac{1736}{1112}$$

$$b = 1.56$$

Substitute the value of b in (i), we get

$$7a + 113 \times 1.56 = 182$$

$$a = \frac{5.72}{7}$$

$$a = 0.82$$

The equation of straight line is

$$y = a + bx$$

$$y = 0.82 + 1.\ 56x$$

This is called regression line of y on x.

(ii) ***Deviation from Actual Mean Method:*** If the arithmetic mean of both the series x and y are not in fraction, this method is suitable for fitting regression lines. This method is easier and simpler to calculate than the previous method, which is a tedious one. We can find out the deviations of x and y series from their respective means.

Regression line of y on x:

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \text{or} \quad \boxed{(y - \bar{y}) = b_{yx}(x - \bar{x})}$$

where $\bar{x}$ is the mean of x series *i.e.,* $\bar{x} = \frac{\Sigma x}{n}$

$\bar{y}$ is the mean of y series *i.e.,* $\bar{y} = \frac{\Sigma y}{n}$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \text{the regression coefficient of } y \text{ on } x$$

or $$\boxed{b_{yx} = \frac{\Sigma d_x d_y}{\Sigma d_x^2}}$$

where $d_x = (x - \bar{x})$ = deviation of x series from $\bar{x}$.

$d_y = (y - \bar{y})$ = deviation of y series from $\bar{y}$.

Regression line of x on y:

$$\boxed{(x - \bar{x}) = b_{xy}\,(y - \bar{y})}$$

where $b_{xy} = r\dfrac{\sigma_x}{\sigma_y}$ = regression coefficient of x on y.

or $$\boxed{b_{xy} = \dfrac{\Sigma d_x d_y}{\Sigma d_y^2}}\ .$$

Example 3: *From the following data of the age of husband and the age of wife, form the two regression lines and calculate the husband's age when the wife's age is 16. And also find the value of coefficient of correlation.*

Husband's age	*36*	*23*	*27*	*28*	*28*	*29*	*30*	*31*	*33*	*35*
Wife's age	*29*	*18*	*20*	*22*	*27*	*21*	*29*	*27*	*29*	*28*

Solution: Let the husband's age be denoted by x and wife's age be denoted by y.

Calculation of Regression Lines

Husband's age x	**Wife's age** y	$d_x = (x - \bar{x})$ **where** $\bar{x} = 30$	d_x^2	$d_y = (y - \bar{y})$ **where** $\bar{y} = 25$	d_y^2	$d_x d_y$
36	29	6	36	4	16	24
23	18	– 7	49	– 7	49	49
27	20	– 3	9	– 5	25	15
28	22	– 2	4	– 3	9	6
28	27	– 2	4	2	4	– 4
29	21	– 1	1	– 4	16	4
30	29	0	0	4	16	0
31	27	1	1	2	4	2
33	29	3	9	4	16	12
35	28	5	25	3	9	15
$\Sigma x = 300$	$\Sigma y = 250$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 138$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 164$	$\Sigma d_x d_y = 123$

Mean:

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{300}{10} = 30$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{250}{10} = 25$$

Regression coefficient of y on x:

$$b_{yx} = \frac{\Sigma d_x d_y}{\Sigma d_x^2} = \frac{123}{138} = 0.89$$

Regression coefficient of x on y:

$$b_{xy} = \frac{\Sigma d_x d_y}{\Sigma d_y^2} = \frac{123}{164} = 0.75$$

Regression line of y on x:

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

$$(y - 25) = 0.89\ (x - 30)$$

$$y = 0.89x - 1.7.$$

Regression line of x on y:

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$(x - 30) = 0.75(y - 25)$$

$$x = 0.75y + 11.25$$

When the wife's age $(y) = 16$, then the husband's age (x) is

$$x = 0.75 \times 16 + 11.25 = 12 + 11.25$$

$$x = 23.25.$$

The coefficient of correlation (r) is given by

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0.89 \times 0.75} = \sqrt{0.6675}$$

$$r = 0.817$$

Hence, the coefficient of correlation is 0.817.

(iii) ***Deviation Taken from the Assumed Mean Method:*** The difference between the above said method and this is that instead of taking deviations from the arithmetic mean, we take deviations from the assumed mean. If the actual mean is in fraction, this method can be used.

The regression line of y on x is:

$$\boxed{(y - \bar{y}) = b_{yx}(x - \bar{x})}$$

where

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

The regression line of x on y is:

$$\boxed{(x - \bar{x}) = b_{xy}(y - \bar{y})}$$

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

and $d_x = (x - A)$ = Deviation taken from assumed mean (A) in series-x.

$d_y = (y - A)$ = Deviation taken from assumed mean (A) in series-y.

Note: It is to be noted that in both b_{yx} and b_{xy}, the numerator is same.

Example 4: *Obtain the regression lines of y on x and x on y from the following data and estimate the blood pressure when the age is 50. Also find the value of coefficient of correlation*:

Age (x)	*56*	*42*	*72*	*36*	*63*	*47*	*55*	*49*	*38*	*42*	*63*	*60*
Blood pressure (y)	*147*	*125*	*160*	*118*	*149*	*128*	*150*	*145*	*115*	*140*	*152*	*155*

Solution: Mean: $$\bar{x} = \frac{1}{n}\Sigma x = \frac{628}{12} = 52.33$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{1684}{12} = 140.33$$

Regression coefficient of y on x:

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

Age (x)	**Blood pressure (y)**	$d_x = x - 62$	d_x^2	$d_y = y - 140$	d_y^2	$d_x d_y$
56	147	– 6	36	7	49	– 42
42	125	– 20	400	– 15	225	300
72	160	10	100	20	400	200
36	118	– 26	676	– 22	484	572
63	149	1	1	9	81	9
47	128	– 15	225	– 12	144	180
55	150	– 7	49	10	100	– 70
49	145	– 13	169	5	25	– 65
38	115	– 24	576	– 25	625	600
42	140	– 20	400	0	0	0
68	152	6	36	12	144	72
60	155	– 2	4	15	225	– 30
$\Sigma x = 628$	$\Sigma y = 1684$	$\Sigma d_x = -116$	$\Sigma d_x^2 = 2672$	$\Sigma d_y = 4$	$\Sigma d_y^2 = 2502$	$\Sigma d_x d_y = 1726$

$$b_{yx} = \frac{12 \times 1726 - (-116) \times (4)}{12 \times 2672 - (-116)^2}$$

$$b_{yx} = 1.14.$$

Regression coefficient of x on y is:

$$b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_y^2 - (\Sigma d_y)^2} = \frac{12 \times 1726 - (-116) \times (4)}{12 \times 2502 - (4)^2}$$

$$b_{xy} = 0.71.$$

Regression line of y on x is:

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

$$y - 140.33 = 1.14(x - 52.33)$$

$$y = 1.14x + 80.67.$$

Regression line of x on y is:

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$(x - 52.33) = 0.71(y - 140.33)$$

$$x = 0.71y - 47.30.$$

The estimate value of y when the age $x = 50$.

$$y = 1.14 \times 50 + 80.67$$

$$y = 137.67.$$

The coefficient of correlation is given by

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{1.14 \times 0.71} = \sqrt{0.8094} = 0.8997$$

The coefficient of correlation is 0.90.

Example 5: *Show that two independent variables are uncorrelated.*

Solution: If x and y are independent variables, then

$$\text{cov}(x, y) = 0$$

$$\therefore \quad r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{0}{\sigma_x \sigma_y} = 0$$

Hence two independent variables are uncorrelated.

Note: But the converse of the above is not true. *i.e.*, two uncorrelated variables may not be independent as the following example illustrates.

							Total
x	– 3	– 2	– 1	1	2	3	$\Sigma x = 0$
y	9	4	1	1	4	9	$\Sigma y = 28$
xy	– 27	– 8	– 1	1	8	27	$\Sigma xy = 0$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{0}{6} = 0,$$

$$\text{cov}\,(x, y) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y} = \frac{1}{6} \times 0 - 0 \times \bar{y} = 0$$

$$\therefore \qquad r = \frac{\text{cov}\,(x, y)}{\sigma_x\, \sigma_y} = 0$$

NOTES

Thus in the above example, the variables x and y are uncorrelated. But on careful examination we find that x and y are not independent but they are connected by the relation $y = x^2$. Hence two uncorrelated variables need not necessarily be independent.

8.7 MULTIPLE REGRESSION

The concepts and techniques for analysing the association among three or more variables are natural extensions of those explored in the bivariate situation discussed so far. In multiple regression model, we assume that a linear relationship exists between some variable y, which we call the dependent variable, and k independent variables $x_1, x_2, ..., x_k$. The independent variables are sometimes referred to as explanatory variables, because of their use in explaining the variation in y; or as predictor variables, because of their use in predicting y. In general, we ought to be able to improve our predicting ability by including more independent variables in such an equation.

Multiple Regression Equation: Regression equation of x_1 on x_2 and x_3 is:

$$\boxed{x_{1.23} = a_{1.23} + b_{12.3}\, x_2 + b_{13.2}\, x_3}$$

The values of $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously the normal equations.

$$\Sigma d_{x_1}.d_{x_2} = b_{12.3}\, \Sigma d_{x_2}^2 + b_{13.2}\, \Sigma d_{x_2} d_{x_3}$$

$$\Sigma d_{x_1}.d_{x_3} = b_{12.3}\, \Sigma d_{x_2} d_{x_3} + b_{13.2}\, \Sigma d_{x_3}^2$$

Partial regression coefficients:

$$b_{12.3} = \frac{\sigma_1}{\sigma_2}\left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right)$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3}\left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)$$

where σ_1, σ_2 and σ_3 are the standard deviation of the variables x_1, x_2, x_3.
r_{12}, r_{13} and r_{23} are the product movement correlation coefficient between x_1 and x_2, between x_1 and x_3, between x_2 and x_3.

Example 6: *In a trivariate distribution*

$$\sigma_1 = 3,\ \sigma_2 = 4,\ \sigma_3 = 5,\ r_{23} = 0.4,\ r_{13} = 0.6,\ r_{12} = 0.7.$$

Determine the regression equation of x_1 on x_2 and x_3 if the variates are measured from their means:

Solution: When the variates are measured from mean, the regression equation of x_1 on x_2 and x_3 is given by:

$$x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3,$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2}\left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right)$$

$$= \frac{3}{4}\left(\frac{0.7 - 0.6 \times 0.4}{1 - (0.4)^2}\right) = \frac{3}{4} \times \frac{0.46}{0.84}$$

$$= \frac{3}{4} \times 0.5476 = 0.41$$

$$b_{12.3} = 0.41$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3}\left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)$$

$$= \frac{3}{5} \times \left(\frac{0.6 - (0.7) \times (0.4)}{1 - (0.4)^2}\right)$$

$$= \frac{3}{5} \times \frac{0.32}{0.84} = \frac{3}{5} \times 0.38$$

$$b_{13.2} = 0.228$$

Therefore, the required equation is

$$x_1 = 0.41x_2 + 0.228x_3.$$

Check Your Progress

State whether the following statements are True or False:

6. Arithmetic mean of the regression coefficients is less than coefficient of correlation.
7. If coefficient of regression $r = 0$, then there is no relationship between the two variables.
8. Regression coefficients are dependent upon change of origin.
9. Regression is a mathematical measure showing the average relationship between two variables.
10. Spearman's Rank correlation method can be used when data are irregular.

8.8 SUMMARY

- In regression analysis, we can predict or estimate the value of one variable from the given value of the other variable. Regression explains the functional form of two variables one as dependent variable and other as independent variable.
- Regression means to 'return' or 'going back'.
- The regression analysis confined to the study of only two variables at a time is termed as simple regression. The regression analysis for studying more than two variables at a time is known as multiple regression.
- If the bivariate data are plotted on a graph paper, a scatter diagram is obtained which indicates some relationship between two variables. The dots of scatter diagram tend to concentrate around a curve. This curve is known as regression curve and its functional form is called regression equation.
- Regression line of y on x, it is denoted by

  $$y = a + bx$$

  where x is independent variable and y is dependent variable.
- Regression line of x on y, it is denoted by

  $$x = a + by$$

  where y is independent variable and x is dependent variable.
- The geometric mean of the regression coefficients is the coefficient of correlation.
- Arithmetic mean of the regression coefficients is greater than or equal to coefficient of correlation.
- If $r = 0$, then there is no relationship between the two variables and they are independent.
- If $r = 1$ or -1, then there is a perfect positive or negative correlation between the two variables.
- In multiple regression model, we assume that a linear relationship exists between some variable y, which we call the dependent variable, and k independent variables $x_1, x_2, ..., x_k$.

8.9 GLOSSARY

- ***Regression:*** Regression means to 'return' or 'going back'.
- ***Multiple Regression:*** The regression analysis for studying more than two variables at a time is known as multiple regression.
- ***Regression:*** When this curve is a straight line, it is called regression line.

8.10 ANSWERS TO CHECK YOUR PROGRESS

1. simple regression
2. $x = a + by$
3. regression equation
4. perfect correlation
5. absolute
6. False
7. True
8. False
9. True
10. True

8.11 TERMINAL AND MODEL QUESTIONS

1. The following data are given for marks in biostatistics and biomathematics:

   Mean marks in biostatistics = 39.5
   Mean marks in biomathematics = 47.5
   S.D. of marks in biostatistics = 10.8
   S.D. of marks in biomathematics = 16.8
   Coefficient of correlation (r) = 0.42

   Find the two regression lines.
2. Two random variables have the regression lines $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Find the mean values of x and y and the coefficient of correlation. If the variance of x is 25, find standard deviation of y from the data given.
3. An experiment is conducted to determine the relationship between rainfall and the wheat yield. Find out a regression line of y on x from the data given below.

Rainfall (inches)	1	2	3	4	5	5	6	7	8	9
Wheat yield (bushels)	1	3	2	5	5	4	7	6	9	8

4. Calculate the correlation coefficient and regression coefficient between two measurements of water quality of a lake.

Salinity (%)	2	4	6	8	10	12	14
Dissolved oxygen (mg/l)	4	2	5	10	4	11	12

5. The following table gives various values of two variables.

x	42	44	58	55	89	98	66
y	56	49	53	58	65	76	58

Determine the regression lines which may be associated with these values and calculate Karl Pearson's coefficient of correlation.

6. Calculate the coefficient of correlation for the following data. Interpret your results.

Fertilizers used (metric tonnes)	15	18	20	24	30	35	40	50
Productivity of land (metric tonnes)	85	93	95	105	120	130	150	160

7. Calculate the coefficient of correlation and regression coefficient from the following data recorded on the number of clusters and the number of pods in a pulse variety.

No. of clusters	10	15	16	20	10	12	14	19
No. of pods	45	80	50	30	25	45	65	75

8. The following zero-order correlation coefficients are given:

$$r_{12} = 0.98,\ r_{13} = 0.44, \text{ and } r_{23} = 0.54.$$

Calculate the coefficient of partial correlation between first and the third variables keeping the effect of second variable constant.

9. A panel of Judges A and B graded seven debators and independently awarded the followig marks:

Debators	1	2	3	4	5	6	7
Marks by A	40	34	28	30	44	38	31
Marks by B	32	39	26	30	38	34	38

An 8^{th} debator was awarded 36 marks by Judge A while Judge B was not present. If Judge B were also present, how many marks wluld you expect him to award to the 3th debator assuming that the same degree of relationship exists in their judgement?

10. From the data given below. Find

(i) The coefficent of correlation between marks in economics and statistics

(ii) The two regression equations

(iii) The most likely marks in statistics when the marks in economics are 30.

Marks in Eco.	25	28	35	32	31	36	29	38	34	32
Marks in Sta.	43	46	49	41	36	32	31	30	33	39

11. You are given the following information about advertising and sales:

	Advertising (*X*)	**Sales (*Y*)**
Mean	10	90
S.D.	3	12

Also coefficient of correlation between X and Y = 0.8. Find the two regression lines.

8.12 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 9: TIMES SERIES ANALYSIS

Structure

9.0 INTRODUCTION

The term **'Time Series'** consists of quantitative data which are arranged in the order of their occurrence. For example, when we collect the data regarding population per capita income, sales and prices of a commodity, etc. for a particular time period and arrange the data so obtained in a series, called **time series.** Thus, according to **'Spiegel'.**

'A time series is a set of observations taken at specified times, usually at equal intervals.'

In the analysis of time series, time is the most important variable which may be either year, month, week, day, hour or even minutes or seconds.

9.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define 'Time Series'
- Explain significance of time series analysis
- Define the term 'Trend' and components of time series
- Explain various methods of measuring trend and their merits and demerits

- Define-short-term fluctuations and various methods of its measurement
- Define 'Seasonal Variation' and various methods to measure seasonal variation

NOTES

9.2 SIGNIFICANCE OF TIME SERIES ANALYSIS

The analysis of time series is of great importance because of the following reasons.

(a) **To understand past behaviour:** With the help of time series analysis, we can observe data over a period of time and easily understand the changes which have taken place in the past. This helps us in predicting the future behaviour.

(b) **To predict future behaviour:** With the help of time series analysis, we can plan for the future business activities. There are so many statistical techniques by means of which we can analyse the time series so obtained regarding the prediction of various future variations in business and economics.

(c) **To evaluate current accomplishments:** By using time series analysis, we can investigate the cause for the growth and decay of the achievements in business activities.

9.3 COMPONENTS OF TIME SERIES

We first explain the term **'Trend'.** The term **'Trend',** is the basic tendency of production, sales and income etc. to grow or decline over a period of time. Trend does not include short-range oscillations, it includes steady movements over a long period of time. In business activities, we come across many economic time series. Some series increase slowly and some increase fast. Some others series decrease at varying rates, some remain constant for a long period of time. That is to say, we have two types of trends which are:

(i) Linear or straight line trend

(ii) Non-linear Trend.

We now discuss the various components of time series. There are mainly four types of components or patterns, or movements of a time series, given as

(a) Secular trend

(b) Seasonal variations

(c) Cyclical variations

(d) Irregular variations

(a) **Secular Trend:** The trends that occur as a result of general tendency of the data to increase or decrease, over a long period of time, are known as *secular trends*.

When we say that secular trend refers to the general tendency of the data to increase or decrease over a long period of time, it means, we are to find what constitutes a long period of time. It does not mean several years.

A particular period can be regarded as long or not in the study of secular trend, depends upon the nature of the data. For example, if we study the yearly salaries of employees for 2005 and 2006, we find that in 2006, the salaries are more as compared to that in 2005. This increment cannot be taken as secular trend since the time period from 2005 to 2006 is too short and in this small period, we cannot conclude that the salaries have shown an increasing tendency. Consider another example. In a culture, the number N of bacteria grows to a rate proportional to N. Suppose it becomes doubles in 10 seconds. We are interested to count the number of bacteria after 8 minutes. This increase in number shows a secular trend. It is clear from this example that in first case, two years could not be regarded as a long period whereas in second example, even 8 minutes constitute a long period. Hence it is the nature of the data which decides whether a particular period would be called long or not.

Also, for secular trend, it is not necessary that the data shows upward tendency or downward tendency. For example, if we study the trend of sales of cars over a period of 20 years, and find that except for a year or two, the sales are increasing continuously, we will call it '*a secular increase in sales*'.

(b) **Seasonal Variations:** The trends that take place during a period of 12 months as a result of change in climate, weather conditions etc. are called *seasonal variations* or season variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself. There are some factors causing *seasonal variation* which are as under:

(i) **Weather and climate changes:** It is the most important factor causing seasonal variations. The change in the climate and weather conditions such as rainfall, humidity, etc. effects the different products differently. For example, there is a greater demand of soft drinks and cotton clothes in summer whereas, there is a greater demand of hot drinks and wollen clothes in winter.

(ii) **Traditions and habits of a culture:** For *seasonal variations* in time series, customs, traditions and habits are important factor. For example, on certain occasions like Deepawali, there is a big demand for sweets and also there is a great demand for cash before the festivals. Similarly, most of the students buy their books in the first few months after the opening of schools and colleges. Hence the sales of books, sweets etc. show seasonal variations.

(c) **Cyclic Variations:** Cyclic variations refer to the oscillatory variations in a time series which have a duration anywhere between 2 to 10 years. These variations arise due to **trade cycles** or **business cycles.** A *business cycle* has four phases namely (i) **Prosperity**, (ii) **Recession**, (iii) **Depression**, (iv) **Recovery.** These phases are illustrated in the figure.

From the given graph, it is clear that the business activities (like production, sales, prices etc.) tend to be at their peak during prosperity. These business activities start

NOTES

recessing or falling and come to the lowest limit of decline. This level is called **depression level**. Thus, each phase changes gradually into the phase which follows it in the order given.

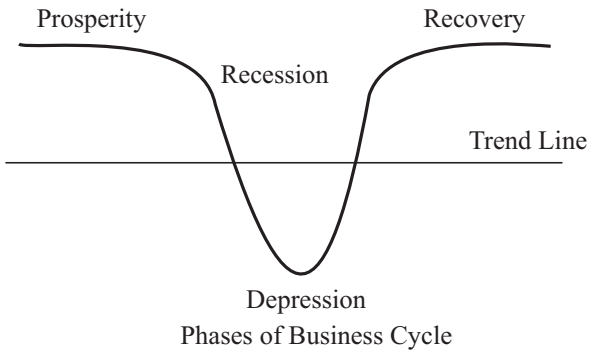


Fig. 9.1

The study of *cyclic variations* is highly useful in framing best policies for stabilizing the level of business activity. For example, their study help us for checking periods of booms and depressions as these periods may bring a complete disaster to the economy.

The cyclic variations are the most typical type of economic fluctuations as they do not show regular periodicity.

(d) **Irregular Variations:** The variations in business activity which do not repeat in a definite pattern are called *irregular variations*. They are also called *erratic variations* or *accidental variations* or *random variations*.

Irregular variations take place due to special causes like floods, earth quakes, strikes and wars etc. The declination in industrial output due to the strike in a factory is an example of irregular variations.

9.4 METHODS OF MEASURING TREND

Given any time series, we can determine its direction which it takes—is it growing or declining?

There are various methods that can be used for determining trend which are listed below:

1. Free hand or graphic method
2. Semi-average method
3. Moving average method
4. Method of least square.

We now discuss in details the above mentioned methods.

1. **Freehand or Graphic Method:** This is the simplest method of studying trend. The various steps involved are given below.

Step I. Plot the given time series on a graph and examine the direction of the trend based on the plotted information.

Step II. Draw a straight line (dotted) which will best fit to the data.

The curve so obtained is called **freehand curve** and this method is also called trend **fitting by inspection.**

Check Your Progress

Fill in the blanks:

1. 'Time series' consists of ..... which are arranged in the order of their occurrence.
2. The business activities start recessing or falling and come to the lowest limit of decline. This level is called .....
3. The variations in business activity which do not repeat in ..... are called irregular variations.
4. Free hand curve method is also called .....
5. Cyclic variations arise due to ..... or .....

Example 1: *Fit a trend line to the following data by Freehand or graphic method:*

Year	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*
Production (million tonnes)	*20*	*22*	*24*	*21*	*23*	*25*	*23*	*26*

Solution: Plot the given time series on a graph paper. The required trend line is shown in the figure given below

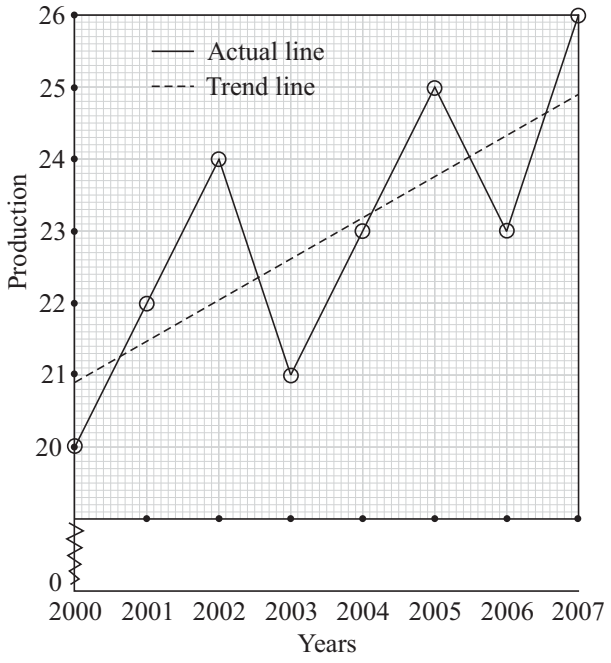


Fig. 9.2

Example 2: *Fit a trend line by 'Free hand curve method' to the data given below:*

Year	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
Sales ('000 ₹)	*150*	*155*	*165*	*152*	*174*	*150*	*174*	*175*	*160*	*180*

Solution: Plot the given time series on a graph paper, the required trend line is shown in the given figure.

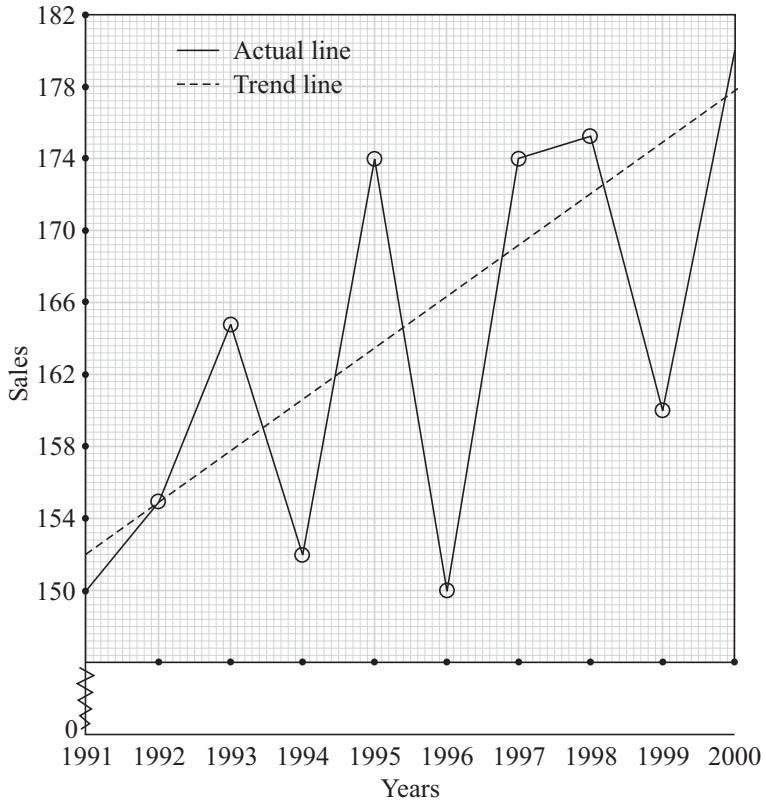


Fig. 9.3

Merits and Demerits of Freehand Curve Method

Merits

(i) It is the simplest method to measure trend.

(ii) It is flexible to use whether the trend is a straight line or a curve.

(iii) It requires no mathematical computations

(iv) It can be used to predict the future behaviour of business activity.

Demerits

(i) This method is subjective in nature as the trend line depends on the personal judgement. It means different persons may draw different trend lines or curves from the same set of given information.

(ii) It is time consuming.

(iii) It leaks accuracy.

2. ***Semi-average Method:*** In this method, the given data is divided into two parts. The various steps involved are

Step I. After dividing the data into two parts, find A.M. of each part.

Step II. Plot the two values of A.M. obtained in step. I on the graph corresponding to the time periods. Join these two points by a straight line, the straight line so obtained is the required trend line.

'Semi-average method' can be applied in two situations given below:

(a) When the number of years given is even

(b) When the number of years given is odd.

Example 3: *Fit a trend line to the following data by using semi-average method:*

Years	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*
Sales ('000 ₹)	*102*	*105*	*114*	*110*	*108*	*116*	*112*

Solution: The number of years given is odd (= 7). So we leave the middle year and we find the A.M. of first three years sales and the last three years sales.

A.M. of first three years sales $= \frac{102 + 105 + 114}{3} = \frac{321}{3} = 107$

A.M. of last three years sales $= \frac{108 + 116 + 112}{3} = \frac{336}{3} = 112.$

Thus we get two points 107 and 112. Plot these points corresponding to their respective middle years *i.e.,* corresponding to 2000 and 2004. Join these two points. The required trend is shown in the given figure.

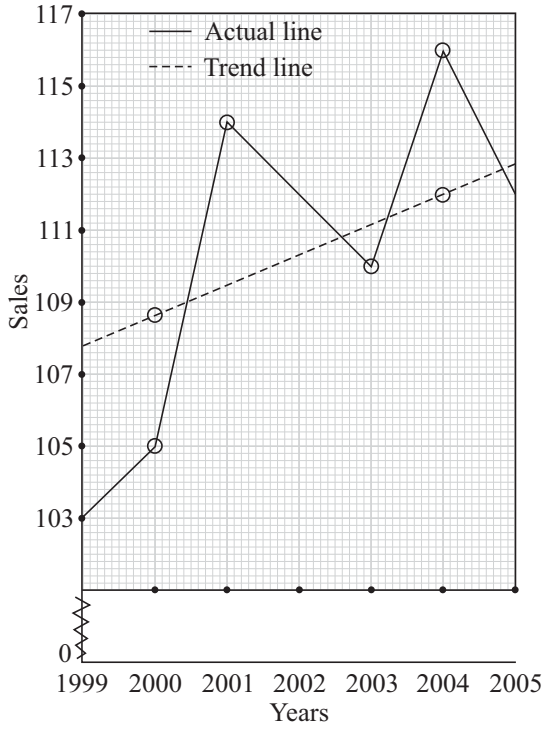


Fig. 9.4

Example 4: *Using semi-average method, fit a trend line for the data given below:*

Years	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*
Sales ('000 ₹)	*20*	*22*	*27*	*26*	*30*	*29*	*40*

Solution: The number of years given is odd (= 7). So we leave the middle year 2002 and calculate the A.M. of the two parts given as.

Year	**Profit ('000 ₹)**	**Semi-Average**	**Middle Year**
1999	20	$\frac{20+22+27}{3} = \frac{69}{3} = 23$	2000
2000	22		
2001	27		
2002	26		
2003	30	$\frac{30+29+40}{3} = \frac{99}{3} = 33$	2004
2004	29		
2005	40		

Thus we get two points 23 and 33. Plot these points corresponding to their respective middle years *i.e.,* corresponding to 2000 and 2004. Join these points, the required trend is shown in the figure.

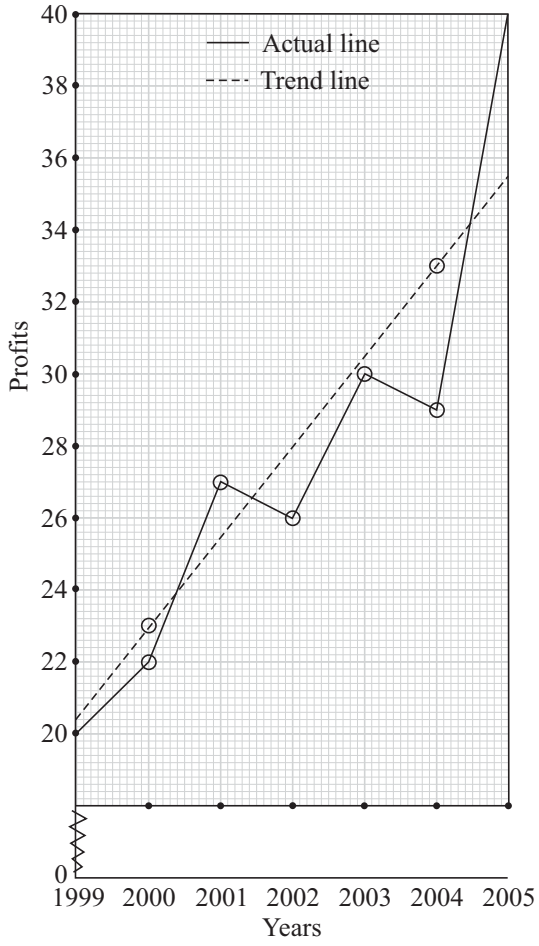


Fig. 9.5

Thus we get two points 275 and 215. Plot these two points corresponding to their respective middle months *i.e.,* at the middle of *March-April* and that of *September–October,* 2005. Join these two points, the required trend is shown in the figure given above.

Merits and Demerits of 'Semi–Average Method'

Merits

1. It is the simplest method as compared to other methods like Moving Average Method and Method of Least Squares.
2. It is objective in nature as everyone will get the same trend for a given data.
3. It is time saving.

Demerits

1. It is based on straight line relationship between the plotted points. But this relationship may or may not exist.
2. It is affected by extreme values. It means that if there are extremes in either half or both halves of the series, the trend line so obtained will not give a true picture of the growth factor.

3. **Moving Average Method:** In this method, we compute moving averages such as 3-yearly moving average, 4-yearly moving average, 5-yearly moving average, etc. The period of moving average is decided by keeping in mind the periodicity of data. The period is determined by plotting the data on the graph paper and the average time interval of successive peaks or toughs are noticed. While selecting the period of moving average, it is necessary to consider that after how many years most of the fluctuations occur in the data.

Moving Average Method is studied in two different situations.

(a) Odd Period Moving Average

(b) Even Period Moving Average.

We now discuss in details the above mentioned method.

Odd Period Moving Average

When the period of moving average is odd, say, K years, where K is odd. The various steps involved are:

Step I. Add all the values corresponding to first K years in the time series and put the sum before the middle year $\left(i.e., \ \frac{K+1}{2}\right)$.

Step II. Leave the first year, apply step I again. Continue this process further till we reach the last value of the series.

Step III. Divide the moving totals obtained in step I and step II by the periods of the moving average.

Step IV. The trend values (or moving averages) of different years.

NOTES

Example 5: *Obtain trend values using 2-yearly moving average for the following data:*

Year	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Production	*412*	*438*	*446*	*454*	*470*	*483*	*490*

Also plot the original and trend values on the same graph paper.

Solution: Here, the period of moving average is odd (= 3). The trend values are obtained in the following table the trend line is shown in the given figure.

Years	**Production**	***3* yearly moving totals**	***3* years moving averages (or Trend values)**
2000	412	—	—
2001	438	→ 412 + 438 + 446 = 1296	$\frac{1296}{3} = 432$
2002	446	→ 438 + 446 + 454 = 1338	$\frac{1338}{3} = 446$
2003	454	→ 446 + 454 + 470 = 1370	$\frac{1370}{3} = 457$
2004	470	→ 454 + 470 + 483 = 1407	$\frac{1407}{3} = 469$
2005	483	→ 470 + 483 + 490 = 1443	$\frac{1443}{3} = 481$
2006	490		

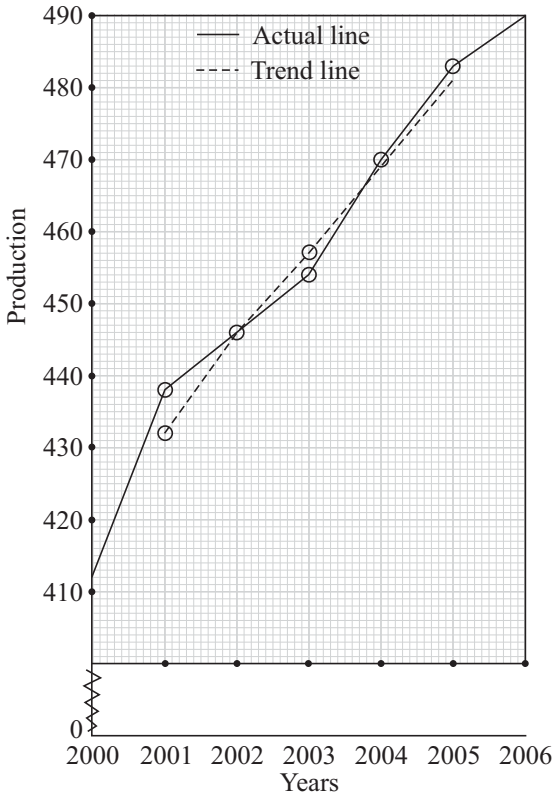


Fig. 9.7

Example 6: *From the data given below, obtain 5 yearly moving average–*

Year	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Sales ('000 ₹)	*16*	*14*	*20*	*18*	*22*	*17*	*19*	*21*	*20*

Also plot the original and trend values on the same graph paper.

Solution: Here, the period of moving average is odd (= 5). The trend values are obtained in the following table. Also the original line and trend line are shown in the given figure.

Years	**Sales ('*000* ₹)**	***5* yearly moving totals**	***5* yearly moving averages (or Trend values)**
1998	16	—	—
1999	14	—	—
2000	20	→ 16 + 14 + 20 + 18 + 22 = 90	$\frac{90}{5} = 18$
2001	18	→ 14 + 20 + 18 + 22 + 17 = 91	$\frac{91}{5} = 18.2$
2002	22	→ 20 + 18 + 22 + 17 + 19 = 96	$\frac{96}{5} = 19.2$
2003	17	→ 18 + 22 + 17 + 19 + 21 = 97	$\frac{97}{5} = 19.4$
2004	19	→ 22 + 17 + 19 + 21 + 20 = 99	$\frac{99}{5} = 19.8$
2005	21	—	—
2006	20	—	—

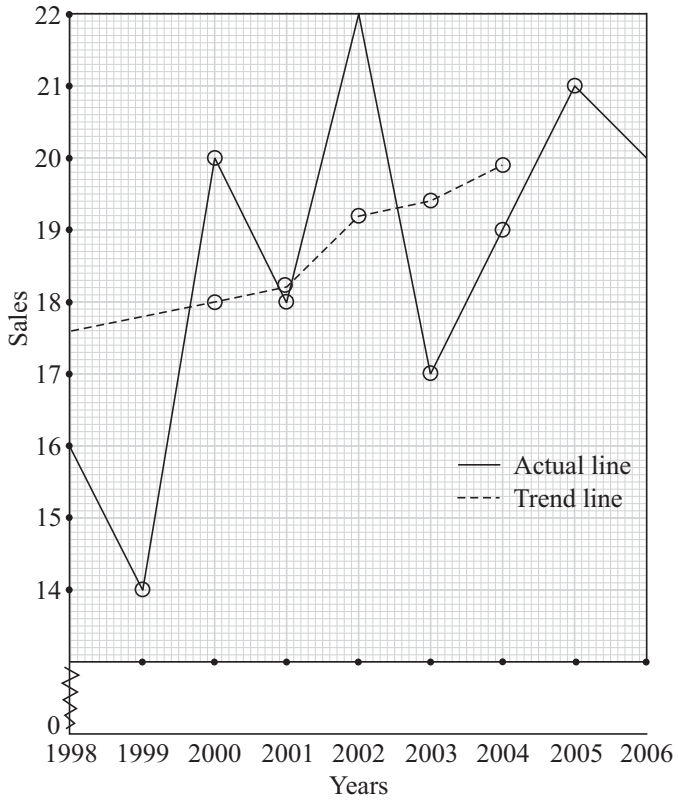


Fig. 9.8

Even Period Moving Average

If the moving average is an even period say, four yearly or six yearly, the moving averages are placed at the centre of the time span. This can be done by the following two methods:

(a) Moving average by centering the totals

(b) Moving average by centering the averages.

We now discuss in details the above mentioned methods.

Moving Average by Centering the Totals

Consider the case of four yearly moving average. The various steps involved are:

Step I. Add all the values corresponding to the first four years and put the sum in between second and third year. After this, the next total (from 2nd to 5th year total) is to be placed in between 3rd and 4th year. Continue this process upto the last value of the series.

Step II. Add the first and second 4 yearly totals and put them in front of 3rd year. Similarly, add second and third 4 yearly total and put them in front of 4th year. Continue this process upto the last value of the series.

Step III. Divide the values obtained in step II by 8, we get the 4 yearly moving averages or trend values.

Example 7: *Obtain trend values using 4 yearly moving average from the data given below*:

Year	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*
Sales (in millions)	*7*	*8*	*9*	*11*	*10*	*12*	*8*	*6*	*5*	*10*

Solution: Here, the period of moving average is even (= 4). We find the trend values by *centering the totals*.

Year	Sales (in millions)	*4* yearly moving totals	*2* period moving totals of *4* yearly moving totals	Trend values or *4* yearly moving average centered
1995	7			
1996	8	→ 7 + 8 + 9 + 11 = 35		
1997	9	→ 8 + 9 + 11 + 10 = 38	→ 35 + 38 = 73	$\frac{73}{8} = 9.12$
1998	11	→ 9 + 11 + 10 + 12 = 42	→ 38 + 42 = 80	$\frac{80}{8} = 10$
1999	10	→ 11 + 10 + 12 + 8 = 41	→ 42 + 41 = 83	$\frac{83}{8} = 10.37$
2000	12	→ 10 + 12 + 8 + 6 = 36	→ 41 + 36 = 77	$\frac{77}{8} = 9.62$
2001	8	→ 12 + 8 + 6 + 5 = 31	→ 36 + 31 = 67	$\frac{67}{8} = 8.37$
2002	6	→ 8 + 6 + 5 + 10 = 29	→ 31 + 29 = 60	$\frac{60}{8} = 7.5$
2003	5	—	—	
2004	10	—	—	

Moving Average by Centering the Average

Consider the case of 4 yearly moving averages. The various steps involved are :

Step I. Add all the values corresponding to the first 4 years and put the sum in between 2nd and 3rd year. After this, the next total (from 2nd to 5th year total) is put in between 3rd and 4th year. Continue this process the last value of the series.

Step II. Divide the 4 yearly totals obtained in step I by 4, we get the 4-yearly uncentered moving averages.

Step III. Add Ist and 2nd 4 yearly moving averages obtained in step II and divide it by 2. Put this sum in front of 3rd year, we get the 4 yearly centred moving averages.

Similarly, add 2nd and 3rd year moving average and divide it by 2 and put this average in front of 4th year. Continue this process upto the last value of the series. These values so obtained are known as *4 yearly centred moving averages or trend values*.

Example 8: *Obtain trend values using 4 yearly moving average from the data given below:*

Year	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*
Sales (in crores)	*7*	*8*	*9*	*11*	*10*	*12*	*8*	*6*	*5*	*10*

Solution: Here, the period of moving average is even (= 4). We find the trend values by *centering the average*.

Year	**Sales (in millions)**	***4* yearly moving totals**	***4* yearly moving average (not centered)**	***2* period moving total of Column IV**	***4* yearly moving averages centered or trend values**
1995	7	—	—	—	—
1996	8	—	—	—	—
		$\rightarrow 7 + 8 + 9 + 11 = 35$	$\frac{35}{4} = 8.75$		
1997	9			$\rightarrow 8.75 + 9.50 = 18.25$	$\frac{18.25}{2} = 9.125$
		$\rightarrow 8 + 9 + 11 + 10 = 38$	$\frac{38}{4} = 9.50$		
1998	11			$\rightarrow 9.50 + 10.50 = 20$	$\frac{20}{2} = 10$
		$\rightarrow 9 + 11 + 10 + 12 = 42$	$\frac{42}{4} = 10.50$		
1999	10			$\rightarrow 10.50 + 10.25 = 20.75$	$\frac{20.75}{2} = 10.375$
		$\rightarrow 11 + 10 + 12 + 8 = 41$	$\frac{41}{4} = 10.25$		
2000	12			$\rightarrow 10.25 + 9 = 19.25$	$\frac{19.25}{2} = 9.625$
		$\rightarrow 10 + 12 + 8 + 6 = 36$	$\frac{36}{4} = 9$		
2001	8			$\rightarrow 9 + 7.75 = 16.75$	$\frac{16.75}{2} = 8.375$
		$\rightarrow 12 + 8 + 6 + 5 = 31$	$\frac{31}{4} = 7.75$		
2002	6			$\rightarrow 7.75 + 7.25 = 15$	$\frac{15}{2} = 7.5$
		$\rightarrow 8 + 6 + 5 + 10 = 29$	$\frac{29}{4} = 7.25$		
2003	5	—	—	—	—
2004	10	—	—	—	—

Period of Moving Average

In time series analysis, we may come across some situations when the period of moving average is not known. The period of the moving average for determining the trend values may be either 3 yearly, or 4 yearly of 5 yearly etc. or some other period. The basic principle in determining the period of moving average is that it should be equal to the period of cyclic variations so that all type of cyclical fluctuations are either eliminated or reduced to minimum.

In some other cases, the period of moving average in a series is not uniform. The cycle may complete in five years or in seven years or in eight or nine years. Under such circumstances, the average duration of the cycle is calculated and this calculated average is taken as the period of moving average.

In most of the cases, the duration of the cycle is found out by plotting original data on a graph paper and reading the time distances between various peaks or troughs. The average of these time distances would give the average duration of the cycle and this is taken as the period of moving average.

The following examples illustrates the above procedure.

Example 9: *Obtain the period of moving average for the following data and hence obtain trend values by using moving average method:*

Year	***Sales (in lakhs)***	***Year***	***Sales in (lakhs)***
1990	*390*	*1998*	*435*
1991	*381*	*1999*	*474*
1992	*372*	*2000*	*459*
1993	*405*	*2001*	*438*
1994	*420*	*2002*	*435*
1995	*396*	*2003*	*492*
1996	*387*	*2004*	*510*
1997	*381*	—	—

Solution: From the given data, we observe that there are peak values *viz.* 390, 420, 474 and 510 for the year 1990, 1994, 1999 and 2004 respectively. Hence the data exhibits a regular cyclic movement with period 5 (length of cycle (1990 – 1994 or 1994 – 1999 or 1999 – 2004 etc.)

Thus, the period of moving average is odd (= 5). The trend values are obtained in the following table.

Year	Sales (in lakhs)	5 yearly moving totals	5 yearly moving Averages or trend values
1990	390	—	—
1991	381	—	—
1992	372	→ 390 + 381 + 372 + 405 + 420 = 1968	$\frac{1968}{5} = 393.6$
1993	405	→ 381 + 372 + 405 + 420 + 396 = 1974	$\frac{1974}{5} = 394.8$
1994	420	→ 372 + 405 + 420 + 396 + 387 = 1980	$\frac{1980}{5} = 396$
1995	396	→ 405 + 420 + 396 + 387 + 381 = 1989	$\frac{1989}{5} = 397.8$
1996	387	→ 420 + 396 + 387 + 381 + 435 = 2019	$\frac{2019}{5} = 403.8$
1997	381	→ 396 + 387 + 381 + 435 + 474 = 2073	$\frac{2073}{5} = 414.6$
1998	435	→ 387 + 381 + 435 + 474 + 459 = 2136	$\frac{2136}{5} = 427.2$
1999	474	→ 381 + 435 + 474 + 459 + 438 = 2187	$\frac{2187}{5} = 437.4$
2000	459	→ 435 + 474 + 459 + 438 + 435 = 2241	$\frac{2241}{5} = 448.2$
2001	438	→ 474 + 459 + 438 + 435 + 492 = 2298	$\frac{2298}{5} = 459$
2002	435	→ 459 + 438 + 435 + 492 + 510 = 2334	$\frac{2334}{5} = 466.8$
2003	492	—	—
2004	510	—	—

Example 10: *Determine the period of moving average for the data given below:*

Year	*Sales (in lakhs)*	*Year*	*Sales in (lakhs)*
1987	*40*	*1997*	*42*
1988	*42*	*1998*	*45*
1989	*40*	*1999*	*46*
1990	*44*	*2000*	*52*
1991	*49*	*2001*	*58*
1992	*46*	*2002*	*56*
1993	*42*	*2003*	*51*
1994	*44*	*2004*	*57*
1995	*44*	*2005*	*54*
1996	*50*	*2006*	*65*

Also find the trend values by using moving average method.

Solution: The given data does not reveal a regular cycle of any fixed period. To determine the most appropriate period of moving average, we first plot the original data as shown in the given figure. Examining the graph carefully, we observe that the data has peaks at the following points.

Year	***1988***	***1991***	***1996***	***1998***	***2001***	***2004***	***2006***
Peak Point	42	49	50	45	58	57	63
Period		3	5	2	3	3	2

The cycle from 1988 to 1991 is of period 3.

The cycle from 1991 to 1996 is of period 5.

The cycle from 1996 to 1998 is of period 2.

The cycle from 1998 to 2001 is of period 3.

The cycle from 2001 to 2004 is of period 3.

The cycle from 2004 to 2006 is of period 2.

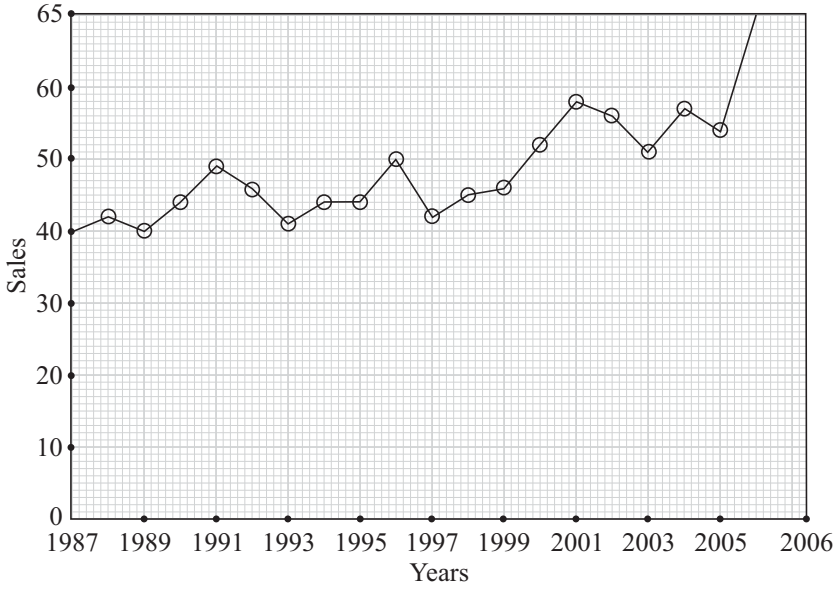


Fig. 9.8

The most appropriate period of moving average is given by the arithmetic mean of periods of different cycles. Thus, the required period of moving average

$$= \frac{3+5+2+3+3+2}{6} = \frac{18}{6} = 3$$

Hence we find 3 yearly moving average. The period of moving average is odd. The trend values are obtained in the following table.

Year	Sales in Lakhs	*3* yearly moving totals	*3* yearly moving averages (or trend values)
1987	40	—	—
1988	42	→ 40 + 42 + 40 = 122	$\frac{122}{3}$ = 40.66
1989	40	→ 42 + 40 + 44 = 126	$\frac{126}{3}$ = 42.00
1990	44	→ 40 + 44 + 49 = 133	$\frac{133}{3}$ = 44.33
1991	49	→ 44 + 49 + 46 = 139	$\frac{139}{3}$ = 46.33
1992	46	→ 49 + 46 + 42 = 137	$\frac{137}{3}$ = 45.66
1993	42	→ 46 + 42 + 44 = 132	$\frac{132}{3}$ = 44.00
1994	44	→ 42 + 44 + 44 = 130	$\frac{130}{3}$ = 43.33
1995	44	→ 44 + 44 + 50 = 138	$\frac{138}{3}$ = 46.00
1996	50	→ 44 + 50 + 42 = 136	$\frac{136}{3}$ = 45.33
1997	42	→ 50 + 42 + 45 = 137	$\frac{137}{3}$ = 45.66
1998	45	→ 42 + 45 + 46 = 133	$\frac{133}{3}$ = 44.33
1999	46	→ 45 + 46 + 52 = 143	$\frac{143}{3}$ = 47.66
2000	52	→ 46 + 52 + 58 = 156	$\frac{156}{3}$ = 52.00
2001	58	→ 52 + 58 + 56 = 166	$\frac{166}{3}$ = 55.33
2002	56	→ 58 + 56 + 51 = 165	$\frac{165}{3}$ = 55.00
2003	51	→ 56 + 51 + 57 = 164	$\frac{164}{3}$ = 54.60
2004	57	→ 51 + 57 + 54 = 162	$\frac{162}{3}$ = 54.00
2005	54	→ 57 + 54 + 63 = 174	$\frac{174}{3}$ = 58.00
2006	63	—	—

9.5 MEASUREMENT OF SHORT-TERM FLUCTUATIONS

The term 'Short-term fluctuations' may be defined as the difference of trend values and the original data. The following examples illustrate the concept of short-term fluctuations more briefly.

Example 11: *From the given data, compute "Trend' and 'short-term variations' by moving average method assuming 'a four yearly cycle'-*

Years	*Sale*
1970	*75*
1971	*60*
1972	*55*
1973	*60*
1974	*65*
1975	*70*
1976	*70*
1977	*75*
1978	*85*
1979	*100*
1980	*70*

Solution: The period of moving average is even (= 4). We find the trend values by *centering* the totals and short-term fluctuations in the following table.

Years	**Sales** *Y*	***4* yearly moving totals**	***2* period moving totals of *4* yearly moving totals**	***4* yearly moving average centered(or trend values)** *T*	**Short-term fluctuations** *Y—T*
1970	75	—	—	—	—
1971	60	—	—	—	—
		→ 75 + 60 + 55 + 60 = 250			
1972	55		→ 250 + 240 = 490	$\frac{490}{8} = 61.25$	– 6.25
		→ 60 + 55 + 60 + 65 = 240			
1973	60		→ 240 + 250 = 490	61.25	– 1.25
		→ 55 + 60 + 65 + 70 = 250			
1974	65		→ 250 + 255 = 455	$\frac{455}{8} = 56.87$	8.13
		→ 60 + 65 + 70 + 70 = 255			
1975	70		→ 255 + 280 = 535	66.87	3.13
		→ 65 + 70 + 70 + 75 = 280			
1976	70		→ 280 + 300 = 580	72.5	– 2.5
		→ 70 + 70 + 75 + 85 = 300			
1977	75		→ 300 + 330 = 630	78.75	– 3.75
		→ 70 + 75 + 85 + 100 = 330			
1978	85		—	—	—
		→ 75 + 85 + 100 + 70 = 330			
1979	100		—	—	—
		—	—	—	
1980	70	—	—	—	—

Example 12: *Using three year moving average determine 'trend' and short-term fluctuations for the following data:*

Years	*Production*
1968	*21*
1969	*22*
1970	*23*
1971	*24*
1972	*24*
1973	*22*
1974	*25*
1975	*26*
1976	*27*
1977	*26*

Solution: Here, the period of moving average is odd (= 3). The trend values and short-term fluctuations are obtained in the following table.

Years	**Production** Y	***3* yearly moving totals**	***3* yearly moving average (or trend values) (T)**	**Short-term fluctuations** $Y—T$
1968	21	—	—	—
1969	22	21 + 22 + 23 = 66	$\frac{66}{3} = 22$	0
1970	23	22 + 23 + 24 = 69	$\frac{69}{3} = 23$	0
1971	24	23 + 24 + 24 = 71	$\frac{71}{3} = 23.66$	0.34
1972	24	24 + 24 + 22 = 70	$\frac{70}{3} = 23.33$	0.67
1973	22	24 + 22 + 25 = 71	$\frac{71}{3} = 23.66$	– 1.66
1974	25	22 + 25 + 26 = 73	$\frac{73}{3} = 24.33$	0.67
1975	26	26 + 26 + 27 = 78	$\frac{78}{3} = 26$	0
1976	27	26 + 27 + 26 = 79	$\frac{79}{3} = 26.33$	0.67
1977	26	—	—	—

Merits and Demerits of Moving Average Method

Merits

The method of *moving average* has the following advantages.

1. It is a simple method if compared to the method of least squares.
2. It is a flexible method. It means, if some more figures are added to the data, the previous calculations will not change. But we will get some more trend values.
3. It is most suitable method in eliminating cyclic fluctuations. If the period of moving average happens to coincide with the period of cyclic fluctuations in the given data, cyclic fluctuations are automatically eliminated.

Demerits

There are some limitations of this method which are given below.

1. Trend values cannot be obtained for all the years. For example, in a 3 yearly moving average, trend values cannot be obtained for the first year and the last year.
2. There is no hard and fast rule in selecting the period of moving average. One has to select the period of moving average based on his own judgement.
3. It cannot be used in forecasting as this method is not represented by a mathematical function.
4. It gives no appropriate computations when the trend is a straight line. The moving average lies either above or below the true sweep of the data.

Method of Least Squares

One of the best method of trend fitting in a time series analysis is the method of least squares. This method is widely used in practice. To fit a trend line, consider the following conditions.

(i) Sum of the deviations of the actual value and computed trend value is zero *i.e.,* $\Sigma(Y - Y_c) = 0$, where

Y = the actual values

Y_c = the trend values

(ii) Sum of the squares of the deviations of the actual values Y and computed trend values Y_c is least from this line, *i.e.,* $\Sigma(Y - Y_c)^2 = 0$.

The line obtained satisfying the conditions (i) and (ii) is known as '**line of best fit**' '**method of least squares**' can also be used to fit **parabolic trend** or exponential trend.

I. **Fitting of a Straight Line Trend:**

Step I. Consider a straight line trend given by

$$Y = a + bX, \quad Y = \text{trend values}$$

$$X = \text{unit of time.}$$

To find the *line of best fit,* we need to determine the constants a and b satisfying the normal equations given below

(*i*) $\Sigma Y = na + b\Sigma X$, n = number of years

(*ii*) $\Sigma XY = a\Sigma X + b\Sigma X^2$

Step II. Find the trend values corresponding to different years and plot these values on a graph paper, we get the required straight line trend.

Year of Origin: To fit a straight line trend by the method of least squares, the first step is to assign any year as the **year of origin.** There are two methods to predict which year should be taken as year of origin.

(a) ***Direct Method:*** Any year can be taken as the year of origin. But generally, first year or before that is taken as year of origin. Let X denotes the time deviations. Then X will take the values 0 for the year of origin, 1 for the next year, 2 for the second next and so on, provided there is no gap in the given data.

(b) ***Short-cut Method:*** In this method, the middle year is taken as the year of origin. It should be noted that in case of odd number of years, when the deviations are taken from the middle year, ΣX would always be zero provided there is no gap in the given data. However, in case of even years also, ΣX will be zero if the year of origin is placed midway between the two middle years.

Example 13: *Below are given the figures of production (in thousand quintals) of a rice factory:*

Years	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Production (in '000 qtls)	*80*	*90*	*92*	*83*	*94*	*99*	*92*

(a) *Fit a straight line trend to the data*

(b) *Show the trend line on a graph paper.*

Solution: (a) Let the required straight line of best fit is

$$Y = a + b\,X \qquad \ldots(1)$$

Here $n = 7$ (odd) $\therefore$ We take 2003 as the year of origin.

Consider the following table. (Apply Direct method)

Years	Production ('*000* qtls) *Y*	Deviation (year of origin = *2003*) *X*	*XY*	X^2
2000	80	– 3	– 240	9
2001	90	– 2	– 180	4
2002	92	– 1	– 92	1
2003	83	0	0	0
2004	94	1	94	1
2005	99	2	198	4
2006	92	3	276	9
	$\Sigma Y = 630$	$\Sigma X = 0$	$\Sigma XY = 56$	$\Sigma X^2 = 28$

The normal equations are

$$\Sigma Y = 7a + b\Sigma X$$

and $$\Sigma XY = a\Sigma X + b\Sigma X^2$$

or $$630 = 7a \quad \Rightarrow \quad a = 90$$

and $$56 = 28b \quad \Rightarrow \quad b = 2$$

$\therefore$ From (1), the required equation of the straight line trend is $Y = 90 + 2X$.

(b) ***To Obtain Trend Values:*** We know that if $Y = a + bX$ represents a straight line trend, then Y represents the trend values corresponding to X unit of time.

$\therefore$ The trend values are obtained in the following table.

Years *Y*	*X*	Trend Values (Y_c) *Y = 90 + 2X*
2000	– 3	84
2001	– 2	86
2002	– 1	88
2003	0	90
2004	1	92
2005	2	94
2006	3	96

The trend line is shown in the given figure

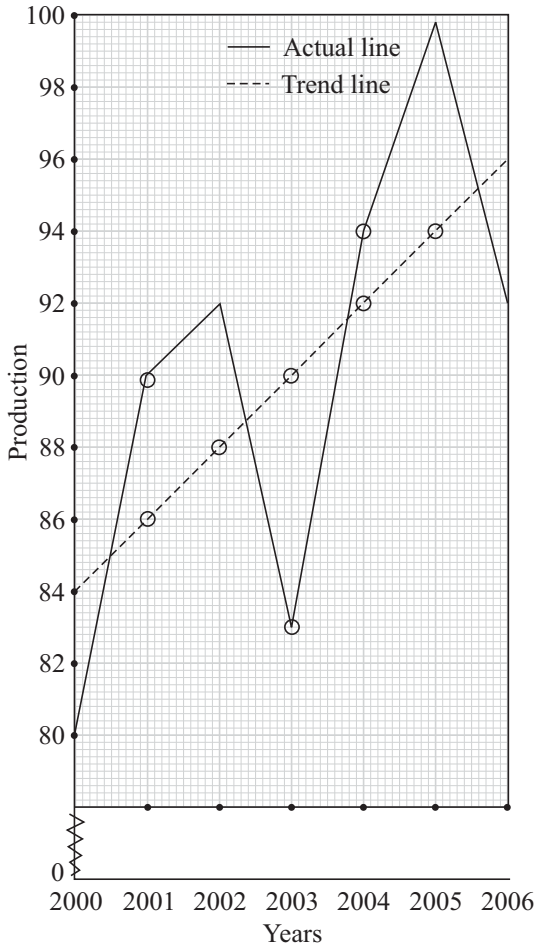


Fig. 9.9

Example 14: *From the data given below:*

Year	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Sales	*80*	*90*	*92*	*83*	*84*	*99*	*92*	*104*

(a) *Fit a straight line trend to the data*

(b) *Show the trend line on a graph paper.*

Solution: (a) Let the required straight line of best fit is

$$Y = a + bX \qquad ...(1)$$

Here $n = 8$ (even), there are two middle years namely 2002 and 2003.

$$\text{A.M. of 2002 and 2003} = \frac{2002 + 2003}{2} = 2002.5$$

Hence we take year of origin as 2002.5, consider the following table.

Years	**Sales** ***Y***	**Deviations from** ***2002.5*** ***x***	**Deviations multiplied by** ***2*** ***X = 2x***	***XY***	X^2
1999	80	– 3.5	– 7	– 560	49
2000	90	– 2.5	– 5	– 450	25
2001	92	– 1.5	– 3	– 276	9
2002	83	– .5	– 1	– 83	1
2003	94	+ .5	1	94	1
2004	99	1.5	3	297	9
2005	92	2.5	5	460	25
2006	104	3.5	7	728	49
	$\Sigma Y = 734$		$\Sigma X = 0$	$\Sigma XY = 210$	$\Sigma X^2 = 168$

The normal equations are

$$\Sigma Y = na + b\Sigma X \quad \text{and} \quad \Sigma XY = a\Sigma X + b\Sigma X^2$$

or $$734 = 8a \quad \Rightarrow \quad a = \frac{734}{8} = 91.75$$

and $$210 = 168b \quad \Rightarrow \quad b = \frac{210}{168} = 1.25$$

$\therefore$ From (1), the required equation of straight line trend is

$$Y = 91.75 + (1.25)X \quad \text{or} \quad Y = 91.75 + 2.5x \qquad | X = 2x$$

(b) ***To Obtain Trend Values:*** We know that if $Y = a + bX$ represents a straight line trend, then Y represents the trend values corresponding to X unit of time. The trend values are obtained in the following table.

Years	***Y***	***X = 2x***	**Trend values** (Y_c) ***Y = 91.75 + (1.25)X***
1999	80	– 7	83
2000	90	– 5	85.5
2001	92	– 3	88
2002	83	– 1	90.5
2003	94	1	93
2004	99	3	95.5
2005	92	5	98
2006	104	7	100.5

The trend line is shown in the given figure.

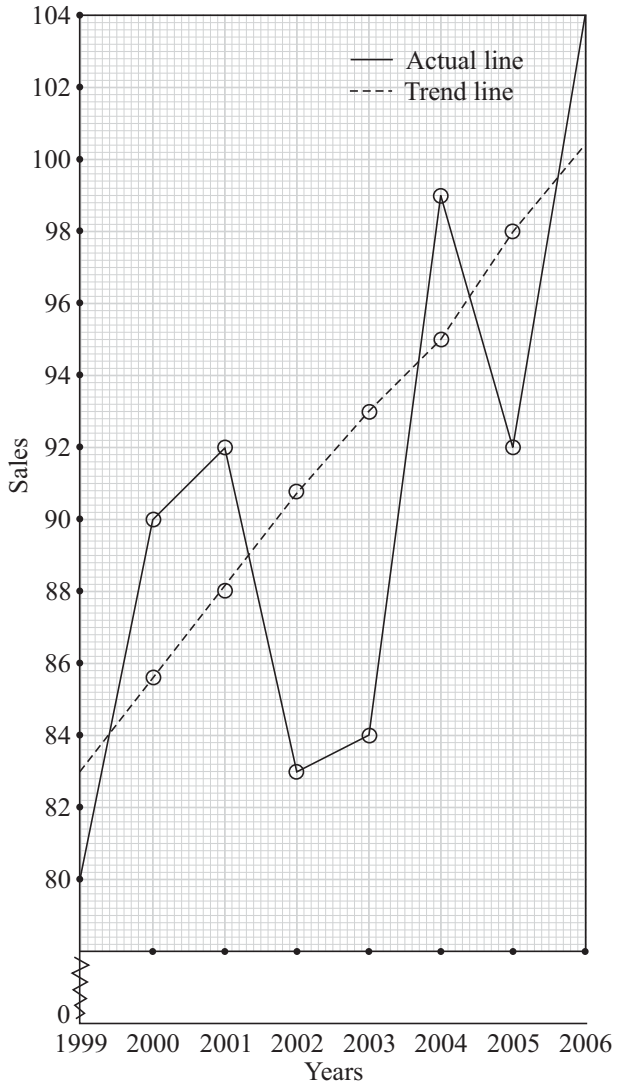


Fig. 9.10

Note: The same result will be obtained if we do not multiply the deviations by 2. But in that case our computations would be more difficult as could be seen below.

Years		**Deviations from *2002.5***		
	Y	***X***	***XY***	X^2
1999	80	– 3.5	– 280	12.25
2000	90	– 2.5	– 225	6.25
2001	92	– 1.5	– 138	2.25
2002	83	– .5	– 41.5	0.25
2003	94	.5	47	0.25
2004	99	1.5	148.5	2.25
2005	92	2.5	230	6.25
2006	104	3.5	364	12.25
		$\Sigma X = 0$	$\Sigma XY = 105$	$\Sigma X^2 = 42$

$\therefore$ The normal equations are

$$\Sigma Y = na + b\Sigma X \quad \text{and} \quad \Sigma XY = a\Sigma X + b\Sigma X^2$$

or $$734 = 8a \implies a = \frac{734}{8} = 91.75$$

and $$105 = 42b \implies b = \frac{105}{42} = 2.5$$

$\therefore$ $Y = 91.75 + 2.5\ X$, which is same.

NOTES

Shifting the Origin

For simplicity of computation, we generally fit the trends to annual data with the middle of the series as origin. However, to study cyclic or seasonal variations, it is necessary to change the origin of the trend equation to some other point in the series. Annual trend may be changed to monthly trend or quarterly trend.

Case I. Shifting the origin for a straight line:

The equation of straight line is $Y = a + bX$.

Here a is called Y-intercept.

After shifting the origin, the new straight line trend is $Y_t = a + b\,(X + k)$, provided k is positive and k = number of times unit shifted forward or $Y_t = a + b\,(X - k)$, provided k is negative and k = number of times unit shifted backward.

Case II. Shifting of origin for parabolic curves: The equation of the parabolic trend is

$$Y = a + bX + cX^2$$

After shifting the origin, the new parabolic trend is

$$Y_t = a + b\,(X + k) + c\,(X + k)^2$$

where k = number of time units shifted.

Also take k positive if the number of times unit shifted forward and negative if shifted backward.

Case III. Shifting of origin for exponential curve: The equation of the exponential trend is

$$Y = ab^X$$

After shifting the origin, the new exponential trend is $Y = ab^{(X+k)}$, if k = number of times unit shifted forward.

or $Y = ab^{(X-k)}$, if k is shifted backward.

Conversion of Annual Trend Equation to Monthly Trend Equation

Given annual trend equation, we can convert it into monthly trend equation.

Case I. When the trend equation is a straight line given by $Y = a + bX$ ...(1)

Assume (1) is given in annual totals. To convert (1) into monthly trend equation, divide ***a*** by 12 and ***b*** by 144 (Since the data are sums of 12 months and hence ***a*** and ***b*** must be divided by 12 and *b* is again divided by 12 so that the time units (X's) will be in months as well).

Thus the monthly trend equation is

$$Y = \frac{a}{12} + \frac{b}{144} X$$

Similarly, the annual trend equation (1) can be converted to quarterly trend equation by dividing ***a*** by 4 and ***b*** by 48. Thus the quarterly trend equation is

$$Y = \frac{a}{4} + \frac{b}{48} X$$

Case II. When the trend equation is a parabola given by

$$Y = a + bX + cX^2 \qquad ...(1)$$

Assume (1) is given in annual totals. To convert (1) into monthly trend equation, divide ***a*** by 12, ***b*** by $12 \times 12 = 144$ and ***c*** by $12 \times 12 \times 12 = 1728$.

Thus the monthly trend equation is

$$Y = \frac{a}{12} + \frac{b}{144} X + \frac{c}{1728} X^2$$

Similarly, the annual trend equation (1) can be converted to quarterly trend equation by dividing ***a*** by 4, ***b*** by 48 and ***c*** by 576. Thus the quarterly trend equation is

$$Y = \frac{a}{4} + \frac{b}{48} X + \frac{c}{576} X^2.$$

Example 15: *Consider the trend equation*

Y = 110 + 2X, where year of origin is 1999 and time unit = 1 year shift the origin to 2003.

Solution: Given year of origin is 1999 and we are required to shift it to 2003, *i.e.,* 4 years forward.

Here $k = 4$ $\therefore$ The required equation is

$$Y_t = 110 + 2 (X + k)$$
$$= 110 + 2(X + 4) = 118 + 2X.$$

Example 16: *Consider the trend equation Y = 210 – 1.5 X, where year of origin is 2003 and time unit = 1 year shift the origin to 1999.*

Solution: Given year of origin is 2003 and we are required to shift the origin from 2003 to 1999 as means going back by 4 years.

$\therefore$ $k = -4$. Hence the required equation is

$$Y_t = 210 - 1.5 (X - k)$$
$$= 210 - 1.5 (X - 4)$$
$$= 216 - 1.5X.$$

Example 17: *You are given the following equation* $Y = 126.55 + 18.04X + 1.786X^2$*, take the year of origin as 2002–2003. Shift the origin to 2003.*

Solution: Year of origin 2002–2003 means half of the annual increment *i.e.,* 0.5. After shifting the origin to 2003, the new equation is

$$Y_t = 126.55 + 18.04 (X + 0.5) + 1.786 (X + 0.5)^2$$
$$= 126.55 + 18.04X + 9.02 + 1.786(X^2 + 0.25 + X)$$
$$= 136.0165 + 19.826X + 1.786X^2.$$

9.6 MEASUREMENT OF SEASONAL VARIATIONS

Most of the phenomena in economics and business show seasonal patterns. But if the data is expressed annually then there is no seasonal variation. Only monthly or quarterly expressed data exhibit strong seasonal movements. To measure seasonal variations, it is necessary to first free the data from the effects of trend, cycles and irregular variation. A measure of seasonal variations which is free from the effects of trend, cycles and irregular variations etc. is known as **seasonal index**, the unit of seasonal index is percent.

There are different methods for measuring seasonal variations. Some of them are

1. Method of simple averages (weekly or monthly or quarterly)
2. Method of moving average
3. Ratio-to-trend method
4. Ratio-to-moving average
5. Link relative method.

We now discuss in details these methods.

I. ***Method of Simple Averages. (Weekly or monthly or quarterly)***

This is the simplest method of obtaining a seasonal index. The various steps involved are

Step I. Arrange the given data by years and monthwise or quarterwise (if quarterly data is given)

Step II. Obtain the totals of each month of quarter for different years. Obtain the average of each month or quarter.

Step III. Obtain an average of monthly averages by dividing the total of monthly averages by 12. In case of quarters, obtain an average of quarterly averages by dividing the total of quarterly averages by 4.

NOTES

Step IV. Compute the percentages of various monthly average (take the general average as base) by using the following formula

(a) Seasonal Index for any month

$$= \frac{\text{Monthly average of that month}}{\text{Average of monthly average}} \times 100$$

(b) Seasonal Index for any quarter

$$= \frac{\text{Average of that quarter}}{\text{General Average}} \times 100.$$

Example 18: *Assume that the trend is absent, determine, if any, seasonality in the data given below:*

Years	*Quarters*			
	I	*II*	*III*	*IV*
2002	*37*	*41*	*33*	*35*
2003	*37*	*39*	*36*	*36*
2004	*40*	*43*	*33*	*31*

Find the seasonal indices for each quarters by using method of simple averages.

Solution: **To find seasonal indices:** Consider the following table.

Years	**Quarters**			
	I	***II***	***III***	***IV***
2002	37	41	33	35
2003	37	39	36	36
2004	40	43	33	31
Total	114	123	102	102
Quarterly Average	$\frac{114}{3} = 38$	$\frac{123}{3} = 41$	$\frac{102}{3} = 34$	$\frac{102}{3} = 34$

Hence average of quarterly averages $= \frac{38 + 41 + 34 + 34}{4} = 36.75$

$\therefore$ Seasonal Index for quarter I $= \frac{38}{36.75} \times 100 = 103.4$

Seasonal Index for quarter II $= \frac{41}{36.75} \times 100 = 111.56$

Seasonal Index for quarter III $= \frac{34}{36.75} \times 100 = 92.52$

Seasonal Index for quarter IV $= \frac{34}{36.75} \times 100 = 92.52.$

Example 19: *Compute the seasonal averages, seasonal variations and seasonal indices for the following time series by using method of simple averages:*

Years	*Months*											
	Jan	*Feb*	*Mar*	*Apr*	*May*	*June*	*July*	*Aug*	*Sept*	*Oct*	*Nov*	*Dec*
2003	*15*	*16*	*18*	*18*	*23*	*23*	*20*	*28*	*29*	*33*	*33*	*38*
2004	*23*	*22*	*28*	*27*	*31*	*28*	*22*	*28*	*32*	*34*	*34*	*44*
2005	*25*	*25*	*35*	*26*	*36*	*30*	*30*	*34*	*38*	*47*	*41*	*53*

Solution: To find seasonal averages, seasonal variations, seasonal indices, consider the following table.

Months	**Years**			***3* yearly totals**	**Monthly average**	**Seasonal Index** $= \frac{\text{Monthly average}}{\text{Average of monthly average}} \times 100$
	2003	***2004***	***2005***			
Jan	15	23	25	15 + 23 + 25 = 63	$\frac{63}{3} = 21$	$\frac{21}{30} \times 100 = 70$
Feb	16	22	25	63	21	$\frac{21}{30} \times 100 = 70$
March	18	28	35	81	27	$\frac{27}{30} \times 100 = 90$
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
Aug	28	28	34	90	30	100
Sept	29	32	38	99	33	110
Oct	33	37	47	117	39	130
Nov	33	34	41	108	36	120
Dec	38	44	53	135	45	150
					360	

Here, Average of monthly average = $\frac{360}{12} = 30$.

Merits and Demerits of Method of Simple Averages

Merits

It is the simplest method as compared to all other methods of measuring seasonality.

NOTES

Demerits

It is not a very good method as it assumes that there is no trend component in the series.

II. ***Method of Moving Averages:*** It is a superior method as compared to method of simple averages. The various steps involved are

Step I. Compute moving average from the given data. If data are given on monthly basis, compute 12-monthly moving averages and in case of quarterly basis. Compute 4-quarterly moving averages.

Step II. Obtain trend values from the moving averages obtained in step I. We denote the trend values by 'T'.

Step III. Obtain short-term fluctuations by using trend values obtained in step-II.

Step IV. To find seasonal variations, use the formula:

Seasonal variations = Quarterly average – General average

where General average = Average of monthly or quarterly averages

Example 20: *Compute seasonal variations for the following data by using method of moving average.*

Years	*Quarters*			
	I	*II*	*III*	*IV*
2001	*30*	*81*	*62*	*119*
2002	*33*	*104*	*86*	*171*
2003	*42*	*153*	*99*	*221*
2004	*56*	*172*	*129*	*235*
2005	*67*	*201*	*136*	*302*

Solution: To find seasonal variations, consider the following table

Years	Quarters	*Y*	*4* quarterly moving totals	*2* period moving totals of *4* quarterly moving totals	Trend values (*T*) or quarterly moving average	Short-term fluctuations *Y–T*
2001	I	30	—	—	—	—
	II	81	—	—	—	—
			→ 30 + 81 + 62 + 119 = 292			
	III	62		→ 299 + 295 = 587	$\frac{587}{8}$ = 73.37	– 11.37
			→ 81 + 62 + 119 + 33 = 295			
	IV	119		→ 295 + 318 = 613	$\frac{613}{8}$ = 76.62	42.38
			→ 62 + 119 + 33 + 104 = 318			

(*Conted.*)

2002	I	33		$\rightarrow$ 318 + 342 = 660	$\frac{660}{8}$ = 82.5	– 49.5
			$\rightarrow$ 119 + 33 + 104 + 86 = 342			
	II	104		$\rightarrow$ 342 + 394 = 736	$\frac{736}{8}$ = 92	12
			$\rightarrow$ 33 + 104 + 86 + 171 = 394			
	III	86		$\rightarrow$ 394 + 403 = 797	$\frac{797}{8}$ = 99.62	13.62
			$\rightarrow$ 104 + 86 + 171 + 42 = 403			
	IV	171		$\rightarrow$ 403 + 452 = 855	$\frac{855}{8}$ = 106.87	64.38
			$\rightarrow$ 86 + 171 + 42 + 153 = 452			
2003	I	42		$\rightarrow$ 452 + 465 = 917	$\frac{917}{8}$ = 114.62	– 72.62
			$\rightarrow$ 171 + 42 + 153 + 99 = 465			
	II	153		$\rightarrow$ 465 + 515 = 980	$\frac{980}{8}$ = 122.5	30.5
			$\rightarrow$ 42 + 153 + 99 + 221 = 515			
	III	99		$\rightarrow$ 515 + 529 = 1044	$\frac{1044}{8}$ = 130.5	– 37.5
			$\rightarrow$ 153 + 99 + 221 + 56 = 529			
	IV	221		$\rightarrow$ 529 + 548 = 1077	$\frac{1077}{8}$ = 134.62	86.38
			$\rightarrow$ 99 + 221 + 56 + 172 = 548			
2004	I	56		$\rightarrow$ 548 + 578 = 1126	$\frac{1126}{8}$ = 140.75	– 84.75
			221 + 56 + 172 + 129 = 578			
	II	172		$\rightarrow$ 578 + 592 = 1170	$\frac{1170}{8}$ = 146.25	25.75
			56 + 172 + 129 + 235 = 592			
	III	129		$\rightarrow$ 592 + 603 = 1195	$\frac{1195}{8}$ = 149.37	– 20.37
			172 + 129 + 235 + 67 = 603			
	IV	235		$\rightarrow$ 603 + 632 = 1235	$\frac{1235}{8}$ = 154.37	80.63
			129 + 235 + 67 + 201 = 632			
2005	I	67		$\rightarrow$ 632 + 639 = 1271	$\frac{1271}{8}$ = 158.87	– 91.87
			235 + 67 + 201 + 136 = 639			
	II	201		$\rightarrow$ 639 + 706 = 1345		
			67 + 201 + 136 + 302 = 706		$\frac{1345}{8}$ = 168.12	32.88
	III	136	—	—		
	IV	302	—	—		

Hence, average of quarterly averages

$$= \frac{-74.68 + 25.28 - 19.215 + 68.44}{4}$$

$$= \frac{-0.175}{4} = -0.04375$$

$\therefore$ Seasonal variations for quarter I

$$= -74.68 - (-0.04375) = 74.63$$

Seasonal variation for quarter II

$$= 25.28 - (-0.04375) = 25.3237$$

Seasonal variation for quarter III

$$= -19.215 - (-0.04375) = -19.17125$$

Seasonal variation for quarter IV

$$= 68.44 - (-0.04375) = 68.48375.$$

III. ***Ratio-to-trend Method:*** It is a method of calculate seasonal index. This method is also known as percentage-to-trend method. The various steps involved are

Step I. Find trend values (T) monthwise or quarterly by the method of least squares.

Step II. Denote the original value by Y. Then find ratio-to-trend value for each period where

$$\text{Ratio-to-trend} = \frac{Y}{T} \times 100$$

Step III. Obtain the A.M. of each quarterly or monthly period ratio-to-trend.

Step IV. The seasonal index is given by seasonal index

$$= \frac{\text{Quarterly (or monthly) average}}{\text{General average}} \times 100.$$

Example 21: *By using Ratio-to-trend method, compute seasonal indices for each quarter of the following data.*

Years	***Quarters***			
	I	***II***	***III***	***IV***
2001	*30*	*40*	*36*	*34*
2002	*34*	*52*	*50*	*44*
2003	*40*	*58*	*54*	*48*
2004	*54*	*76*	*68*	*62*
2005	*80*	*92*	*86*	*82*

Solution: We first find the trend values by using method of least squares. Consider the following table. Here, $n = 5$(odd) $\therefore$ year of origin is 2003

Years	Yearly totals	Yearly average Y	Deviations from 2003 X	XY	X^2
2001	30 + 40 + 36 + 34 = 140	$\frac{140}{4} = 35$	– 2	– 70	4
2002	34 + 52 + 50 + 44 = 180	$\frac{180}{4} = 45$	– 1	– 45	1
2003	40 + 58 + 54 + 48 = 200	$\frac{200}{4} = 50$	0	0	0
2004	54 + 76 + 68 + 62 = 260	$\frac{260}{5} = 65$	1	65	1
2005	80 + 90 + 86 + 82 = 340	$\frac{340}{5} = 85$	2	170	4
		$\Sigma Y = 280$	$\Sigma X = 0$	$\Sigma XY = 120$	$\Sigma X^2 = 10$

We first find the trend values yearly and will it convert it to quarterly. The equation of the straight line is

$$Y = a + bX \qquad ...(1)$$

To find *a, b*. The normal equations are

$$\Sigma Y = 5a + b\Sigma X$$

and $$\Sigma XY = a\Sigma X + b\Sigma X^2$$

or $$280 = 5a \quad \Rightarrow \quad a = \frac{280}{5} = 56$$

and $$120 = 10b \quad \Rightarrow \quad b = 12$$

$\therefore$ From (1), $$Y = 56 + 12X \qquad ...(2)$$

To find the trend values. Consider the following table.

Year	X	Trend values (Y_c) $Y = 56 + 12X$
2001	– 2	32
2002	– 1	44
2003	0	56
2004	1	68
2005	2	80

Conversion to quarterly trend values. From (2), b = yearly increment = 12

$\Rightarrow$ quarterly increment = $\frac{12}{4} = 3$

NOTES

Consider the year 2001. Trend values for the middle quarter *i.e.,* half of the II quarter and half of the III quarter is 32. Also quarterly increment is 3. Therefore, the trend value of quarter II is $32 - \frac{3}{2} = 30.5$ and for quarter III is $32 + \frac{3}{2} = 33.5$. Similarly trend value for the quarter I is 30.5 – 3 = 27.5 and for the quarter IV is 33.5 + 3 = 36.5. Thus quarterly trend values for each year is given in the following table.

Trend Values

Years	Quarters			
	I	***II***	***III***	***IV***
2001	27.5	30.5	33.5	36.5
2002	39.5	42.5	45.5	48.5
2003	51.5	54.5	57.5	60.5
2004	63.5	66.5	69.5	72.5
2005	75.5	78.5	81.5	84.5

Quarterly values as % of trend values. Consider the following table.

Years	Quarters			
	I	***II***	***III***	***IV***
2001	$\frac{30}{27.5} \times 100 = 109.09$	$\frac{40}{30.5} \times 100 = 131.14$	107.46	93.15
2002	$\frac{34}{39.5} \times 100 = 86.07$	$\frac{52}{42.5} \times 100 = 122.35$	109.89	90.72
2003	$\frac{40}{51.5} \times 100 = 77.66$	$\frac{58}{54.5} \times 100 = 106.42$	93.91	79.34
2004	$\frac{54}{63.5} \times 100 = 85.03$	$\frac{76}{66.5} \times 100 = 114.28$	97.84	85.52
2005	$\frac{80}{75.5} \times 100 = 105.96$	$\frac{92}{78.5} \times 100 = 117.19$	105.52	97.04
Total	463.81	591.38	514.62	445.77
Quarterly Average	$\frac{463.84}{5} = 92.77$	$\frac{591.38}{5} = 118.28$	$\frac{514.62}{5} = 102.92$	$\frac{445.77}{5} = 89.15$

Now the general average $= \frac{92.77 + 118.28 + 102.92 + 89.15}{4}$

$= \frac{403.12}{4} = 100.78$

$\therefore$ Seasonal Index for quarter I $= \frac{92.77}{100.78} \times 100 = 92.05$

Seasonal Index for quarter II $= \frac{118.28}{100.78} \times 100 = 117.36$

Seasonal Index for quarter III $= \frac{102.29}{100.78} \times 100 = 101.49$

Seasonal Index for quarter IV $= \frac{89.15}{100.78} \times 100 = 88.46.$

IV. ***Ratio-to-moving Average Method:*** It is the most commonly used method to measure seasonal variations. It is also known as **percentages of moving average** method. The various steps involved are

Step I. Obtain the trend values by applying **moving average method.** If the data are quarterly, find 4-quarterly moving averages. In case of monthly data, find 12 monthly moving averages.

Step II. Find ratio-to-moving averages by using the formula

$$\text{Ratio-to-moving average} = \frac{Y}{T} \times 100,$$

where Y = original value

T = moving average.

Step III. Obtain quarterly averages after arranging the ratio-to-moving averages corresponding to different periods.

Step IV. Obtain general average.

Step V. Obtain seasonal indices by using the formula.

$$\text{Seasonal Index} = \frac{\text{Quarterly average}}{\text{General average}} \times 100$$

Example 22: *Compute seasonal indices by Ratio-to-moving average method from the following data:*

Years	*Quarters*			
	I	*II*	*III*	*IV*
2004	*68*	*62*	*61*	*63*
2005	*65*	*58*	*66*	*61*
2006	*68*	*63*	*63*	*67*

Solution: We first find the trend values by using *moving average method*. Consider the following table.

Years	Q	Given value Y	4 quarterly moving totals	2 quarterly moving totals	Trend values or 4 quarterly moving average T	Ratio-to-moving average $= \frac{Y}{T} \times 100$
2004	I	68	—	—	—	—
	II	62	—	—	—	—
			68 + 62 + 61 + 63 = 254			
	III	61		505	$\frac{505}{8} = 63.125$	$\frac{61}{63.125} \times 100 = 96.63$
			62 + 61 + 63 + 65 = 251			
	IV	63		498	$\frac{498}{8} = 62.25$	$\frac{63}{62.25} \times 100 = 101.21$
			61 + 63 + 65 + 58 = 247			
2005	I	65		499	$\frac{499}{8} = 62.375$	$\frac{65}{62.375} \times 100 = 104.2$
			63 + 65 + 58 + 66 = 252			
	II	58		502	$\frac{502}{8} = 62.75$	$\frac{58}{62.75} \times 100 = 92.43$
			65 + 58 + 66 + 61 = 250			
	III	66		503	$\frac{503}{8} = 62.875$	$\frac{66}{62.875} \times 100 = 104.97$
			58 + 66 + 61 + 68 = 253			
	IV	61		511	$\frac{511}{8} = 63.875$	$\frac{61}{63.875} \times 100 = 95.49$
			66 + 61 + 68 + 63 = 258			
2006	I	68		513	$\frac{513}{8} = 64.125$	$\frac{68}{64.125} \times 100 = 106.04$
			61 + 68 + 63 + 63 = 255			
	II	63		516	$\frac{516}{8} = 64.5$	$\frac{63}{64.5} \times 100 = 97.67$
			68 + 63 + 63 + 67 = 261			
	III	63	—	—	—	—
	IV	67	—	—	—	—

We now obtain quarterly average for each quarter. Consider the following table :

Years	Quarters			
	I	*II*	*III*	*IV*
2004	—	—	96.63	101.2
2005	104.2	92.43	104.97	95.49
2006	106.04	97.67	—	—
Totals	210.24	190.1	201.16	196.69
Quarterly Average	$\frac{210.24}{2}$ = 105.12	$\frac{190.1}{2}$ = 95.05	$\frac{201.16}{2}$ = 100.58	$\frac{196.69}{2}$ = 98.345

$$\therefore \quad \text{General average} = \frac{\text{Quarterly average}}{4}$$

$$= \frac{105.12 + 95.05 + 100.58 + 98.345}{4}$$

$$= \frac{399.095}{4} = 99.77$$

$$\therefore \quad \text{Seasonal Index for quarter I} = \frac{105.12}{99.77} \times 100 = 105.36$$

$$\text{Seasonal Index for quarter II} = \frac{95.05}{99.77} \times 100 = 95.26$$

$$\text{Seasonal Index for quarter III} = \frac{100.58}{99.77} \times 100 = 100.81$$

$$\text{Seasonal Index for quarter IV} = \frac{98.345}{99.77} \times 100 = 98.57$$

Merits and Demerits of Ratio-to-trend Method

Merits

1. It is a simple method and easy to compute and understand.

2. For measuring seasonal variations, it is more logical as compared to method of monthly average. For, it has a ratio-to-trend value for each month for which data are available. Hence there is no loss of data as occurs in the case of moving averages.

Demerits

In case of cyclic fluctuations: The trend (a straight line or a curve) does not follow the actual data as closely as a 12-month moving average does. Hence the seasonal index so obtained is more biased as compared to ratio-to-moving average method.

V. ***Link Relative Method:*** It is the most difficult method as compared to all other methods of measuring seasonal variations. The various steps involved are

Step I. Obtain link relatives of the seasonal figure monthly or quarterly by using the formula

$$\text{Link relative} = \frac{\text{Current season's figure}}{\text{Previous season's figure}} \times 100$$

where the word 'season' means time period. In case of monthly data, season refers to a 'month' and in case of quarterly data to a 'quarter'.

Step II. Compute average of the link relatives for each month or quarter.

Step III. Convert the averages obtained in step II into chain relatives by using the formula.

Chain relatives

$$= \frac{[\text{Average of link relative of the current season's figure}] \times [\text{Chain relative of the previous seasons's figure}]}{100}$$

Assume chain relative of the first quarter = 100

Step IV. Calculate chain relative for the first term on the basis of chain relative of the last season's figure by using the formula

Chain relatives of the first term =

$$\frac{[\text{Chain relative of the last seasons's figure}] \times [\text{Average of link relative of the first season's figure}]}{100}$$

Step V. Correct chain related. Chain relative of the first period should be 100, But due to influence of the trend, it can be more or less than 100. Obtain the difference by deducing 100 from the revised chain relative of the first term. Divide this difference by the number of periods and multiply the quotient by 1, 2, 3 and so on. Subtract these values from the chain relative of second term, chain relative of third term chain relative of fourth term and so on.

The above process is called *correcting of chain relative.*

Step VI. Compute A.M. of the corrected chain relatives.

Step VII. Obtain seasonal index by using the formula.

$$\text{Seasonal Index} = \frac{\text{Correct chain relatives}}{\text{General average}} \times 100$$

Example 23: *Obtain seasonal indices by using link relative method for the following data:*

Link Relative

Years	*Quarters*			
	I	*II*	*III*	*IV*
2001	—	*120*	*133*	*83*
2002	*80*	*117*	*113*	*89*
2003	*88*	*129*	*111*	*93*
2004	*80*	*125*	*115*	*96*
2005	*83*	*117*	*120*	*79*

Solution: Since link relatives is given to us, we find average of the link relatives for each quarter.

Now we find chain relatives by using the formula:

Chain relatives =

$$\frac{[\text{Average of link relative of the current season's figure}] \times [\text{Chain relative of the previous season's figure}]}{100}$$

[Assume chain relative of the first quarter = 100] the chain relatives so obtained are shown in the following table.

Years	Quarters			
	I	*II*	*III*	*IV*
2001	—	120	133	83
2002	80	117	113	89
2003	88	129	111	93
2004	80	125	115	96
2005	83	117	120	79
Total of link of relatives	331	608	592	440
Average of link relatives	$\frac{331}{4} = 82.75$	$\frac{608}{5} = 121.6$	$\frac{592}{5} = 118.4$	$\frac{440}{5} = 88$
Chain relatives	100	$\frac{121.6 \times 100}{100} = 121.6$	$\frac{118.4 \times 121.6}{100} = 143.97$	$\frac{88 \times 143.97}{100} = 126.69$
Corrected chain relatives	100	$121.6 - 1 \times 1.21 = 120.39$	$143.97 - 2 \times 1.21 = 141.55$	$126.69 - 3 \times 1.21 = 123.06$

We next compute corrected (or adjusted) chain relatives (see the table). For this, chain relative of the first quarter (on the basis of last quarter) =

$$\frac{[\text{Chain relative of the last season's figure}] \times [\text{Average of the link relative of the first season's figure}]}{100}$$

$$= \frac{126.69 \times 82.75}{100} = 104.84$$

Difference between chain relative of the first quarter based on the last quarter and chain relative of the first quarter based on the first quarter

$$= 104.84 - 100 = 4.84$$

$$\Rightarrow \quad \text{Difference per quarter} = \frac{4.84}{4} = 1.21.$$

NOTES

∴ Corrected (or adjusted) chain relatives for each quarter are obtained by subtracting 1×1.21, 2×1.21, 3×1.21, from the chain relatives of the II, III and IV quarters respectively. (See the table).

Finally, general average (Average of corrected chain relatives)

$$= \frac{100 + 120.39 + 141.55 + 123.06}{4} = 121.25$$

∴ Seasonal Index for quarter I

$$= \frac{\text{Corrected chain relative of quarter I}}{\text{General average}} \times 100$$

$$= \frac{100}{121.25} \times 100 = 82.47$$

Seasonal Index for quarter II $= \frac{120.39}{121.25} \times 100 = 99.29$

Seasonal Index for quarter III $= \frac{141.55}{121.25} \times 100 = 116.74$

Seasonal Index for quarter IV $= \frac{123.06}{121.25} \times 100 = 101.5.$

Example 24: *Apply the method of link relatives to the following data and calculate seasonal indices.*

Years	***Quarters***			
	I	***II***	***III***	***IV***
2003	*6*	*6.5*	*7.8*	*8.7*
2004	*5.4*	*7.9*	*8.4*	*7.3*
2005	*6.8*	*6.5*	*9.3*	*6.4*
2006	*7.2*	*5.8*	*7.5*	*8.5*
2007	*6.6*	*7.3*	*8*	*7.1*

Solution: We first find the link relatives of each quarter by using the formula.

$$\text{Link relative} = \frac{\text{Current season's figure}}{\text{Previous year's figure}} \times 100$$

The link relatives of each figure is shown in the following Table.

Link Relatives

Years	Quarters			
	I	*II*	*III*	*IV*
2003	—	$\frac{6.5}{6} \times 100 = 108.33$	$\frac{7.8}{6.5} \times 100 = 120$	$\frac{8.7}{7.8} \times 100 = 111.53$
2004	$\frac{5.4}{8.7} \times 100 = 62.06$	$\frac{7.9}{5.4} \times 100 = 146.29$	$\frac{8.4}{7.9} \times 100 = 106.32$	$\frac{7.3}{8.4} \times 100 = 86.9$
2005	$\frac{6.8}{7.3} \times 100 = 93.15$	$\frac{6.5}{6.8} \times 100 = 95.58$	$\frac{9.3}{6.5} \times 100 = 143.07$	$\frac{6.4}{9.3} \times 100 = 68.81$
2006	$\frac{7.2}{6.4} \times 100 = 112.15$	$\frac{5.8}{7.2} \times 100 = 80.55$	$\frac{7.5}{5.8} \times 100 = 129.31$	$\frac{8.5}{7.5} \times 100 = 113.33$
2007	$\frac{6.6}{7.1} \times 100 = 77.64$	$\frac{7.3}{6.6} \times 100 = 110.6$	$\frac{8}{7.3} \times 100 = 109.58$	$\frac{7.1}{8} \times 100 = 88.75$
Average of link relatives	$\frac{345.35}{4} = 86.33$	$\frac{541.35}{5} = 108.27$	$\frac{608.28}{5} = 121.65$	$\frac{469.32}{5} = 93.86$
Chain relatives	100	$\frac{100 \times 108.27}{100}$ = 108.27	$\frac{108.27 \times 121.65}{100}$ = 131.71	$\frac{131.71 \times 93.86}{100}$ = 123.62
Corrected chain relatives	100	108.27 – 1.68 = 106.59	131.71 – 2 × 1.68 = 128.35	123.62 – 3 × 1.68 = 118.58

Calculation of Corrected Chain Relatives

Chain relative of quarter I on the basis of first quarter = 100

Chain relative of quarter I on the basis of last quarter = $\frac{86.33 \times 123.62}{100} = 106.72$

Difference between these chain relatives = 106.72 – 100 = 6.72

$\Rightarrow$ Difference per quarter = $\frac{6.72}{4} = 1.68$

$\therefore$ Corrected (or adjusted) chain relatives are obtained by subtracting 1 × 1.68, 2 × 1.68, 3 × 1.68 from the chain relatives of quarters II, III and IV respectively.

Hence average of corrected chain relatives (or General Average)

$$= \frac{100 + 106.59 + 128.35 + 118.58}{4} = \frac{453.52}{4} = 113.38$$

NOTES

$\therefore$ Seasonal Index for quarter I

$$= \frac{\text{Corrected chain relatives of quarter}}{\text{General average}} \times 100$$

$$= \frac{100 \times 100}{113.38} = 88.19$$

Seasonal Index for quarter II $= \frac{106.59}{113.38} \times 100 = 94.01$

Seasonal Index for quarter III $= \frac{128.35}{113.38} \times 100 = 113.2$

Seasonal Index for quarter IV $= \frac{118.58}{113.38} \times 100 = 104.58.$

Check Your Progress

State whether the following statements are True or False:

6. Method of least squares can also be used to fit parabolic trend or exponential trend.
7. Equation of the exponential curve is
   $Y = a + bX.$
8. Measure of seasonal variations is free from the effects of trend.
9. Formula for finding seasonal variations is
   Seasonal variations = Quarterly average + General average
10. Ratio-to-trend method is also known as percentage-to-trend method.

9.7 SUMMARY

- 'A time series is a set of observations taken at specified times, usually at equal intervals.'
- The term **'Trend',** is the basic tendency of production, sales and income etc. to grow or decline over a period of time. Trend does not include short-range oscillations, it includes steady movements over a long period of time.
- The trends that occur as a result of general tendency of the data to increase or decrease, over a long period of time, are known as *secular trends*.
- The trends that take place during a period of 12 months as a result of change in climate, weather conditions etc. are called *seasonal variations* or season

variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself.

- It is the most important factor causing seasonal variations. The change in the climate and weather conditions such as rainfall, humidity, etc. effects the different products differently.
- Cyclic variations refer to the oscillatory variations in a time series which have a duration anywhere between 2 to 10 years. A *business cycle* has four phases namely (*i*) **Prosperity,** (*ii*) **Recession,** (*iii*) **Depression** (*iv*) **Recovery.**
- These business activities start recessing or falling and come to the lowest limit of decline. This level is called **depression level.**
- *Irregular variations* take place due to special causes like floods, earth quakes, strikes and wars etc.
- The curve so obtained is called **freehand curve** and this method is also called trend **fitting by inspection.**
- **Semi-average Method.** In this method, the given data is divided into two parts.
- 'Semi-average method' can be applied in two situations given below:

  (a) When the number of years given is even

  (b) When the number of years given is odd.
- **Moving Average Method.** In this method, we compute moving averages such as 3 yearly moving average, 4 yearly moving average, 5 yearly moving average, etc.
- If the moving average is an even period say, four yearly or six yearly; The moving averages are placed at the centre of the time span.
- The term 'Short-term fluctuations' may be defined as the difference of trend values and the original data.
- To fit a straight line trend by the method of least squares, the first step is to assign any year as the **year of origin.**
- A measure of seasonal variations which is free from the effects of trend, cycles and irregular variations etc. is known as **seasonal index**, the unit of seasonal index is per cent.
- To find seasonal variations, use the formula :

  Seasonal variations = Quarterly average – General average
- **Ratio-to-Trend Method.** It is a method of calculate seasonal index. This method is also known as percentage-to-trend method.

9.8 GLOSSARY

- ***'Trend'*:** is the basic tendency of production, sales and income etc. to grow or decline over a period of time.
- ***Secular Increase in Sale:*** If we study the trend of sales of cars over a period of 20 years, and find that except for a year or two, the sales are increasing continuously, we will call it.
- ***Season Variations:*** Season variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself.
- ***Erratic Variations*** The variations in business activity which do not repeat in a definite pattern are called.
- ***Year of Origin:*** To fit a straight line trend by the method of least squares, the first step is to assign any year as the **year of origin.**

9.9 ANSWERS TO CHECK YOUR PROGRESS

1. quantitative data
2. depression level
3. definite pattern
4. trend fitting by inspection
5. trade cycles, business cycles.
6. True
7. False
8. True
9. False
10. True

9.10 TERMINAL AND MODEL QUESTION

1. What is meant by time series ? Discuss its importance in business.
2. Explain the meaning and importance of time series.
3. Discuss the various components of time series.
4. Distinguish between secular trend and periodic movement of time series.

5. Fit a trend line for the following data, using 'Free hand curve method'.

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004
Sales (in lakhs)	22	28	24	30	18	26	20	32	16

6. Fit a trend line by the method of semi-average to the following data:

Year	1998	1999	2000	2001	2002	2003
Production ('000 units)	22	26	24	30	28	32

7. Calculate the trend values by the method of 4-yearly moving averages for the following data.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Production	464	515	518	467	502	540	557	571	586	612

8. Calculate 3-yearly moving averages of the production figures given below and draw the trend.

Year	**Production**	**Year**	**Production**
1991	15	1999	63
1992	21	2000	70
1993	30	2001	74
1994	36	2002	82
1995	42	2003	90
1996	46	2004	95
1997	50	2005	102
1998	56		

9. Fit a straight line trend by the method of least square to the following data.

Years	1999	2000	2001	2002	2003	2004	2005	2006
Sales (in lakhs)	38	40	65	72	69	60	87	95

Assume that the same rate of change will continue, predict the sale for the year 2008. Also plot the trend value on the graph and show the trend line.

10. The sales of a company in lakhs varied from 1998 to 2005 in the following manner.

Years	1998	1999	2000	2001	2002	2003	2004	2005
Sales (₹ Lakhs)	412	438	444	454	470	482	490	500

Fit a trend line by the method of semi-average. Estimate the sales for the year 2006, if the actual sales for that year is ₹520 lakhs. Justify the difference between the two figures.

11. The sale of a commodity in tonnes varied from January 2005 to December 2005 in the manner given below:

Jan.	Feb.	March	April	May	June	July	Aug.	Sep.	Oct.	Non.	Dec.
280	300	280	280	270	240	230	230	220	200	210	200

Fit a trend line by using semi-average method.

12. Calculate the seasonal index for the following data by using simple average method assuming trend is absent:

Years	**Ist Quarter**	**2nd Quarter**	**3rd Quarter**	**4th Quarter**
1987	3.7	4.1	3.3	3.5
1988	3.7	3.9	3.6	3.6
1989	4.0	4.1	3.3	3.1
1990	3.3	4.4	4.0	4.0

13. Calculate the seasonal index for the following data by using simple average method for the following data:

Years	**Summer**	**Monsoon**	**Autumn**	**Winter**
1981	112	110	120	115
1982	80	145	105	90
1983	95	100	140	80
1984	110	90	130	100

14. From the following data, calculate 3-yearly, 5-yearly and 7-yearly moving averages and plot the data on a graph paper.

Year	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**
Cyclic fluctuations	+ 2	+ 1	0	– 2	– 1	+ 2	+ 1

Year	1998	1999	2000	2001	2002	2003	2004	2005
Cyclic fluctuations	0	– 2	– 1	+ 2	+ 1	0	– 2	– 1

15. Using three-yearly moving averages, determine the 'trend' and short-term fluctuations. Plot the original data and trend values on the same graph paper.

Year	**Production (in '000 tonnes)**	**year**	**Production ('000 tonnes)**
1997	21	2002	22
1998	22	2003	25
1999	23	2004	27
2000	25	2005	27
2001	24	2006	26

16. Determine the period of moving average for the data given below

Year	1980	1981	1982	1983	1984	1985	1986	1987
Sales (in lakhs)	130	127	124	135	140	132	129	127

Year	1987	1988	1989	1990	1991	1992	1993
Sales (in lakhs)	145	158	153	146	145	164	170

17. Calculate the trend values of the following data by using 4-yearly moving average

Year	**Value**	**Year**	**Value**
1992	41	2000	67
1993	61	2001	73
1994	55	2002	78
1995	48	2003	76
1996	53	2004	84
1997	67	—	—
1998	62	—	—
1999	60	—	—

18. Consider the trend equation $Y = 280 + 10X + 0.6X^2$, where origin = 1976, X unit = 1 year, Y = annual production of steel. Sift the origin to 1979–79.
19. Consider the trend equation $Y = 25\ (2.8)^X$, where origin = 2003, X unit = 1 year Shift the origin forward by 2 years.

NOTES

20. Consider the trend equation Y = 20 + 0.8X, where origin = 2002, X units = one year, Y units = production in million tonnes.

    Shift the origin to January 1, 2003.

9.11 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

MS 104

Quantitative Techniques in Management

Block - III
Block Title : Probability and Distribution

UTTARAKHAND OPEN UNIVERSITY
SCHOOL OF MANAGEMENT STUDIES AND COMMERCE
University Road, Teenpani By pass, Behind Transport Nagar, Haldwani- 263 139
Phone No: (05946)-261122, 261123, 286055
Toll Free No.: 1800 180 4025
Fax No.: (05946)-264232, e-mail: info@uou.ac.in, som@uou.ac.in
http://www.uou.ac.in
www.blogsomcuou.wordpress.com

Board of Studies

Programme Coordinator

Dr. Manjari Agarwal
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

Units Written By		Unit No.
Text material developed by	Devashish Dutta	
Typeset by	Goswami Associates, Delhi	

Editor(s)

Dr. Hitesh Kumar Pant
Assistant Professor
Department of Management Studies
Kumaun University
Bhimtal Campus

Dr. Manjari Agarwal
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

Er. Sumit Prasad
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

ISBN : 978-93-85740-10-7

Edition : 2016 (Restricted Circulation)
Published by : Uttarakhand Open University, Haldwani, Nainital - 263 139
Printed at : Laxmi Publications (P) Ltd., New Delhi
DUO-8158-67.26-QUAN TECH MGMT B-III

CONTENTS

UNIT 10: PROBABILITY — DEFINITION AND CLASSIFICATION

Structure

10.0 INTRODUCTION

The word 'Probability' and 'Chance' are quite familiar to everyone. Many a times, we come across statements like, "Probably it may rain today", "chances of hitting the target are very few". "It is possible that he may top the examination". In the above statements, the probably, chances, possible, etc. convey the sense of uncertainty about the occurrence of some event. Ordinarily, it appears that there cannot be any exact measurement for these uncertainties, but in Mathematical Statistics, we have methods for calculating the degree of certainty of events in numerical value, under certain conditions. When, we perform experiments in science and engineering, repeatedly under identical conditions, we get almost the same result. There also exist experiment in which the outcome may be different even if the experiment is performed under identical conditions. In such experiments, the outcome of each experiment depends on chance.

If an experiment is repeated under essentially homogeneous and similar conditions, we generally come across two types of situations :

(i) The result or what is usually known as the 'outcome' is unique or certain.

(ii) The result is not unique but may be one of the several possible outcomes.

The phenomena covered by (i) are known as 'deterministic' or 'predictable' phenomena. By a deterministic phenomenon we mean one in which the result can be predicted with certainty. For example,

(a) For a perfect gas,

$$V \propto \frac{1}{P}, \quad i.e., \quad PV = \text{constant},$$

provided the temperature remains same.

(b) The velocity 'v' of a particle after time t is given by

$$v = u + at,$$

where u is the initial velocity and a is the acceleration. This equation uniquely determines v if the right-hand quantities are known.

(c) Ohm's Law,

$$C = \frac{E}{R},$$

where C is the flow of current, E the potential difference between the two ends of the conductor and R the resistance, uniquely determines the value of C as soon as E and R are given.

A deterministic model is defined as a model which stipulates that the condition under which an experiment is performed determines the outcome of the experiment. For a number of situations the deterministic model suffices. However, there are phenomena [as covered by (ii) above] which do not lend themselves to deterministic approach and are known as 'unpredictable' or 'probabilistic' phenomena. For example,

(a) In tossing of a coin one is not sure if a head or a tail will be obtained.

(b) If a light-tube has tested for t hours, nothing can be said about its further life. It may fail to function any moment.

In such cases we talk of chance or probability which is taken to be a quantitative measure of certainty.

The concept of probability is basic to many principles of classical genetics, *e.g.*, segregation, independent assortment etc. However, the principles of population genetics are based on the consideration of laws of probability.

Short History

The theory of probability has its origin in the games of chance related to gambling. It was in the first half of the sixteenth century that serious thought was given to the

problem of probability by eminent mathematicians. An Italian mathematician, J. Cardon was the first man to write a book titled ''Book on games of chance'' in 1663. However, the classical treatment of probability dates back to the seventeenth century after the work of two mathematicians, B. Pascal (1623–62) and P. Fermat (1601–65). Much of this theory developed out of attempts to solve problems related to games of chance such as rolling of dice. J. Bernoulli (1654–1705) also extended the application of the theory of probability much beyond the restricted field of games of chance. In the nineteenth century, P.S. Laplace (1749–1827) compiled the first general theory of probability. R.A. Fisher and Von Mises introduced the empirical approach to probability. The modern theory of probability was developed by Russian mathematicians like Chebyshev (1821–94), who founded the Russian school of statisticians; A. Markoff (1856–1922) and A. Kolmogorov (1933). The theory of probability has been used very extensively in all disciplines of sciences. Now it becomes difficult to discuss biostatistics without an understanding of the meaning of probability. A knowledge of probability theory makes it possible to interpret biostatistical results and we can express numerically the inevitable uncertainties in the resulting conditions.

Mathematical (or Classical) Definition of Probability: If an event can happen in n ways which are equally likely, exhaustive and mutually exclusive and out of these n ways, m ways are favourable to an event A, then the probability of happening of A is given by

$$p \text{ or } P(A) = \frac{m}{n}$$

If A happens in m ways, it will fail in $(n - m)$ ways so that the probability of its failure

$$q \quad \text{or} \quad P(\overline{A}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p$$

$$\Rightarrow \qquad p + q = 1 \quad i.e., \quad P(A) + P(\overline{A}) = 1$$

$$0 \le p \le 1 \; ; \; 0 \le q \le 1$$

If P(A) = 1, then A is called a certain event. If P(A) = 0, then A is called an impossible event.

Statistical (or Empirical) Definition of Probability: If in n trials, an event A happens m times then the probability of happening A is given by

$$p \text{ or } P(A) = \lim_{n \to \infty} \frac{m}{n}.$$

NOTES

10.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define the term 'Probability'
- Define various terms related with probability
- Explain various mathematical tools related with probability
- Explain algebra of an event and probability of an event
- Classify probability into conditional and unconditional

10.2 DEFINITION OF IMPORTANT TERMS

1. **Experiment:** Any operation that results in two or more outcomes is called an experiment, and performing of an experiment is called trial.

2. **Random Experiment:** A random experiment is defined as an experiment in which all possible outcomes are known and which can be repeated under identical conditions but it is not possible to predict the outcome of any particular trial in advance.

e.g. Tossing a coin or throwing a die is random experiment.

3. **Sample Space:** The sample space of a random experiment is defined as the set of all possible outcomes of the experiment. The possible outcomes are called sample points. The sample space is generally denoted by the letter S.

e.g. In throwing a fair die, sample space is S = {1, 2, 3, 4, 5, 6}. In tossing of two unbiased coins sample space is S = {HH, HT, TH, TT}.

4. **Event:** Any subset of the sample space is defined as an event. An event is called an elementary (or simple) event if it contains only one sample point. In the experiment of throwing a die, the event A of getting 2 is a simple event. We write A = {2}. Also an event is called an impossible event if it can never occur. In the above experiment, event B = {7} of getting 7 is an impossible event. An event which is sure to occur is called a certain event.

e.g. In throwing a die, the event of getting a number less than 7 is a certain event.

5. **Exhaustive Events:** The total number of all possible outcomes in any trial are known as exhaustive events or cases.

e.g. In tossing a coin, there are two exhaustive events, head and tail. In throwing a die, there are 6 exhaustive cases, any one of the six faces may turn up.

For examples,

(i) In tossing an unbiased or uniform coin, head and tail are equally likely events.

(ii) In throwing an unbiased die, all the six faces are equally likely to come.

6. **Independent Events:** Several events are said to be 'independent' if the happening (or non-happening) of an event is not affected by the happening (or non-happening) of the other events. For examples,

(i) The sex of a calf to be born is independent of the sex of the calf in the previous calfing by the same dam.

(ii) In tossing an unbiased coin the event of getting a head in first toss is independent of getting a head in the second, third and subsequent throws.

(iii) If we draw a card from a pack of well-shuffled cards and replace it before drawing the second card, the result of second draw is independent of the first draw. But, however, if the first card drawn is not replaced then the second draw is dependent on the first draw.

7. **Favourable Events or Cases:** The number of cases 'favourable' to an event in a trial is the number of outcomes which entail the happening of the event. For examples,

(i) In drawing a card from a pack of cards the number of cases favourable to drawing of an ace is 4, for drawing a spade is 13 and for drawing a red card is 26.

(ii) In throwing of two dice, the number of cases favourable to getting a sum 5 is: (1, 4), (4, 1), (2, 3), (3, 2), *i.e.,* 4.

8. **Mutually Exclusive events:** Events are said to be 'mutually exclusive' if the happening of any one of them precludes the happening of all others, *i.e.,* if no two or more of them can happen simultaneously in the same trial. For examples,

(i) In throwing a die all the 6 faces numbered 1 to 6 are mutually exclusive since if any one of these faces comes, the possibility of others, in the same trial, is ruled out.

(ii) In tossing a coin the events head and tail are mutually exclusive.

(iii) Similarly, a new born can be either a boy or a girl.

9. **Equally Likely Events:** Events are said to be equally likely, if there is no reason to expect any one in preference to any other.

e.g. If we draw a card from a well-shuffled pack, we may get any card, then the 52 different cases are equally likely.

10. **Compound Events:** Events obtained by combining together two or more elementary events are known as the compound events.

e.g. In throwing a die, getting 5 or 6 is called a compound event.

Check Your Progress

Fill in the blanks:

1. If P(A) = 0 then A is called an .....
2. The total number of all possible outcomes in any trial are known as .....
3. Tossing a coin or throwing a die is ..... experiment.
4. Any subset of the sample space is defined as an .....
5. Events obtained by combining together two or more elementary events are known as .....

10.3 MATHEMATICAL TOOLS

Preliminary Notions of Sets

The set theory was developed by the German mathematician, G. Cantor (1845–1918).

Sets and elements of sets: A set is a well defined collection or aggregate of all possible objects having given properties and specified according to a well defined rule. In other words, a set is a collection of well defined objects. The objects comprising a set are called elements, members or points of the set. Sets are often denoted by capital letters, viz., *A, B, C* etc. and the elements are denoted by small letters, viz., *a, b, c* etc. If x is an element of the set A, we write symbolically $x \in A$ (x belongs to A). If x is not a member of the set A, we write $x \notin A$ (x does not belong to A).

$\in$ = Belong to (is an element of)

$\notin$ = Does not belong to (is not a member of)

$\subset$ = Contained in

$\supset$ = Contains

: = Such that

$\cup$ = Union

$\cap$ = Intersection.

If every element of the set A belongs to the set B, *i.e.*, if $x \in A \implies x \in B$, then we say that A is a subset of B and write symbolically $A \subseteq B$ (A is contained in B) or $B \supseteq A$ (B contains A). Two sets A and B are said to be equal or identical if $A \subseteq B$ and $B \subseteq A$ and write $A = B$ or $B = A$.

A null or an empty set is one which does not contain any element at all and is denoted by ϕ.

Remarks:

1. Every set is a subset of itself.
2. An empty set is a subset of every set.
3. A set containing only one element is conceptually distinct from the element itself, but will be represented by the same symbol for the sake of convenience.

Operations on sets: The 'union' of the two given sets A and B, denoted by $A \cup B$, is defined as a set consisting of all those points which belong to either A or B or both. Thus symbolically,

$$A \cup B = \{x: x \in A \text{ or } x \in B\}$$

Similarly, $\bigcup_{i=1}^{n} A_i = \{x: x \in A_i \text{ for at least one } i = 1, 2,..., n\}$

The 'intersection' of the two sets A and B, denoted by $A \cap B$, is defined a set consisting of those elements which belong to both A and B. Thus symbolically,

$$A \cap B = \{x: x \in A \text{ and } x \in B\}$$

Factorial

The product of all the positive integers, say 1 to n is denoted by $n!$ or $\lfloor \underline{n}$. It is read as 'n' factorial. This notation helps us to write such products in a compact form.

Thus, $n\,! = 1 \times 2 \times 3 \times 4 \times ... \times n.$

For example, $6\,! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

or $6\,! = 6 \times 5\,! = 720$

or $6\,! = 6 \times 5 \times 4\,! = 720$

Permutation

The permutation means any of the ways in which a set of objects can be arranged. In other words, permutation refers to different arrangement of objects in a set where all elements are different and distinguishable. For example, two books, A and B, taken at a time can be arranged in two ways, *i.e.*, AB and BA. Another example, we have 3 books and can arrange them in 6 ways, as follows:

ABC	ACB	BAC
BCA	CAB	CBA

It is to be noted that the order of arrangement is important and taken into account, when the order is changed, a different permutation results.

Notes:

1. When out of total n objects, p_1 objects be of first kind, p_2 objects be of second kind and so on p_k objects be of other kind, then the number of permutations of these n objects are $\dfrac{(p_1 + p_2 + ... + p_k)}{p_1\,!\, p_2\,! ... p_k\,!} = \dfrac{n\,!}{p_1\,!\, p_2\,! ... p_k\,!}$,

   where $n = p_1 + p_2 + ... + p_k$.

NOTES

2. **Permutation in a ring or circle:** When things are arranged in a row, we find two ends in each arrangement, while when the things are arranged in circle, there is no such ends.

Thus the number of ways in which n different things can be arranged in a circle taking all together is $(n-1)$, since any one of the things placed first is fixed and the remaining $(n-1)$ things can now be arranged in $(n-1)$, ways. For example, in how many ways 7 students form a ring ? In this case we are concerned with the relative positions of the students and therefore, 1 student is kept fixed. The remaining 6 students can be arranged taken all together in 6 ways. Hence, the required number of ways = 6 ! = 720 ways.

Permutation of n different things taken r at a time: suppose, we have n different objects and r space to be filled; it is difficult to write down permutations. Therefore, a formula has been derived symbolically

$$^nP_r = \frac{n!}{(n-r)!}, \text{ similarly } {}^nP_n = n!, 0! = 1.$$

Example 1: (i) *In how many different ways can the letters of the word SUNDAY be arranged ?*

(*ii*) *How many of these arrangement begin with S?*

(*iii*) *How many begin with A?*

(*iv*) *How many begin with S but do not end with A?*

Solution: (i) There are six different letters *S, U, N, D, A, Y* in the given word and these six letters can be arranged taking all at a time in $^6P_6$ ways. Hence the required number of ways,

$$^6P_6 = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720 \text{ ways.}$$

(ii) For the arrangements beginning with *S*, *S* is kept fixed and we can arrange the other 5 letters *U, N, D, A* and *Y* taking 5 at time in $^5P_5$ ways. Hence the required number of ways, $^5P_5 = 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ ways.

(iii) The number of arrangements beginning with $A = 120$.

(iv) For the arrangements which begin with *S* and end with *A*, both *S* and *A* are fixed and we can arrange the other 4 letters *U, N, D* and *Y* taken 4 at a time in $^4P_4$ ways *i.e.*,

$$^4P_4 = 4! = 4 \times 3 \times 2 \times 1 = 24 \text{ ways.}$$

Hence, the required number of arrangements which begin with *S* but do not end with $A = 120 - 24 = 96$ ways.

Example 2: (i) *How many words with or without meaning can be formed by using all the letters of the word BOTANY, using each letter exactly once?*

(ii) *In how many ways can the letters of the word BIOSTATISTICS be arranged?*

Solution: (i) The word BOTANY contains 6 different letters

∴ The number of required words

= Number of arrangements of 6 letters taken all at a time

$= {}^{6}P_{6} = 6\,! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720.$

(ii) In the word BIOSTATISTICS, there are 13 alphabets, out of which *I*, *S* and *T* appears thrice and rest alphabets appear once only. Thus, the number of permutations according to formula

$$\frac{n\,!}{p_1\,!\,p_2\,!\,p_3\,!} = \frac{13\,!}{3\,!\,3\,!\,3\,!}$$

$$= \frac{13 \times 12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)\,(3 \times 2 \times 1)\,(3 \times 2 \times 1)}$$

$= 2,88,28,800.$

Combination

In permutation, order of the objects was important, But in many problems, we are interested only in selecting or choosing object without regard to order. Such selections are called combinations. In other words, 'permutation is an arrangement and combination is a selection of the objects'. For example, *ABC* and *BCA* are the same combination because order is irrelevant.

A combination of *n* different objects taken *r* at a time is denoted by ${}^{n}C_{r}$. It is a selection of only *r* objects out of *n* objects and we do not consider the order of arrangement.

Symbolically, $\boxed{{}^{n}C_{r} = \frac{n\,!}{r\,!\,(n-r)\,!}}$

Similarly $${}^{n}C_{n} = \frac{n\,!}{n\,!\,(n-n)\,!} = 1.$$

Comparison between Permutation and Combination

Number of objects ***(n)***	**Taking at a time** ***(r)***	**Combination** ${}^{n}C_{r}$	**Permutation** ${}^{n}P_{r}$
AB	Two	AB	AB BA
ABC	Three	ABC	ABC ACB BCA BAC CAB CBA
ABC	Two	AB AC BC	AB BA CA AC BC CB

To find the value of ${}^nP_r$:

We know that $$\boxed{{}^nC_r = \frac{n!}{r!(n-r)!}}$$

and $$\boxed{{}^nP_r = \frac{n!}{(n-r)!}}$$

Thus, $${}^nC_r = \frac{{}^nP_r}{r!}$$

Note: ${}^nC_n = 1$, ${}^nP_n = n!$, ${}^nP_1 = n$, ${}^nC_1 = n$, ${}^nC_0 = 1$, ${}^nP_0 = 1$.

Example 3: *A man has 6 friends. In how many ways he can invite one or more of them to a dinner ?*

Solution: He may invite one or more friends by selecting either 1 friend or 2 friends or 3 friends or 4 friends or 5 friends or 6 friends out of 6 friends.

1 friend can be selected out of 6 in ${}^6C_1$ ways $= \frac{6!}{1!(6-1)!} = 6$ ways

2 friends can be selected out of 6 in ${}^6C_2$ ways $= \frac{6!}{2!(6-2)!} = 15$ ways

3 friends can be selected out of 6 in ${}^6C_3$ ways $= \frac{6!}{3!(6-3)!} = 20$ ways

4 friends can be selected out of 6 in ${}^6C_4$ ways $= \frac{6!}{4!(6-4)!} = 15$ ways

5 friends can be selected out of 6 in ${}^6C_5$ ways $= \frac{6!}{5!(6-5)!} = 6$ ways

6 friends can be selected out of 6 in ${}^6C_6$ ways $= \frac{6!}{6!(6-6)!} = 1$ way

Hence the required number of ways = 6 + 15 + 20 + 15 + 6 + 1 = 63 ways.

10.4 EVENTS AS SUBSETS OF SAMPLE SPACE

Event

An **event** of a random experiment is defined as a subset of the sample space of the random experiment. If the outcome of an experiment is an element of an event A, we say that the event A has occurred. An event is called an **elementary** (*or* **simple**) **event,** if it contains only one sample point. In the experiment of rolling a die, the event A of getting '3' is a simple event. We write A′ = {3}. An event is called an

impossible event, if it can never occur. In the above example, the event B = {7} of getting '7' is an impossible event. On the other hand, an event which is sure to occur is called a **sure event.** In the above example of rolling a die, the event C of getting a number less than 7 is a sure event. A sure event is also called a **certain event.**

Algebra of Events

We know that the events of a random experiments are sets, being subsets of the sample space. Thus, we can use set operations to form new events.

Let A and B be any two events associated with a random experiment.

The event of occurrence of either A or B or both is written as 'A or B' and is denoted by the subset $A \cup B$ of the sample space. In other words, $A \cup B$ represents the event of occurrence of at least one of A and B.

The event of occurrence of both A and B is written as 'A and B' and is denoted by the subset $A \cap B$ of the sample space. For simplicity the event $A \cap B$ is also denoted by 'AB'.

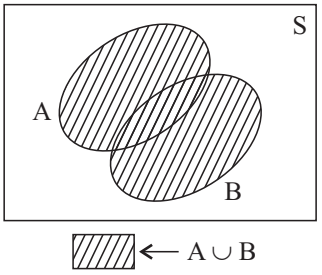


Fig. 10.3

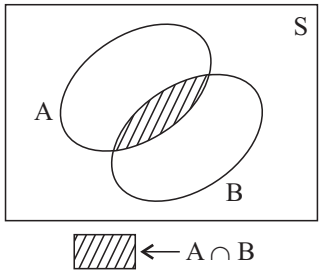


Fig. 10.4

The event of non-occurrence of event A is written as 'not A' and is denoted by the set A′, which is the complement of set A. The event A′ is called the **complementary event** of the event A.

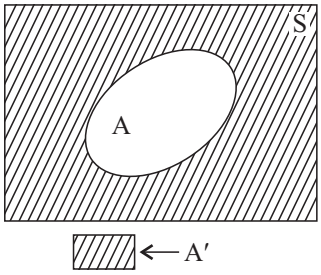


Fig. 10.5

Example 4: *8 tickets are numbered 1 to 8. A ticket is drawn at random. Write the sample space and the event A of getting a ticket with odd number.*

Solution: Here S = {1, 2, 3, 4, 5, 6, 7, 8}. The odd numbers from 1 to 8 are 1, 3, 5, 7.

$\therefore$ A = {1, 3, 5, 7}.

Example 5: *There are 2 children in a family. Find the events that:*

(*i*) *both children are boys* (*ii*) *only one of the children is a girl*

(*iii*) *there is at least one girl* (*iv*) *the older child is a boy.*

Solution: Here S = {BB, BG, GB, GG}.

(i) Let A be the event that both children are boys.

$\therefore$ A = {BB}.

(ii) Let A be the event that only one of the children is a girl.

$\therefore$ A = {BG, GB}.

(iii) Let A be the event that there is at least one girl.

$\therefore$ A = {BG, GB, GG}.

(iv) Let A be the event that the older child is a boy.

$\therefore$ A = {BB, BG}.

Example 6: *An urn contains 4 red and 6 yellow balls. Two balls are drawn at random from the urn. Find the number of elements in the sample space. Also find the number of elements in the event of getting:*

(*i*) *both balls red* (*ii*) *one ball red and one ball yellow*

(*iii*) *both balls yellow.*

Solution: Total number of balls = 4 + 6 = 10.

No. of elements is S = no. of ways of selecting 2 balls out of 10 balls

= no. of combinations of 10 things taking 2 at a time

$$= {}^{10}C_2 = \frac{10 \times 9}{1 \times 2} = 45.$$

(i) Let A be the events of getting 2 red balls.

$\therefore$ No. of elements in $A = {}^4C_2 \times {}^6C_0 = \frac{4 \times 3}{1 \times 2} \times 1 = 6.$

(ii) Let B be the event of getting one red and one yellow balls.

$\therefore$ No. of elements in $B = {}^4C_1 \times {}^6C_1 = 4 \times 6 = 24.$

4 Red
6 Yellow

(iii) Let C be the event of getting 2 yellow balls.

$\therefore$ No. of elements in $C = {}^4C_0 \times {}^6C_2 = 1 \times \frac{6 \times 5}{1 \times 2} = 15.$

Example 7: *A card is drawn from a pack of 52 playing cards, each card being equally likely to be drawn. Let A be the event, "card drawn is red". B the event "card drawn is a king or a queen" and C the event "card drawn is a king". Find the value of* $P(A \cap B \cap C)$.

Solution: A = event of getting a red card.

B = event of getting a king or a queen.

C = event of getting a king.

$\therefore$ $B \cap C$ = event of getting a king.

$\therefore$ $A \cap (B \cap C)$ = event of getting a red king *i.e.,* king of hearts or king of diamonds.

$$\therefore P(A \cap B \cap C) = \frac{2}{52} = \frac{1}{26}.$$

Example 8: *Two unbiased coins are tossed simultaneously. Find the probability of getting:*

(*i*) *one head* (*ii*) *one tail*

(*iii*) *at most one head* (*iv*) *at least one tail*

(*v*) *more than 2 heads* (*vi*) *less than 3 tails.*

Solution: Here S = (HH, HT, TH*, TT).

NOTES

(i) Let A = event of getting one head. ∴ A = {HT, TH}

$$\therefore \quad \text{P(one head)} = \text{P(A)} = \frac{n(\text{A})}{n(\text{S})} = \frac{2}{4} = \frac{1}{2}.$$

(ii) Let A = event of getting one tail. ∴ A = {HT, TH}

$$\therefore \quad \text{P(one tail)} = \text{P(A)} = \frac{n(\text{A})}{n(\text{S})} = \frac{2}{4} = \frac{1}{2}.$$

(iii) Let A = event of getting at most one head. ∴ A = {HT, TH, TT}

$$\therefore \quad \text{P(at most one head)} = \text{P(A)} = \frac{n(\text{A})}{n(\text{S})} = \frac{3}{4}.$$

(iv) Let A = event of getting at least one tail. ∴ A = {HT, TH, TT}

$$\therefore \quad \text{P(at least one tail)} = \text{P(A)} = \frac{n(\text{A})}{n(\text{S})} = \frac{3}{4}.$$

(v) Let A be the event of getting more than 2 heads. ∴ A = ϕ.

$$\therefore \quad \text{P(more than 2 heads)} = \text{P(A)} = \frac{n(\text{A})}{n(\text{S})} = \frac{0}{4} = 0.$$

(vi) Let A be the event of getting less than 3 tails. ∴ A = {HH, HT, TH, TT}

$$\therefore \quad \text{P(less than 3 tails)} = \text{P(A)} = \frac{n(\text{A})}{n(\text{S})} = \frac{4}{4} = 1.$$

Probability of an Event

Suppose in a random experiment, there are *n* exhaustive, equally likely outcome. Let A be an event and there are *m* outcomes (cases) favourable to the happening of it. The **probability** P(A) of the happening of the event A is defined as:

$$\mathbf{P(A) = \frac{Total\ number\ of\ cases\ favourable\ to\ the\ happening\ of\ A}{Total\ number\ of\ exhaustive\ equally\ likely\ cases}}$$

$$= \frac{\mathbf{m}}{\mathbf{n}}.$$

It may be observed from this definition, that

$$0 \le m \le n.$$

$$\therefore \quad 0 \le \frac{m}{n} \le 1 \quad \text{or} \quad \mathbf{0 \le P(A) \le 1.}$$

The number of cases favourable of the non-happening of the event A is $n - m$.

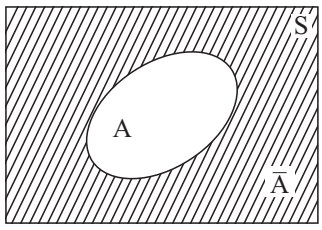


Fig. 10.4

$\therefore$ $P(\text{not A}) = \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(A).$

$\therefore$ $P(A) + P(\text{not A}) = 1$ *i.e.,* $P(A) + P(\overline{A}) = 1$

NOTES

If A is a sure event, then $P(A) = \frac{n}{n} = 1$ and if A happens to be an impossible event, then $P(A) = \frac{0}{n} = 0$.

From now onward, we shall always assume that the outcomes of any given random experiment are equally likely unless the contrary is stated explicitely.

Example 9: *A coin is tossed twice. If A denotes the event "number of heads is odd" and B denotes the event "number of tails is at least one". Find the cases favourable to the event A $\cap$ B.*

Solution: Here S = {HH, HT, TH, TT}

A = {HT, TH}, B = {HT, TH, TT}

$\therefore$ A $\cap$ B = event of occurring both A and B

= {HT, TH}.

Example 10: *A, B and C are three events associated with the sample space S of a random experiment. If A, B and C also denote the subsets of S representing these events, what are the sets representing the events:*

(*i*) *Out of the three events, only A occurs*

(*ii*) *Out of the three events, not more than two occur*

(*iii*) *Out of the three events, only one occurs*

(*iv*) *Out of the three events, exactly two events occur*

(*v*) *Out of the three events, at least two events occur.*

Solution: (i) In this event, A occurs and B, C do not occur ?

$\therefore$ Required event = $A \cap B' \cap C'$.

(ii) In this event, all events do not occur simultaneously.

$\therefore$ Required event = $(A \cap B \cap C)'$.

(iii) In this event, either only A occur or only B occur or only C occur.

$\therefore$ Required event = $(A \cap B' \cap C') \cup (A' \cap B \cap C') \cup (A' \cap B' \cap C)$.

(iv) In this event either only A, B occur or only B, C occur or only A, C occur.

$\therefore$ Required event = $(A \cap B \cap C') \cup (A' \cap B \cap C) \cup (A \cap B' \cap C)$.

(v) In this event, either only A, B occur or only B, C occur or only A, C occur or all occur.

$\therefore$ Required event = $(A \cap B \cap C') \cup (A' \cap B \cap C) \cup (A \cap B' \cap C) \cup (A \cap B \cap C)$.

'Odds in Favour' and 'Odds Against' an Event

Let A be an event of a random experiment. The ratio P(A): P($\overline{A}$) is called the **odds in favour** of happening of the event A. The ratio P($\overline{A}$): P(A) is called the **odds against** the happening of the event A.

Let odds in favour of an event A be *m*: *n*.

Let $P(A) = p$. $\therefore$ $p: 1 - p = m: n$

$\Rightarrow$ $\frac{p}{1-p} = \frac{m}{n}$ $\Rightarrow$ $np = m - mp$

$\Rightarrow$ $p = \frac{m}{m+n}$ *i.e.,* $P(A) = \frac{m}{m+n}$.

$\therefore$ **If odds in favour of A are m: n, then P(A) = $\frac{\mathbf{m}}{\mathbf{m+n}}$.**

Similarly, if odds against A are *m: n* then odds in favour of A are *n: m* and $P(A) = \frac{n}{n+m}$.

Remark: Odds in favour of event A are same as odds against the complement A′ of A and *vice-versa.*

Example 11: *Find the probability of the event A if* (i) *odds in favour of event A are 5 : 7* (ii) *odds against A are 3 : 4.*

Solution: (i) Odds in favour of event A are 5 : 7.

Let $P(A) = p$. $\therefore$ $p : 1 - p = 5 : 7$

$\Rightarrow$ $\frac{p}{1-p} = \frac{5}{7}$ $\Rightarrow$ $7p = 5 - 5p$

$\Rightarrow$ $12p = 5$ $\Rightarrow$ $p = \frac{5}{12}$

$\therefore$ $P(A) = \frac{5}{12}$.

(ii) Odds against event A are 3 : 4. Let $P(A) = p$. $\therefore$ $1 - p : p = 3 : 4$

$\Rightarrow$ $\frac{1-p}{p} = \frac{3}{4}$ $\Rightarrow$ $4 - 4p = 3p$

$\Rightarrow$ $7p = 4$ $\Rightarrow$ $p = \frac{4}{7}$

$\therefore$ $P(A) = \frac{4}{7}$.

Check Your Progress

State whether the following statements are True or False:

6. The permutation mean any of the way in which a set of objects can be arranged.
7. An elementary event contains more than one sample point.
8. Complement of set A is denoted by A′.
9. A ∪ B represents the events of occurance of both A and B.
10. If odds in favour of A are $m : n$, then $P(A) = \frac{m}{m+n}$.

10.5 CLASSIFICATION OF PROBABILITY

INDEPENDENT EVENTS

Conditional Probability

Let us consider the random experiment of throwing a die. Let A be the event of getting an odd number on the die.

$\therefore$ S = {1, 2, 3, 4, 5, 6} and A = {1, 3, 5}.

$\therefore$ $P(A) = \frac{3}{6} = \frac{1}{2}$.

Let B = {2, 3, 4, 5, 6}. If, after the die is thrown, we are given the information that the event B has occurred, then the probability of event A will no more be $\frac{1}{2}$, because in this case, the favourable cases are three and the total number of possible outcomes will be five and not six. The probability of event A, with the condition that event B has happened will be 3/5. This conditional probability is denoted as P(A/B). Let us define the concept of conditional probability in a formal manner.

Let A and B be any two events associated with a random experiment. The probability of occurrence of event A when the event B has already occurred is called the **conditional probability** of A when B is given and is denoted as P(A/B). The conditional probability P(A/B) is meaningful only when P(B) ≠ 0, *i.e.,* when B is not an impossible event.

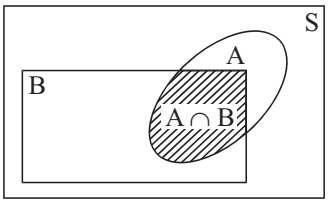


Fig. 10.5

By definition,

P(A/B) = Probability of occurrence of event A when the event B as already occurred.

$$= \frac{\text{No. of cases favourable to B which are also favourable to A}}{\text{No. of cases favourable to B}}$$

$$\therefore \quad \mathbf{P(A/B) = \frac{No.\ of\ cases\ favourable\ to\ A \cap B}{No.\ of\ cases\ favourable\ to\ B}}$$

NOTES

$$\text{Also,} \quad P(A/B) = \frac{\dfrac{\text{No. of cases favourable to A} \cap \text{B}}{\text{No. of cases in the sample space}}}{\dfrac{\text{No. of cases favourable to B}}{\text{No. of cases in the sample space}}}$$

$$\therefore \quad \mathbf{P(A/B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) \neq 0.}$$

Similarly, we have

$$\mathbf{P(B/A) = \frac{P(A \cap B)}{P(A)}, \text{ provided } P(A) \neq 0.}$$

Independent Events

Let A and B be two events associated with a random experiment. We have

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \text{ provided } P(A) \neq 0.$$

$$\therefore \quad P(A \cap B) = P(A)\, P(B/A).$$

In general P(B/A) may or may not be equal to P(B). When P(B/A) and P(B) are equal, then the events A and B are of special importance.

Two events associated with a random experiment are said to be **independent events** if the occurrence or non-occurrence of one event does not affect the probability of the occurrence of the other event. For example, the events A and B are independent events when P(A/B) = P(A) and P(B/A) = P(B).

Theorem: *Let A and B be events associated with a random experiment. The events A and B are independent if and only if P(A ∩ B) = P(A) P(B).*

Proof: Let A and B be independent events.

$$\therefore \quad P(A \cap B) = \left(\frac{P(A \cap B)}{P(B)}\right) P(B) = P(A/B)\, P(B)$$

$$\left(\because \quad P(A/B) = \frac{P(A \cap B)}{P(B)}\right)$$

$$= P(A)\, P(B) \qquad (\because \quad P(A/B) = P(A))$$

$$\therefore \quad P(A \cap B) = P(A)\, P(B).$$

Conversely, let $P(A \cap B) = P(A)\, P(B)$.

$$\therefore \quad P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\, P(B)}{P(B)} = P(A)$$

NOTES

and $$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A)\,P(B)}{P(A)} = P(B).$$

$\therefore$ $P(A/B) = P(A)$ and $P(B/A) = P(B)$.

$\therefore$ A and B are independent events.

Remark 1: $P(A \cap B) = P(A)\,P(B)$ is the necessary and sufficient condition for the events A and B to be independent.

Remark 2: Let A and B be events associated with a random experiment.

(i) Let A and B be (mutually exclusive) $\therefore$ $P(A \cap B) = 0$

$\therefore$ $P(A \cap B) \neq P(A)\,P(B)$ *i.e.,* A and B are not independent events.

$\therefore$ **Mutually exclusive events cannot be independent.**

(ii) Let A and B be independent.

$\therefore$ $P(A \cap B) = P(A)\,P(B)$ *i.e.,* $P(A \cap B) \neq 0$.

$\therefore$ A and B are not mutually exclusive events.

$\therefore$ **Independent events cannot be mutually exclusive.**

Important observation: If A and B be any two events associated with a random experiment, then their physical description is not sufficient to decide if A and B are independent events or not. A and B are declared to be independent events only when we have $P(A \cap B) = P(A)\,P(B)$.

Dependent Events

Let A and B be two events associated with a random experiment. If A and B are not independent events, then these are called **dependent events.**

$\therefore$ In case of dependent events, we have **$P(A \cap B) = P(A)\,P(B/A)$.**

10.6 SUMMARY

- By a deterministic phenomenon we mean one in which the result can be predicted with certainty.
- However, there are phenomena [as covered by (ii) above] which do not lend themselves to deterministic approach and are known as 'unpredictable' or 'probabilistic' phenomena.
- ***Mathematical (or Classical) Definition of Probability:*** If an event can happen in n ways which are equally likely, exhaustive and mutually exclusive and out of these n ways, m ways are favourable to an event A, then the probability of happening of A is given by

$$p \text{ or } P(A) = \frac{m}{n}.$$

- Any operation that results in two or more outcomes is called an experiment, and performing of an experiment is called trial.
- A random experiment is defined as an experiment in which all possible outcomes are known and which can be repeated under identical conditions but it is not possible to predict the outcome of any particular trial in advance.
- The sample space of a random experiment is defined as the set of all possible outcomes of the experiment.
- Any subset of the sample space is defined as an event.
- The total number of all possible outcomes in any trial are known as exhaustive events or cases.
- Several events are said to be 'independent' if the happening (or non-happening) of an event is not affected by the happening (or non-happening) of the other events.
- The number of cases 'favourable' to an event in a trial is the number of outcomes which entail the happening of the event.
- Events are said to be 'mutually exclusive' if the happening of any one of them precludes the happening of all others, *i.e.,* if no two or more of them can happen simultaneously in the same trial.
- Events are said to be equally likely, if there is no reason to expect any one in preference to any other.
- Events obtained by combining together two or more elementary events are known as the compound events.
- The product of all the positive integers, say 1 to n is denoted by $n!$ or $\lfloor \underline{n}$. It is read as 'n' factorial. This notation helps us to write such products in a compact form.
- The permutation means any of the ways in which a set of objects can be arranged.
- In permutation, order of the objects was important, But in many problems, we are interested only in selecting or choosing object without regard to order. Such selections are called combinations.
- The ratio $P(A) : P(\overline{A})$ is called the **odds in favour** of happening of the event A. The ratio $P(\overline{A}) : P(A)$ is called the **odds against** the happening of the event A.
- The probability of occurrence of event A when the event B has already occurred is called the **conditional probability** of A when B is given and is denoted as P(A/B).

NOTES

- Two events associated with a random experiment are said to be **independent events** if the occurrence or non-occurrence of one event does not affect the probability of the occurrence of the other event.

10.7 GLOSSARY

- ***Deterministic Model:*** A deterministic model is defined as a model which stipulates that the condition under which an experiment is performed determines the outcome of the experiment.
- ***Elements:*** The objects comprising a set are called elements.
- ***Sure Event:*** An event which is sure to occur is called a sure event.

10.8 ANSWERS TO CHECK YOUR PROGRESS

1. Impossible event
2. exhaustive events
3. random
4. event
5. compound event
6. True
7. False
8. True
9. False
10. True

10.9 TERMINAL AND MODEL QUESTIONS

1. From a group of 3 boys and 2 girls, we select two children. What would be the sample space of this random experiment ? Also, write the events of getting (i) both girls (ii) both boys.
2. Two dice are thrown simultaneously. Find the number of elements in the event of getting:

   (i) sum 4 (ii) sum 7

   (iii) sum 11 (iv) sum not greater than 5.

3. Four coins are tossed simultaneously. Write the events of getting:

   (i) one head (ii) two tails

   (iii) at least two heads (iv) more than four tails.

4. Two cards are drawn at random from a well shuffled pack of cards. Find the number of elements in the sample space. Also find the number of elements in the event of getting:

   (i) two ace cards (ii) two red cards

   (iii) two heart cards (iv) one king and one queen cards.

5. A coin is tossed. Find the events A', B', $A \cup B$, $A \cap B$, where:

   A = event of getting no head and B = event of getting one head.

6. A die is thrown. If:

   A = event of getting a prime number and B = event of getting number greater than 3, find the events A^c, B^c, $A \cup B$ and $A \cap B$.

7. Three coins are tossed simultaneously. Let:

   A = event of getting 2 heads and B = event of getting at most one tail. Find the events, 'A or B' and 'A and B'.

8. A and B are events associated with a random experiment. Write the following events symbolically:

   (i) only A occurs (ii) only B occurs

   (iii) none of them occurs (iv) at least one of them occur.

9. A die is thrown. Observe the number that appears on the top face. Describe the following events:

   (i) A : a number less than 7 (ii) B : a number greater than 7

   (iii) C : a multiple of 3 (iv) D : a number less than 4

   (v) E : an even number greater than 4 (vi) F : a number not less than 3.

   Also, find $A \cup B$, $A \cap B$, $B \cup C$, $E \cap F$, $D \cap E$, F'.

10. Two unbiased dice are thrown simultaneously. Find the probability of:

    (i) getting sum 10

    (ii) not getting same number on the dice

    (iii) getting a multiple of 3 as the sum

    (iv) an even number on the first die and an odd number on the second die.

11. Three unbiased coins are tossed simultaneously. Find the probability of getting:

    (i) one tail

    (ii) more heads than tails

(iii) at the most one head

(iv) head on first coin and tail on the third coin.

12. An unbiased die is thrown. Find the probability of getting:

(i) number greater than 4 (ii) an even number

(iii) a number less than 8 (iv) a multiple of 4

13. Two dice are thrown. Find the probability that

(i) the total of the numbers on the dice is 8,

(ii) the first die shows 6,

(iii) the total of the numbers on the dice is greater than 8,

(iv) the total of the numbers on the dice is 13,

(v) both the dice show the same number,

(vi) the sum of the numbers shown by the dice is less than 5,

(vii) the sum of the numbers shown by the dice is exactly 6.

14. Five cards are drawn from a pack of cards. What is the probability of its being 4 kings and the remaining card be other?

15. From a pack of 52 cards, 4 cards are drawn at random. What is the probability that

(i) all of them are spades (ii) one is of each unit

(iii) all are diamonds

(iv) there are two spades and two hearts?

10.10 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 11: LAWS OF PROBABILITY

Structure

11.0 INTRODUCTION

The concept of probability originated in the beginning of eighteenth century in problems pertaining to games of chance such as throwing a die, tossing a coin, drawing a card from a pack of cards etc. Starting with games of chance, 'Probability' today has become one of the basic tools of statistics and has wide range of applications in Science, Engineering and Business. In this chapter we will learn about the laws of probability.

The laws of probability may be applied to any subject which involves chance or random happenings. The study of probability is closely related to genetics because it involves a statistical analysis of the ratios of various characteristics among the off spring of parents of a known phenotype. Multiplication law of probability is useful to find out the results of more than one generation also.

11.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Explain general additive law and additive law for mutually exclusive events
- Explain multiplication law of probability and to derive various theorems related to it
- Define conditional probability
- Explain various applications of probability
- Define Bay's theorem of probability
- Explain total probability rule

11.2 ADDITIVE LAW OF PROBABILITY

Mutually Exclusive Events

Two events associated with a random experiment are said to be **mutually exclusive,** if both cannot occur together in the same trial. In the experiment of throwing a die, the events A = {1, 4} and B = {2, 5, 6} are mutually exclusive events. In the same experiment, the events A = {1, 4} and C = {2, 4, 5, 6} are not mutually exclusive because, if 4 appear on the die, then it is favourable to both events A and C. The definition of mutually exclusive events can also be extended to more than two events. We say that more than two events are mutually exclusive, if the happening of one of these, rules out the happening of all other events. The events A = {1, 2}, B = {3} and C = {6}, are mutually exclusive in connection with the experiment of throwing a single die.

n events A_1, A_2, ..., A_n associated with a random experiment are said to **mutually exclusive events** if $A_i \cap A_j = \phi$ for all *i, j* and $i \neq j$.

For example, let a pair of dice be thrown and let A, B, C be the events that the sum is 7, sum is 8, sum is greater than 10 respectively,

$\therefore$ A = {(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)}

B = {(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)}

and C = {(5, 6), (6, 5), (6, 6)}

The events A, B and C are mutually exclusive.

If two events *A* and *B* are mutually exclusive, *i.e.,* $P(A \cap B) = 0$, then probability of occurrence of either *A* or *B* is the sum of the individual probability of *A* and *B*, Symbolically,

$$\boxed{P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)}\ .$$

The formula for probability of union of two events can be extended to any number of events. For three mutually exclusive events A, B and C, the probability of their union is given as

$$P(A \cup B \cup C) = P(A) + P(B) + P(C).$$

Example 1: *A die is tossed. What is the probability of getting 2 or 4 or 6?*

Solution: The probability of $2 = \frac{1}{6}$

The probability of $4 = \frac{1}{6}$

The probability of $6 = \frac{1}{6}$

Therefore, the probability of 2 or 4 or 6 is

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

Example 2: *The distribution of blood types in India is roughly as follows:*

A : 39% *B : 7%*

AB : 6% *O : 48%*

An individual is brought into the emergency room after an accident. He is to be blood typed. What is the probability that he will be of type A, B or AB?

Solution: Let E_1, E_2 and E_3 denote the events that the patients has type A, B and AB blood respectively.

$$P(E_1) = 39\% = \frac{39}{100} = 0.39$$

$$P(E_2) = 7\% = \frac{7}{100} = 0.07$$

$$P(E_3) = 6\% = \frac{6}{100} = 0.06.$$

Since it is impossible for an individual to have two different blood types, these events are mutually exclusive.

Therefore, the required probability

$$\begin{aligned} P(E_1 \text{ or } E_2 \text{ or } E_3) &= P(E_1) + P(E_2) + P(E_3) \\ &= 0.39 + 0.07 + 0.06 \\ &= 0.52. \end{aligned}$$

Hence, there is a 52% chance that the patient will have one of the three blood types.

Additive Law of Probability (For not Mutually Exclusive)

If A and B are two events (not mutually exclusive), then the probability of the union of A and B is governed by the law:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A, B and C are three events then $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

For example, if the probability of a person's buying a shoe is 0.6 and that of buying a cap is 0.3, we cannot calculate the probability of his buying either a shoe or a cap by adding the two probabilities. These events are not mutually exclusive, he can very well buy a shoe as well as a cap.

Example 3: *What is the probability of getting either an ace or a spade from a pack of 52 cards?*

Solution: The events, ace and spade can occur together because we can draw the ace of spades. Therefore, ace and spade are not mutually exclusive events.

$$P(\text{ace or spade}) = P(\text{ace}) + P(\text{spade}) - P(\text{ace and spade})$$

$$= \frac{4}{52} + \frac{13}{52} - \left(\frac{4}{52} \times \frac{13}{52}\right)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}.$$

Example 4: *A card is drawn at random from a well shuffled pack of cards. What is the probability that the card is a spade or a queen?*

Solution: Let A denotes a spade and B denotes a queen.

The probability of a spade card is

$$P(A) = \frac{13}{52} = \frac{1}{4}$$

The probability of a queen is

$$P(B) = \frac{4}{52} = \frac{1}{13}$$

Therefore, the required probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{1}{4} + \frac{1}{13} - \left(\frac{1}{4} \times \frac{1}{13}\right)$$

$$= \frac{4}{13}.$$

Example 5: *It was recently reported that 20% of all college students at some point in their college careers suffers from depression, that 4% consider suicide, and*

that 17% suffer from depression or consider suicide. What is the probability that a randomly selected college student suffers from depression and has considered suicide?

Solution: Let events A and B denote depression and consider suicide respectively, then it is given that

$$P(A) = \frac{20}{100} = 0.20,\ P(B) = \frac{4}{100} = 0.04,\ P(A \cup B) = \frac{17}{100} = 0.17$$

By addition theorem,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

or
$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$
$$= 0.20 + 0.04 - 0.17$$
$$= 0.07 \text{ or } 7\%.$$

Example 6: *It is thought that 30% of all people in a city are obese (A_1) and that 3% suffer from diabetes (A_2). 2% are obese and suffer from diabetes. What is the probability that a randomly selected person is obese or suffers from diabetes?*

Solution: We have given that

$$P(A_1) = 30\% = \frac{30}{100} = 0.30,$$

$$P(A_2) = 3\% = \frac{3}{100} = 0.03 \text{ and}$$

$$P(A_1 \cap A_2) = 2\% = \frac{2}{100} = 0.02.$$

The probability of a person to be obese or diabetic is given by

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$
$$= 0.30 + 0.03 - 0.02$$
$$= 0.31 \text{ or } 31\%.$$

11.3 MULTIPLICATION THEOREM OF PROBABILITY

If two events A and B are independent, the probability of their product (intersection) is equal to the product of their individual probabilities. Notationally,

$$\boxed{P(AB) = P(A \cap B) = P(A) \,.\, P(B)}\ .$$

This law can be extended to any number of events.

For three independent events A, B and C,

$$P(ABC) = P(A \cap B \cap C) = P(A) \,.\, P(B) \,.\, P(C).$$

For example, in tossing 2 coins,

NOTES

Probability of head in one coin = $\frac{1}{2}$

Probability of head in another coin = $\frac{1}{2}$

Thus probability of heads in both coins = $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

Example 7: *What is the probability of the heads on two or three successive tosses?*

Solution: (i) The probability of the head in first toss is

$$P(A) = \frac{1}{2}.$$

The probability of the head in second toss is

$$P(B) = \frac{1}{2}.$$

The probability of heads in both the tosses,

$$P(A \text{ and } B) = P(A \cap B) = P(A) \,.\, P(B)$$

$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

(ii) $$P(A) = \frac{1}{2},\ P(B) = \frac{1}{2},\ P(C) = \frac{1}{2}$$

The probability of heads in all the three coins,

$$P(ABC) = P(\text{A} \cap \text{B} \cap \text{C}) = P(A) \,.\, P(B) \,.\, P(C)$$

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.$$

Example 8: *The probability that a man aged 60 will survive 10 years is* $\frac{2}{5}$ *and a woman aged 50 surviving 10 years is* $\frac{3}{4}$*. What are the chances that they will both survive 10 years?*

Solution: Surviving of a man does not affect the surviving of a woman. Thus both the events are independent. The probability that they will both survive 10 years is

$$P(A \text{ and } B) = P(A \cap B) = P(A) \,.\, P(B)$$

$$= \frac{2}{5} \times \frac{3}{4} = \frac{3}{10}.$$

Example 9: *The probability that a man suffers from arthritis is 0.45 and has cold is 0.30. What is the probability that he suffers from arthritis and cold both?*

Solution: The events arthritis and cold are independent because the occurrence of one does not affect the occurrence of the other. Therefore, the required probability, *i.e.,*

$$P(\text{arthritis and cold}) = P(\text{arthritis}) \times P(\text{cold})$$
$$= 0.45 \times 0.30 = 0.135.$$

11.4 CONDITIONAL PROBABILITY

Many times the information is available that an event has occurred and one is required to find out the probability of occurrence of another event B utilising the information about A. Such a probability is known as conditional probability and is denoted by $P(B/A)$ *i.e.,* the probability of the event B given A. For example, suppose we know that a newly born baby will be a male (A), then one is interested to know the probability of a strile birth (B) *i.e.,* we want to calculate $P(B/A)$. The formula is,

$$\boxed{P(B/A) = \frac{P(A \cap B)}{P(A)}} = \frac{P(\text{both})}{P(\text{given})}$$

If A and B are independent [$P(A \cap B) = P(A)\ P(B)$], then

$$P(B/A) = \frac{P(A)\ P(B)}{P(A)} = P(B).$$

Notes:

1. The multiplication law explained in section 11.3 is not applicable in case of dependent events. For example, the chance that a patient with some disease survives the next year depends, of course, on his having survived to the present time and the current status of his disease. Such probability is called conditional probability.
2. The general rule of multiplication in its modified form in terms of conditional probabilities becomes

$$\boxed{P(A \text{ and } B) = P(A \cap B) = P(B) \times P(A/B) \text{ or } P(A) \times P(B/A)}\ .$$

Check Your Progress

Fill in the blanks:

1. Two events are said to be ..... , if both cannot occur together in the same trial.
2. Additive law for 2 mutually exclusive events is $P(A \text{ or } B)$ = ...
3. Conditional probability for 2 events A and B can denoted by ..... *i.e.*, the probability of event B given A.
4. Possible outcomes of an random experiment are called .....
5. The events which ensure the required happening are called .....

Example 10: *From past experience with the illness of his patients, a doctor has gathered the following information in a population:*

5% feel that they have cancer and do have cancer ;

45% feel that they have cancer and do not have cancer ;

10% do not feel that they have cancer and do have it ; and the remaining

40% feel that they do not have cancer and really do not have it.

Find the probability (i) a patient has cancer, given that he feels he has it. (ii) a patient feels he has cancer, given that he does have it.

Solution: Denoting the events, as

A when the patient feels he has cancer, and

B when the patient has cancer, we have

$$P(A \cap B) = P(AB) = 0.05$$
$$P(A) = 0.5$$
$$P(B) = 0.15$$

(i) The probability that a patient has cancer, given that he feels he has it, is given by

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.05}{0.5} = 0.1.$$

(ii) The probability he feels he has cancer, given that he does have it, is given by

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.05}{0.15} = 0.33.$$

Example 11: *Two dice are thrown. Find the probability that the sum of the numbers in the two dice is 10, given that the first die shows six.*

Solution: In throwing of two dice, there are 36 possible outcomes. Therefore exhaustive number of cases is 36.

Then the event A that the sum of numbers is 10 *i.e.*,

$$A = \{(4, 6), (5, 5), (6, 4)\}$$

and the event B that first die shows 6 is

$$B = \{(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}.$$

Thus, $$P(A) = \frac{3}{36} = \frac{1}{12}, P(B) = \frac{6}{36} = \frac{1}{6}$$

Also $$A \cap B = \{(6, 4)\} \quad \text{and} \quad P(A \cap B) = \frac{1}{36}$$

$P(A/B)$ = Probability that sum is 10 when first die shows 6

= Probability of A when B has occurred

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}.$$

Example 12: *It is estimated that among the U.S. population as a whole, 55% are over weight (A), 20% have high blood pressure (B), and 60% are overweight or have high blood pressure. Is the fact that a person is overweight independent of the state of his or her blood pressure?*

Solution: We have given that,

$$P(A) = 55\% = \frac{55}{100} = 0.55,$$

$$P(B) = 20\% = \frac{20}{100} = 0.20,$$

$$P(A \cup B) = 60\% = \frac{60}{100} = 0.60.$$

We are to show that overweight is independent of blood pressure or not. Using the addition theorem,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

or
$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$
$$= 0.55 + 0.20 - 0.60 = 0.15$$

Thus,
$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.15}{0.55} = 0.27.$$

If A (overweight) and B (blood pressure) are independent, then $P(B/A)$ should be equal to $P(B)$.

Here $P(B/A) = 0.27$ and $P(B) = 0.20$

i.e., $P(B/A) \neq P(B)$,

we may conclude that the events are not independent. (Practically an overweight person increases the probability of having high blood pressure).

11.5 PROBABILITY APPLICATIONS

The laws of probability may be applied to any subject which involves chance or random happenings. The study of probability is closely related to genetics because it involves a statistical analysis of the ratios of various characteristics among the offspring of parents of a known phenotype. Multiplication law of probability is useful to find out the results of more than one generation also. Probability applications in genetics will be more clear if we analyse the results of first, second and third generations of human families. We can make the probabilities of events even more explicit using a probability tree of all three generations. Figure 11.1 shows the possible outcome of first generation. Figure 11.2 will give four possible outcome of the second generation. The complete probability tree is shown in figure 11.3. However, it is clear that no event is affected by the events preceding or following it.

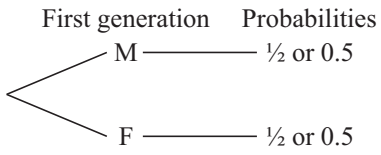


Fig. 11.1. Probability Tree of First Generation

Analysis of results of figure 11.1.

(i) Probability of getting a male = $\frac{1}{2}$ or 0.5 or 50%

(ii) Probability of getting a female = $\frac{1}{2}$ or 0.5 or 50%.

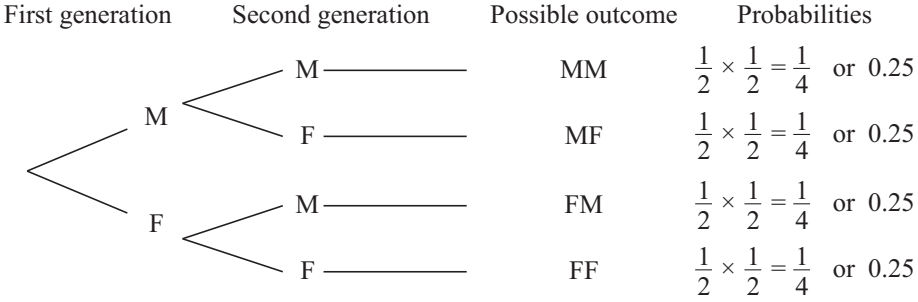


Fig. 11.2. Probability Tree of the First Generation and the Second Generation

Analysis of results of figure 11.2.

(i) Probability of getting two males = $\frac{1}{4}$ or 0.25 or 25%

(ii) Probability of getting two females = $\frac{1}{4}$ or 0.25 or 25%

(iii) Probability of getting one of either (male or female) sex

$$= \frac{1}{4} + \frac{1}{4} = \frac{1}{2} = 50\%.$$

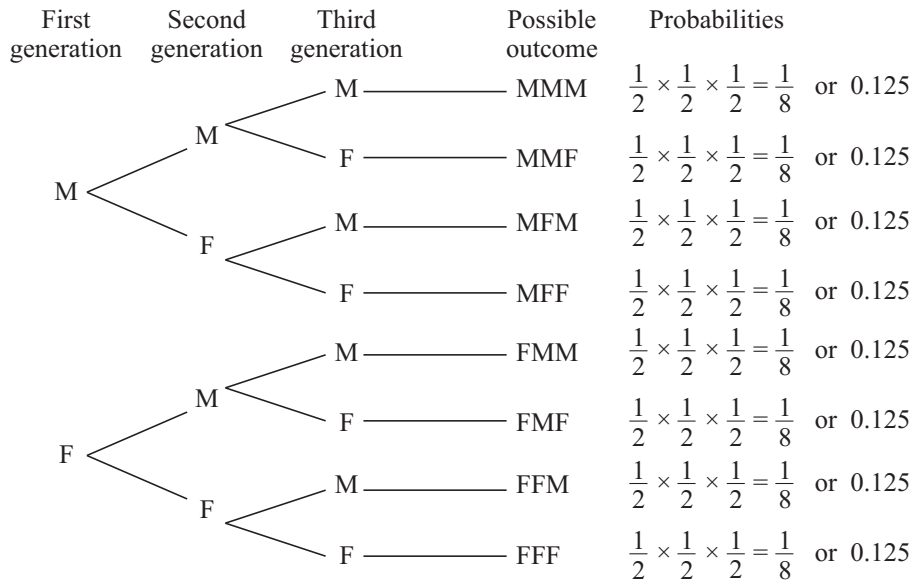


Fig. 11.3. A Complete Probability Tree

Possible Outcomes and Their Probabilities

	First generation		Second generation		Third generation	
	Possible outcomes	**Probabilities**	**Possible outcomes**	**Probabilities**	**Possible outcomes**	**Probabilities**
	M	0.5	*MM*	0.25	*MMM*	0.125
	F	0.5	*MF*	0.25	*MMF*	0.125
			FM	0.25	*MFM*	0.125
			FF	0.25	*MFF*	0.125
					FMM	0.125
					FMF	0.125
					FFM	0.125
					FFF	0.125
Total	2	1.0	4	1.0	8	1.0

By using the above table, we can obtain the following probabilities:

(i) Probability of getting 3 males

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \quad \text{or} \quad 0.125 \quad \text{or} \quad 12.5\%.$$

(ii) Probability of getting 3 females

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \quad \text{or} \quad 0.125 \quad \text{or} \quad 12.5\%.$$

(iii) Probability of getting 2 males and 1 female

$$= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} \quad \text{or} \quad 0.375 \quad \text{or} \quad 37.5\%.$$

(iv) Probability of getting 2 females and 1 male

$$= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} \quad \text{or} \quad 0.375 \quad \text{or} \quad 37.5\%.$$

(v) Total probability $= \frac{1}{8} + \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{8}{8} = 1$ or 100%.

(vi) Probability of getting *MFM* at the end of the third generation

$$P(MFM) = P(M) \times P(F) \times P(M) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$$

$$= 0.5 \times 0.5 \times 0.5$$

$$= 0.125 \quad \text{or} \quad 12.5\%.$$

Remarks:

1. We can understand the law of probability more clearly when it may be applied to genetic problems. In human beings, albinism is a recessive genetic trait, the birth of an affected child of normal looking parents indicates that the parents are heterozygous (carriers) for that trait. Parents are normal and healthy because of

recessive nature of gene. Since, the trait is monogenic recessive, they can expect affected children and normal children in the ratio of 1 : 3. The probability of having an affected child is thus $\frac{1}{4}$ at each birth and so also in all future pregnancies. In other words, the probability for normal child at each birth is $\frac{3}{4}$. If, parents decide to have three more children and want to know the chance that all the three will be albinos. The answer would be $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{64}$; not a very great chance. On the other hand, if they want to know the probability of all three being normal, the answer would be $\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$. One has to bear in mind that the occurrence of an albino in the first pregnancy will not influence the heredity of future children for this trait. By applying the laws of probability one can know that he/she may be carrying some serious hereditary defect.

2. One may often note the working of the laws of probability for predicting the ratio of boys and girls born in a family. The probability for a boy at any birth is $\frac{1}{2}$ and for a girl also $\frac{1}{2}$, since the human male produces an equal number of *X* and *Y* sperms. What is the probability that the first two children born in a family will be males? For this, one has to determine the product of separate probabilities at each conception; $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. What is the probability that the third child in the family in which the first two are males will also be a male? The probability of third child is also $\frac{1}{2}$ because the sex of any child is independent of the sex of the other children.
3. The principles of probability can be applied in solving the Mendel's problems of heredity. Probability can be used to obtain the phenotypic ratios to be expected from dihybrid, trihybrid or polyhybrid crosses etc.
4. Practical applications of the laws of probability are made by human geneticists and in some instances by animal breeders in analysing pedigrees. The first step in such an analysis is to determine whether the trait in question is behaving as a dominant or recessive. For example, deafness may behave as dominant in some families and recessive in others. Recessive genes are difficult to follow because they may remain hidden by their dominant alleles generation after generation. Carriers in the population usually cannot be identified until an affected child is born. Recessives are expressed more frequently in families in which the father and the mother are more closely related than parents in the general population.

11.6 BAYE'S PROBABILITY

Baye's probability is also known as 'inverse probability'. The idea of inverse probability was given by Sir Thomas Bayes in 1763 which was reprinted in

Biometrika in the year 1958. The problem of inverse probability arise when we have an outcome and we want to know the probability of its belonging to a specified trial or population out of many alternative trials or populations. To be more specific, let us consider three urns containing white (W), black (B) and red (R) balls as follows:

URN I : $2W$, $3B$ and $4R$ balls,

URN II : $3W$, $1B$ and $2R$ balls,

URN III : $4W$, $2B$ and $5R$ balls.

Two balls are drawn from a urn and they happen to be one white and one red balls. Now the interest lies to know the probability that the balls are drawn from URN III. Such a probability is Baye's probability. This type of problem is tackled by the fundamental theorem of inverse probability, given by Bayes.

Baye's theorem: If $E_1, E_2,..., E_n$ are mutually disjoint events with $P(E_i) \neq 0$, $(i = 1, 2,..., n)$, then for any arbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have

$$P(E_i/A) = \frac{P(E_i)\, P(A/E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A/E_i)} \quad .$$

Usually the physician knows the conditional probability of a particular symptom (or positive test) for a particular disease. However, it is important that he knows the conditional probability of the disease for an individual patient, given the particular symptom (or positive test). The Baye's theorem provides the means to derive the latter probability from the former probability. This theorem is illustrated by an example.

An example to develop Baye's theorem: This example concerns bacteriuria, and pyelonephritis in pregnancy. Suppose it is known that roughly 6 percent of pregnant women attending a prenatal clinic at a large urban hospital have bacteriuria (bacteria in the urine).

Consider two events E_1 a pregnant woman has bacteriuria, and E_2 she does not have bacteriuria.

Since E_1 and E_2 are mutually exclusive events,

$$P(E_1) = 0.06,\ P(E_2) = 1 - 0.06 = 0.94.$$

Suppose it is further known that 30 percent of bacteriuric and 1 percent of non-bacteriuric pregnant women proceed to develop this disease.

Using A to denote the occurrence of pyelonephritis, then

$P(A/E_1)$ = P(Pyelonephritis given that the pregnant woman was bacteriuric)

= 0.30

NOTES

$P(A/E_2)$ = P(Pyelonephritis given that the pregnant woman was non-bacteriuric)

= 0.01.

1. With these definitions consider the following probability questions:

(i) What is the chance of a pregnant woman having both bacteriuria and pyelonephritis ?

Using Multiplication Law:

$$P(E_1 \text{ and } A) = P(A/E_1) \,.\, P(E_1)$$
$$= (0.30) \,.\, (.06) = 0.0180.$$

(ii) What is the chance of a pregnant woman not having bacteriuria but having pyelonephritis?

Using multiplication Law:

$$P(E_2 \text{ and } A) = P(A/E_2) \,.\, P(E_2)$$
$$= (0.01)\ (0.94) = 0.0094.$$

2. What is the chance of pyelonephritis?

In this particular example, pyelonephritis can occur in two mutually exclusive ways, with or without bacteriuria. Hence, application of the additive law to the probabilities determined in 1(i) and (ii) gives,

$$P(\text{Pyelonephritis}) = P(A)$$
$$= P(E_1 \text{ and } A) + P(E_2 \text{ and } A)$$
$$= 0.0180 + 0.0094 = 0.0274$$

3. Finally, with the knowledge that a pregnant woman has developed pyelonephritis, what is the chance she had been bacteriuric?

Using the notation developed, the question is: what is the probability of (E_1/A) *i.e.,* the presence of bacteriuria given that the pregnant woman has pyelonephritis.

$$P(E_1/A) = \frac{P(E_1 \text{ and } A)}{P(A)}$$
$$= \frac{P(A/E_1) \,.\, P(E_1)}{P(A/E_1) \,.\, P(E_1) + P(A/E_2) \,.\, P(E_2)}$$
$$= \frac{(0.30)\ (0.06)}{(0.30)\ (0.06) + (0.01)\ (0.94)} = \frac{0.0180}{0.0274}$$
$$= 0.6569.$$

In other words, if a pregnant woman has developed pyelonephritis, there is a 65.7% chance that she had been bacteriuric.

Example 13: *The contents of urns I, II and III are as follows:*

URN I : 1 white, 2 black and 3 red balls,

URN II : 2 white, 1 black and 1 red balls and

URN III : 4 white, 5 black and 3 red balls.

One urn is chosen at random and two balls drawn. They happen to be white and red. What is the probability that they come from urns I, II or III?

Solution: Let E_1, E_2 and E_3 denote the events that the urn *I*, *II* and *III* is chosen, respectively, and let *A* be the event that the two balls taken from the selected urn are white and red. Then

$$P(E_1) = \frac{1}{3},\ P(E_2) = \frac{1}{3},\ \text{and}\ P(E_3) = \frac{1}{3},$$

$$P(A/E_1) = \frac{{}^1C_1 \times {}^3C_1}{{}^6C_2} = \frac{1 \times 3 \times 2 \times 1}{6 \times 5} \qquad \left({}^6C_2 = \frac{6 \times 5}{2 \times 1}\right)$$

$$= \frac{1}{5},$$

$$P(A/E_2) = \frac{{}^2C_1 \times {}^1C_1}{{}^4C_2} = \frac{2 \times 1 \times 2 \times 1}{4 \times 3} \qquad \left({}^4C_2 = \frac{4 \times 3}{2 \times 1}\right)$$

$$= \frac{1}{3},$$

$$P(A/E_3) = \frac{{}^4C_1 \times {}^3C_1}{{}^{12}C_2} \qquad \left({}^{12}C_2 = \frac{12 \times 11}{2 \times 1}\right)$$

$$= \frac{4 \times 3 \times 2 \times 1}{12 \times 11} = \frac{2}{11}$$

Hence,

$$P(E_1/A) = \frac{P(E_1)\,.\,P(A/E_1)}{P(E_1)\,P(A/E_1) + P(E_2)\,.\,P(A/E_2) + P(E_3)\,.\,P(A/E_3)}$$

$$= \frac{\frac{1}{3} \times \frac{1}{5}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{33}{118}.$$

$$P(E_2/A) = \frac{P(E_2)\,.\,P(A/E_2)}{P(E_1)\,P(A/E_1) + P(E_2)\,.\,P(A/E_2) + P(E_3)\,.\,P(A/E_3)}$$

$$= \frac{\frac{1}{3} \times \frac{1}{3}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{55}{118}$$

NOTES

and $$P(E_3/A) = \frac{P(E_3) \,.\, P(A/E_3)}{P(E_1)\, P(A/E_1) + P(E_2) \,.\, P(A/E_2) + P(E_3) \,.\, P(A/E_3)}$$

$$= \frac{\frac{1}{3} \times \frac{2}{11}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{30}{118}.$$

11.7 JOINT PROBABILITY

Definition of 'Joint Probability'

A statistical measure where the likelihood of two events occurring together and at the same point in time are calculated. Joint probability is the probability of event B occurring at the same time event A occurs.

Notation for joint probability takes the form:

$$P(A \cap B) \text{ or } P(A, B)$$

which reads: the joint probability of A and B.

Example: The probability that a card is a four and red = p(four and red) = 2/52 = 1/26.

(There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

For example, a joint probability cannot be calculated when tossing a coin on the same flip. However, the joint probability can be calculated on the probability of rolling a 2 and a 5 using two different dice.

We'll introduce a simple, concrete example, and define joint probability in terms of that example.

Table 11.1 shows the number of male and female members of the standing faculty in the departments of Mathematics and English. We learn that the Math department has 1 woman and 37 men, while the English department has 17 women and 20 men. The two departments between them have 75 members, of which 18 are women and 57 are men.

Table. 11.1

	Math	English	Total
Female	1	17	18
Male	37	20	57
Total	38	37	75

Table 2 (below) shows the same information as proportions (of the total of 75 faculty in the two departments). If we wrote the name, sex and department affiliation of each of the 75 individuals on a ping-pong ball, put all 75 balls in a big urn, shook it up, and chose a ball at random, these proportions would represent the probabilities of picking a female Math professor (about 0.013, or 13 times in a thousand tries), a female English professor (0.227), a male Math professor (0.493), and so on.

In formula form, we would write

P(female, math) = 0.013,

P(female, english) = 0.227, etc. These are called "joint probabilities"; thus

P(female, english) is "the joint probability of *female* and *english*". Note that joint probabilities (like logical conjunctions) are symmetrical, so that **P(english, female)** means the same thing and **P(female, english)**–though often we chose a canonical order in which to write down such categories.

Table 11.2 represents the "joint distribution" of sex and department.

Table 11.2

	Math	English	Total
Female	0.013	0.227	0.240
Male	0.493	0.267	0.760
Total	0.506	0.494	1.00

The bottom row and rightmost column in Table 2 give us the proportions in the single categories of sex and department:

P(female) = 0.240, P(male) = 0.760, P(math) = 0.506, etc.

As before, these proportions can also be seen as the probabilities of picking a ball of the designated category by random selection from our hypothetical urn.

Check Your Progress

State whether the following statements are True or False:

6. The laws of probability may be applied to any subject which involves chance or random happenings.
7. Baye's probability is also known as inverse probability.
8. Joint probability of two events A and B can be denoted as $P(A \cup B)$.
9. Multiplication law of probability is useful to find out the result of one generation only.

11.8 SUMMARY

- Two events associated with a random experiment are said to be **mutually exclusive,** if both cannot occur together in the same trial.

- The definition of mutually exclusive events can also be extended to more than two events.
- If two events A and B are mutually exclusive, *i.e.,* $P(A \cap B) = 0$, then probability of occurrence of either A or B is the sum of the individual probability of A and B, Symbolically,

$$\boxed{P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)}\ .$$

- If A and B are two events (not mutually exclusive), then the probability of the union of A and B is governed by the law:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A, B and C are three events then $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.

- If two events A and B are independent, the probability of their product (intersection) is equal to the product of their individual probabilities. Notationally,
- For three independent events A, B and C,

$$P(ABC) = P(A \cap B \cap C) = P(A) \,.\, P(B) \,.\, P(C).$$

- Many times the information is available that an event has occurred and one is required to find out the probability of occurrence of another event B utilising the information about A. Such a probability is known as conditional probability and is denoted by $P(B/A)$ *i.e.,* the probability of the event B given A.
- The laws of probability may be applied to any subject which involves chance or random happenings.
- Baye's probability is also known as 'inverse probability'.
- **Baye's theorem.** If $E_1, E_2, ..., E_n$ are mutually disjoint events with $P(E_i) \neq 0$, $(i = 1, 2, ..., n)$, then for any arbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have

$$\boxed{P(E_i/A) = \frac{P(E_i)\, P(A/E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A/E_i)}.}$$

- A statistical measure where the likelihood of two events occurring together and the at the same point in time are calculated. Joint probability is the probability of event B occurring at the same time event A occurs.

11.9 GLOSSARY

- ***Inverse Probability:*** Baye's probability is also known as 'inverse probability'.

11.10 ANSWERS TO CHECK YOUR PROGRESS

1. mutually exclusive
2. P(A) + P(B)
3. $P\left(\frac{B}{A}\right)$
4. sample points
5. favourable events
6. True
7. True
8. False
9. True

11.11 TERMINAL AND MODEL QUESTIONS

1. Explain the concept of independent and mutually exclusive events in probability.
2. What is conditional probability? Explain with the help of examples.
3. Explain with an example the Baye's theorem of probability.
4. What is the probability of getting 3 white balls in a draw of 3 balls from a box containing 6 white and 5 red balls?
5. A bag contains 6 white and 9 black balls. Four balls are drawn at a time. Find the probability for the first draw to give 4 white and the second to give 4 black balls in each of the following cases:

   (i) The balls are not replaced before the second draw.

   (ii) The balls are replaced before the second draw.
6. Assume that a factory has two machines. Past records show that machine No.1 produced 30% of the items of output and machine No. 2 produced 70% of the items. Further, 5% of the items produced by machine No.1 were defective and only 1% produced by machine No. 2 were defective. If an item is drawn

at random and found to be defective, what is the probability that the defective item was produced by machine No. 1 or by machine No.2?

7. A box contains 6 red, 4 white and 5 black balls. A person draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each colour.
8. A bag contains 5 green and 10 white balls. If one ball is drawn from it, find the chance that the ball drawn is green.
9. A bag contains 3 red, 6 white and 7 blue balls. What is the probability that two balls drawn are white and blue.
10. Two cards are randomly drawn from a pack of 52 cards and thrown away. What is the probability of drawing an ace in a single draw from the remaining 50 cards.
11. If a single draw is made from a well-shuffled pack of cards, what is the probability for the drawn card to be a queen or any card of hearts?
12. What is the chance that a leap year, selected at random will contain 53 Sundays ?
13. Two students *X* and *Y* work independently on a problem. The probability that *X* will solve it is $\frac{3}{4}$ and the probability that *Y* will solve it is $\frac{2}{3}$. What is the probability that the problem will be solved?
14. A man wants to marry a girl having qualities: White complexion–the probability of getting such girl is 1 in 20. Handsome dowry–the probability of getting is 1 in 50. Westernised style–the probability is 1 in 100. Find the probability of his getting married to such a girl, who has all the three qualities.
15. What is the probability of getting all the heads in four throws of a coin?
16. What is the chance of getting two sixes in two rolling of a single die?
17. Find the chance of throwing head or tail alternatively in 3 successive tossings of a coin?
18. Two cards are drawn at random from a well shuffled pack of 52 cards. What is the probability that:

    (a) both are aces,

    (b) both are red,

    (c) at least one is an ace.
19. The probability (i) that *A* can solve a problem in biostatistics is $\frac{4}{5}$, (ii) that *B* can solve it is $\frac{2}{3}$, (iii) that *C* can solve it is $\frac{3}{7}$. If all of them try independently, find the probability that the problem will be solved.

20. A university has to select an examiner from a list of 50 persons, 20 of them are women and 30 men, 10 of them knowing Hindi and 40 not, 15 of them being teacher and the remaining 35 not. What is the probability of the university selecting a Hindi-knowing woman teacher?

21. The probability that a person has blood pressure is $\frac{2}{3}$, and the probability that he will not have diabetes is $\frac{5}{9}$. If probability of having at least one disease is $\frac{4}{5}$, what is the probability that he will have both the diseases?

22. From a pack of 52 cards, a card is drawn at random. What is the probability of getting the card of heart or club or seven?

11.12 REFERENCES

1. Cheema, Col. D.S. (2011), *Operations Research, University* Science Press (An imprint of Laxmi Publications), Delhi
2. Sharma Dr. J.K. (2013), *Operations Research Theory and Applications*, Trinity Press (An imprint of Laxmi Publications), Delhi.

UNIT 12: PROBABILITY DISTRIBUTION

NOTES

Structure

12.0 INTRODUCTION

We have already studied a lot about frequency distributions. These distributions are based upon observations, *i.e.*, the frequencies for different values of the variable, under consideration, are based on actual observation. For example, if an unbiased coin is tossed 100 times, we may get head 57 times. Here, 57 is the observed frequency but theoretically we shall expect 'head', 50 times. In this chapter, we shall study *probability distributions* and *frequency distributions* which are based upon theoretical considerations.

12.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Define random variables and their types
- Explain probability distribution of a discrete random variable
- Explain mean and variance of random variable
- Define continuous probability distribution and its various types

12.2 RANDOM VARIABLE

Let S be the sample space of a given random experiment. A real valued function 'x' defined on the sample space S is called a **random variable.**

Thus, if $s \in S$, then $x(s)$ is a unique real number.

Remark: The values of a random variable are real numbers, connected with the outcomes of the random experiment, under consideration.

In the random experiment of toss of two coins, if we define the random variable (x) as the *number of heads,* then the values of the random variable x are 0, 1, 1, 2 corresponding to the outcomes TT, TH, HT, HH respectively.

We write, $x(TT) = 0^*$, $x(TH) = 1$, $x(HT) = 1$, $x(HH) = 2$.

In case, there are three coins, then the values of this random variable are 0, 1, 1, 1, 2, 2, 2, 3 corresponding to the outcomes TTT, TTH, THT, HTT, HHT, HTH, THH, HHH respectively.

We can define any number of random variables on the same sample space. If x denotes the random variable, defined as the cube of the number of tails, in the experiment of toss of two coins, then we have

Sample points	**HH**	**HT**	**TH**	**TT**
x	$(0)^3 = 0$	$(1)^3 = 1$	$(1)^3 = 1$	$(2)^3 = 8$

Random variables are of two types: (i) discrete random variable and (ii) continuous random variable.

(i) A random variable is called a **discrete random variable** if it can take only finitely many values. For example, in the experiment of drawing three cards from a pack of playing cards, the random variable "number of kings drawn" is a discrete random variable taking value either 0 or 1 or 2 or 3.

(ii) A random variable is called a **continuous random variable** if it can take any value between certain limits. For example, height, weight of students in a class are continuous random variables.

12.3 PROBABILITY DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

Let x be a discrete random variable assuming values $x_1, x_2, x_3, ..., x_n$ corresponding to the various outcomes of a random experiment. If the probability of occurrence

of $x = x_i$ is $P(x_i) = p_i$, $1 \le i \le n$ such that $p_1 + p_2 + p_3 + ... + p_n = 1$, then the function, $P(x_i) = p_i$, $1 \le i \le n$ is called the **probability function** of the random variable x and the set $\{P(x_1), P(x_2), P(x_3), ..., P(x_n)\}$ is called the **probability distribution** of x.

The graph of a probability distribution is also drawn as shown in the diagram Fig. 12.1. This is also known as a **bar-chart.**

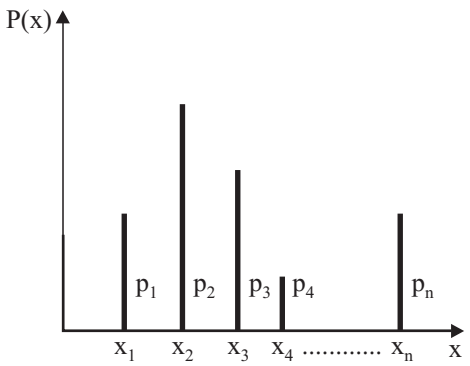


Fig. 12.1: Graph of Probability Distribution

Working Rules for Finding Probability Distribution

I. Identify the random variable and put it as x.

II. Find the possible values of x.

III. Find $P(x)$ for all possible values of x and write P.D. of x.

IV. Check that the sum of all probabilities in the P.D. is one. If this sum is not one, then some mistake is bound to have occurred in the calculation work. Remove the mistake and again verify that the sum of all probabilities is one.

Example 1: *A fair die is tossed once. If the random variable x is the "number of even numbers", find the probability distribution of x.*

Solution: Here S = {1, 2, 3, 4, 5, 6}.

Let x denotes the random variable "number of even numbers".

$\therefore$ x can take values 0 and 1, because at most we can get one even number.

$$P(x = 0) = P(\text{no even number})$$

$$= P(1 \text{ or } 3 \text{ or } 5) = \frac{3}{6} = \frac{1}{2}$$

$$P(x = 1) = P(\text{one even number})$$

$$= P(2 \text{ or } 4 \text{ or } 6) = \frac{3}{6} = \frac{1}{2}$$

$\therefore$ The probability distribution of x is

x	***0***	***1***
P(x)	$\frac{1}{2}$	$\frac{1}{2}$

Example 2: *Find the probability distribution of the random variable "number of heads" when:*

(i) *two coins are tossed* (ii) *one coin is tossed twice.*

Solution: (i) Let S be the sample space.

$\therefore$ S = {HH, HT, TH, TT}

Let x denotes the discrete random variable "*number of heads*".

$\therefore$ The possible values of x are 0, 1, 2.

We have $P(x = 0) = P(\{TT\}) = \frac{1}{4}$

$P(x = 1) = P(\{HT, TH\}) = \frac{2}{4} = \frac{1}{2}$

$P(x = 2) = P(\{HH\}) = \frac{1}{4}$.

$\therefore$ The required probability distribution (P.D.) is

x	0	1	2
$P(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

(ii) Let H be the event of getting a head.

Let x denote the discrete random variable "*number of heads*" in two tosses.

$\therefore$ The possible values of x are 0, 1, 2.

We have $P(x = 0) = P(\overline{H}_1 \overline{H}_2) = P(\overline{H}_1)\, P(\overline{H}_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

$P(x = 1) = P(H_1 \overline{H}_2 \text{ or } \overline{H}_1 H_2) = P(H_1 \overline{H}_2) + P(\overline{H}_1 H_2)$

$= P(H_1)\, P(\overline{H}_2) + P(\overline{H}_1)\, P(H_2) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

$P(x = 2) = P(H_1 H_2) = P(H_1)\, P(H_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

$\therefore$ The required probability distribution (P.D.) is

x	0	1	2
$P(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example 3: *Two cards are drawn successively with replacement from a well shuffled pack of 52 cards. Find the probability distribution of number of queens.*

Solution: Let Q be the event of drawing a queen from the pack of cards.

Let x denotes the discrete random variable *"number of queens"* in two draws.

$\therefore$ The possible values of x are 0, 1, 2.

We have $P(x = 0) = P(\overline{Q}_1 \overline{Q}_2) = P(\overline{Q}_1)\, P(\overline{Q}_2) = \frac{48}{52} \times \frac{48}{52} = \frac{144}{169}$

NOTES

$$P(x = 1) = P(Q_1\overline{Q}_2 \text{ or } \overline{Q}_1 Q_2) = P(Q_1\overline{Q}_2) + P(\overline{Q}_1 Q_2)$$

$$= P(Q_1)\,P(\overline{Q}_2) + P(\overline{Q}_1)\,P(Q_2) = \frac{4}{52}\times\frac{48}{52} + \frac{48}{52}\times\frac{4}{52} = \frac{24}{169}$$

$$P(x = 2) = P(Q_1Q_2) = P(Q_1)\,P(Q_2) = \frac{4}{52}\times\frac{4}{52} = \frac{1}{169}.$$

$\therefore$ The required probability distribution (P.D.) is

x	0	1	2
$P(x)$	$\frac{144}{169}$	$\frac{24}{169}$	$\frac{1}{169}$

12.4 MEAN AND VARIANCE OF A RANDOM VARIABLE

We know the method of finding the mean and variance of frequency distributions. In a frequency distribution, we have frequencies corresponding to different values of the variable. Similarly in a probability distribution, we have probabilities corresponding to different admissible values of the discrete random variable.

Now, we shall extend the idea of mean and variance for probability distributions.

Let x be a discrete random variable assuming values $x_1, x_2, \ldots\ldots, x_n$ with respective probabilities $p_1, p_2, \ldots\ldots, p_n$ with $p_1 + p_2 + \ldots\ldots + p_n = 1$.

We define, $$\text{mean } (\mu) = \frac{\sum_{i=1}^{n} p_i x_i}{\sum_{i=1}^{n} p_i} = \sum_{i=1}^{n} p_i x_i \qquad \left(\because \quad \sum_{i=1}^{n} p_i = 1\right)$$

and $$\text{variance} = \frac{\sum_{i=1}^{n} p_i (x_i - \mu)^2}{\sum_{i=1}^{n} p_i} = \sum_{i=1}^{n} p_i (x_i - \mu)^2.$$

We have $$\sum_{i=1}^{n} p_i (x_i - \mu)^2 = \sum_{i=1}^{n} p_i (x_i^2 + \mu^2 + 2\mu x_i)$$

$$= \sum_{i=1}^{n} p_i x_i^2 + \mu^2 \sum_{i=1}^{n} p_i - 2\mu \sum_{i=1}^{n} p_i x_i$$

$$= \sum_{i=1}^{n} p_i x_i^2 + \mu^2.1 - 2\mu.\mu = \sum_{i=1}^{n} p_i x_i^2 - \mu^2.$$

$\therefore$ Mean $(\mu) = \sum_{i=1}^{n} \mathbf{p_i x_i}$ and **variance** $(\sigma^2) = \sum_{i=1}^{n} \mathbf{p_i x_i^2} - \mu^2$.

In short, we write, $\mu = \Sigma \mathbf{px}$ and variance $= \Sigma \mathbf{px^2} - \mu^2$.

The mean of random variable x is called the **expected value** of x and is denoted by E(x). The mean and variance of a random variable are also referred to as the mean and variance of the corresponding P.D.

Remark: S.D. of probability distribution $= \sqrt{\text{variance}} = \sqrt{\Sigma px^2 - \mu^2}$.

Working Rules for Solving Problems

I. Identify the random variable (x) and its possible values x_1, x_2, ...

II. Find probabilities for all values of the variable x.

III. Draw table and find Σpx and Σpx^2.

IV. Find mean and variance by using the formulae $\mu = \Sigma px$ and variance $= \Sigma px^2 - \mu^2$.

Check Your Progress

Fill in the blanks:

1. A real valued function x defined on the ..... is called a random variable.

2. A ..... can take any value between certain limits.

3. The mean of random variable x is called ..... of x and is denoted by E(x).

4. Discrete random variable can take only .....

5. Standard deviation of probability distribution

$= \sqrt{\quad}$.

Example 4: *A random variable x has the following probability distribution:*

x	*– 2*	*– 1*	*0*	*1*	*2*	*3*
P(x)	*0.1*	*k*	*0.2*	*2k*	*0.3*	*k*

(*i*) *Find the value of k.*

(*ii*) *Calculate mean and variance of x.*

Solution:

Calculation of Mean and Cariance

x	p	px	px^2
– 2	0.1	– 0.2	0.4
– 1	k	$-k$	k
0	0.2	0	0
1	$2k$	$2k$	$2k$
2	0.3	0.6	1.2
3	k	$3k$	$9k$
	$\Sigma p = 4k + 0.6$	$\Sigma px = 4k + 0.4$	$\Sigma px^2 = 12k + 1.6$

(i) In a P.D., we have $\Sigma p = 1$.

$\therefore \quad 4k + 0.6 = 1 \quad$ or $\quad 4k = 0.4$ or $k = 0.1$.

(ii) Mean $(\mu) = \Sigma px = 4k + 0.4 = 4(0.1) + 0.4 = 0.8$.

Variance $= \Sigma px^2 - \mu^2 = (12k + 1.6) - (0.8)^2 = (12(0.1) + 1.6) - 0.64 = 2.16$.

Example 5: *A die is tossed twice. A "success" is "getting an odd number" on a random toss. Find the mean and variance of the number of successes.*

Solution: Let E be the event of getting a success, *i.e.,* of getting an odd number in the toss of a die. On a die, odd numbers are 1, 3, 5.

$$\therefore \quad P(E) = \frac{3}{6} = \frac{1}{2} \text{ and } P(\overline{E}) = 1 - \frac{1}{2} = \frac{1}{2}.$$

Let x denotes the random variable "*number of successes*".

$\therefore$ The possible values of x are 0, 1, 2.

$$P(x = 0) = P(\overline{E}_1 \overline{E}_2) = P(\overline{E}_1)\, P(\overline{E}_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(x = 1) = P(E_1 \overline{E}_2 \text{ or } \overline{E}_1 E_2) = P(E_1)\, P(\overline{E}_2) + P(\overline{E}_1)\, P(E_2)$$

$$= \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{2}$$

$$P(x = 2) = P(E_1 E_2) = P(E_1)P(E_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

Calculation of Mean and Variance

x	p	px	px^2
0	$\frac{1}{4}$	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{1}{2}$	1
	$\Sigma p = 1$	$\Sigma px = 1$	$\Sigma px^2 = \frac{3}{2}$

Mean $(\mu) = \Sigma px = 1$

Variance $= \Sigma px^2 - \mu^2 = \frac{3}{2} - (1)^2 = 0.5.$

Example 6: *Find the mean and variance of the number of heads in the two tosses of a coin.*

Solution: Let x denotes the random variable, "*number of heads*" in the two tosses.

$\therefore$ The possible values of x are 0, 1, 2.

We have $P(x = 0) = P(\text{no head}) = P(\overline{H}_1\overline{H}_2) = P(\overline{H}_1)P(\overline{H}_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

where H_i is the event of getting head in the ith toss, $i = 1, 2$.

$P(x = 1) = P(\text{one head}) = P(H_1\overline{H}_2 \text{ or } \overline{H}_1H_2) = P(H_1)P(\overline{H}_2) + P(\overline{H}_1)\,P(H_2)$

$$= \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{2}$$

$P(x = 2) = P(\text{both heads}) = P(H_1H_2) = P(H_1)P(H_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$

Calculation of Mean and Variance

x	p	px	px^2
0	$\frac{1}{4}$	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{1}{2}$	1
	$\Sigma p = 1$	$\Sigma px = 1$	$\Sigma px^2 = \frac{3}{2}$

$\therefore$ Mean (μ) = $\Sigma px = 1$

Variance = $\Sigma px^2 - \mu^2 = \frac{3}{2} - (1)^2 = 0.5$.

NOTES

12.5 CONTINUOUS PROBABILITY DISTRIBUTIONS

Since continuous random variables such as height, time, weight, monetary values, length of life of a particular product, etc. can take large number of both integer and non-integer values. The sum of the probability to each of these values is no longer sum to 1.

Unlike discrete random variables, continuous random variables do not have probability distribution functions specifying the exact probabilities of their specified values. Instead, probability distribution is created by distributing one unit of probability along the real line. Such a distribution (also called *probability density function*) determines probabilities that the random variable falls into a specified interval of values.

Certain characteristics of probability density function for the continuous random variable x, are follows:

(i) Area under a continuous probability distribution is equal to 1.

(ii) Probability P($a \leq x \leq b$) of random variable, x, value will fall in an interval from a to b is equal to the area under the *probability density function curve* between the points (values) a and b.

Since nature follows a predictable pattern for many kinds of measurements, therefore most numerical values of a random variable are spread around the center. A frequency distribution of values of random variable observed in nature which follows this pattern is approximately bell shaped. Thus, such distribution of measurements is called a **normal curve (or distribution).**

German mathematician Karl Friedrich Gauss developed the concept of normal distribution (also known as *Gaussian distribution*). Normal distribution is used to study a continuous phenomenon or process such as daily changes in the stock market index, frequency of arrivals of customers at a bank, frequency of telephone calls into a switch board, customer servicing times and so on.

Normal distribution: A continuous probability distribution in which the mean of the distribution lies at the center of the curve and the curve is symmetrical around a vertical line erected at the mean. The tails of the curve extend indefinitely parallel to the horizontal axis.

Normal Probability Distribution Function

The formula for normal probability distribution is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-1/2)[(x-\mu)/\sigma]^2}, \quad -\infty < x < \infty \qquad \ldots(12.1)$$

where π is constant 3.1416, e is constant 2.7183, μ is the mean of the normal distribution, σ is standard of normal distribution and $f(x)$ = relative frequencies (height of the curve) within which values of random variable x fall.

The graph of a normal probability distribution with mean μ and standard deviation σ is shown in Fig. 12.2. The distribution is symmetric about its mean μ that falls at the centre of curve. Since the total area under the normal probability distribution is equal to 1, the symmetry implies that the area on either side of μ is 50 per cent or 0.5. The *shape* of the distribution is determined by μ and σ values.

In symbols, if a random variable x follows normal probability distribution with mean μ and standard deviation σ, then it is also expressed as $x \sim N(\mu, \sigma)$.

Characteristics of normal probability distribution: The shape of normal distribution varies according to the value of mean, μ and/or standard deviation σ. Larger the value of the standard deviation σ, the wider and flatter is the normal curve showing more variability in the data. Thus, standard deviation σ determines the range of values that any random variable is likely to assume. Figure 12.2 (a) shows three normal distributions with different values of the mean μ and a fixed standard deviation σ while in Fig. 12.2 (b) normal distributions are shown with different values of the standard deviation σ and a fixed mean μ.

From Figs. 12.2 (a) and 12.2 (b), the following characteristics of a normal distribution and its density function may be derived:

(i) For every pair of values of μ and σ, the normal probaility density function curve is bell shaped and symmetric. The mean μ determines the *central location* of the normal distribution while standard deviation σ determines its *spread.*

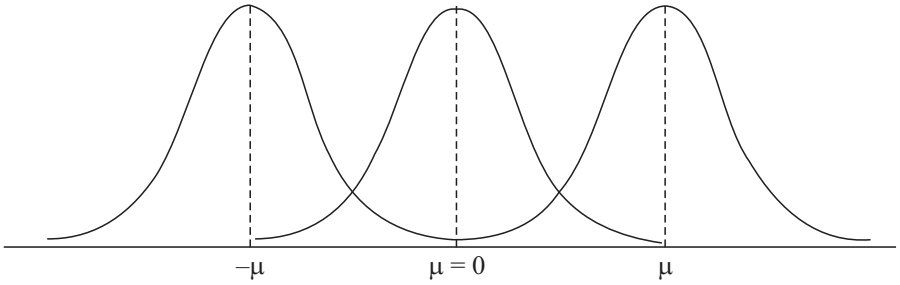


Fig. 12.2: (a) Normal Distributions with Different Mean Values but Fixed Standard Deviation

NOTES

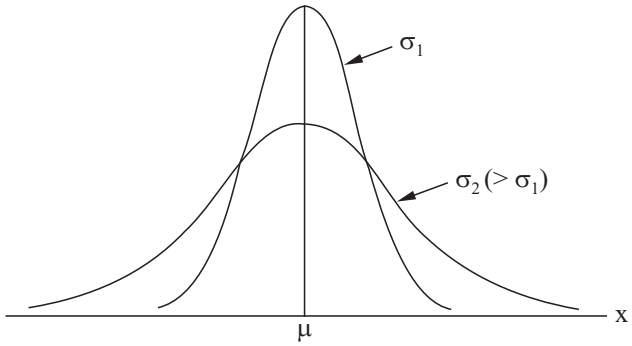


Fig. 12.2: (b) Normal Distributions with Fixed Mean and Variable Standard Deviation

(ii) The normal curve is symmetrical around a vertical line erected at the mean μ with respect to the area under it, *i.e.* 50 per cent of the area of the curve lies on both sides of the mean, μ. This implies that the probability of any random variable whose value is above or below the mean will be same. Thus, for any normal random variable x, $P(x \leq \mu) = P(x \geq \mu) = 0.50$.

(iii) The values of mean, median and mode for the normal distribution are equal because the highest value of the probability density function occurs when value of a random variable, $x = \mu$.

(iv) The two tails of the normal curve extend to infinity in both directions and never touch the horizontal axis.

(v) The mean of the normal distribution may be negative, zero or positive as shown in Fig. 12.2 (a).

(vi) The area under the normal curve represents probabilities for the normal random variable, and therefore, the total area under the curve for the normal probability distribution is 1.

Standard normal probability distribution: A normal probability distribution with mean equal to zero and standard deviation equal to one.

Standard normal probability distribution: To deal with problems where the normal probability distribution is applicable, the value of random variable x is standarized by expressing it as the number of standard deviations (σ) lying on both sides of its mean (μ). Such *standardized normal random variable, z* (also called *z-statistic, z-score* or *normal variate*) is defined as

$$z = \frac{x - \mu}{\sigma}$$

or equivalently $x = \mu + z\sigma$

The z*-statistic* measures the number of standard deviations that any value of the random variable x falls from the mean.

(i) When x is less than the mean (μ), the value of z is negative.

(ii) When x is more than the mean (μ), the value of z is positive.

(iii) When $x = \mu$, the value of $z = 0$.

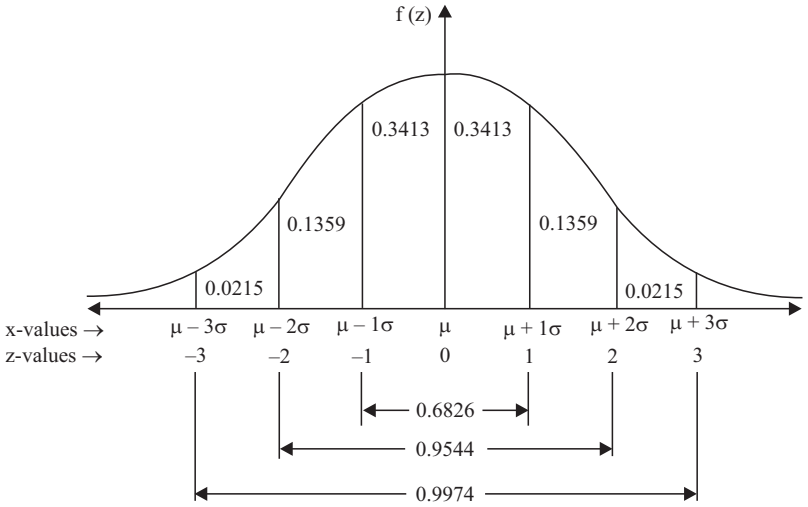


Fig. 12.3: Standard Normal Distribution

Any normal probability distribution with parameters μ and σ can be converted into another distribution called **standard normal probability distribution** as shown in Fig. 12.3 with mean $\mu_z = 0$ and standard deviation $\sigma_z = 1$ with the help of equation 12.1.

Since *z-statistic* measures the number of standard deviations that any value of the random variable x falls from the mean, therfore value of z obtained represents the area under the normal curve. For example, $z = \pm 2$ implies that the value of x is 2 standard deviations above or below the mean (μ).

Area under the normal curve: Since tails of normal curve does not touch x-axis, the range of normal distribution is infinite in both the directions away from μ. This implies that as x moves away from μ, the pdf $f(x)$ approaches x-axis but never actually touches it.

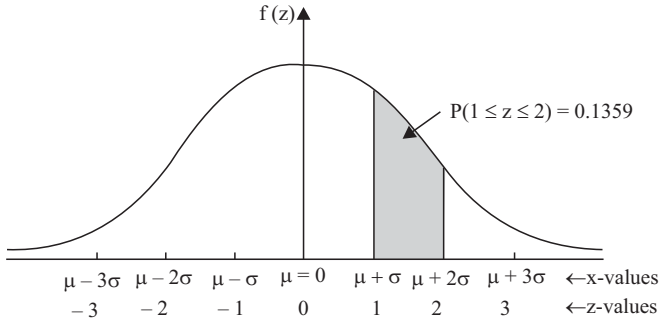


Fig. 12.4: Diagram for Finding P(1 < z < 2)

The area under the standard normal distribution between the mean $z = 0$ and $z = z_0$ (a specified positive value of z) can be read from standard normal (z) table. For

example, area between $1 \leq z \leq 2$ is the proportion of the area under the curve which lies between the vertical lines erected at two points along the x-axis. Also, if the distance between x and μ is one standard deviation or $(x - \mu)/\sigma = 1$, then 34.134 per cent of the distribution lies the between x and μ. Similarly, if x is at 2σ away from μ, *i.e.* $(x - \mu)/\sigma = 2$, then the area will include 47.725 per cent of the distribution as shown in Table 12.1.

Table 12.1: Area Under the Normal Curve

$z = \frac{x - \mu}{\sigma}$	**Area Under Normal Curve Between x and μ**
1.0	0.34134
2.0	0.47725
3.0	0.49875
4.0	0.49997

The percentage of the area of the normal distribution lying within the given range is shown in Table 12.2 and Fig. 12.4.

Table 12.2: Percentage of the Area of the Normal Distribution Lying within the Given Range

Number of Standard Deviations from Mean	**Approximate Percentage of Area under Normal Curve**
$x \pm \sigma$	68.26
$x \pm 2\sigma$	95.45
$x \pm 3\sigma$	99.75

Since standard normal distribution is a symmetrical distribution, therefore

$$P(0 \leq z \leq a) = P(-a \leq z \leq 0) \text{ for any value } a.$$

For example, $P(1 \leq z \leq 2) = P(z \leq 2) - P(z \leq 1) = 0.9772 - 0.8413 = 0.1359$

Approximation of Binomial and Poisson Distributions to Normal Distribution

The binomial distribution approaches a normal distribution with standardized variable, z, that is,

$$z = \frac{x - np}{\sqrt{npq}} \sim N(0, 1)$$

However, this approximation works well when both $np \geq 10$ and $npq \geq 10$

Similarly, Poisson distribution also approaches a normal distribution with standardized variable, z, that is,

$$z = \frac{x - \lambda}{\sqrt{\lambda}} \sim N(0, 1)$$

Check Your Progress

State whether the following statements are True or False:

6. Mean of the normal distribution is zero (0) only.
7. For every pair of values of mean (μ) and standard deviation (σ), the normal probability density function curve is bell shaped.
8. The range of normal distribution is finite in both the directions away from μ.
9. Tails of normal curve touches *x*-axis at 2 points.
10. The mean μ determines the central location of the normal distribution.

Example 7: *How would you use the normal distribution to find approximate frequency of exactly 5 successes in 100 trials, the probability of success in each trial being p = 0.1.*

Solution: Let n = number of trials, p = probability of success, and q = probability of failure. Given $n = 100$, $p = 0.1$, and $q = 0.9$.

Hence, for binomial distribution, mean $\mu = np = 100 \times 0.1 = 10$,

and standard deviation $(\sigma) = \sqrt{npq} = \sqrt{100 \times 0.1 \times 0.9} = 3$.

When number of trials is large, binomial distribution tends to approximate normal distribution. Frequency of exactly 5 successes in 100 trials of binomial distribution will correspond to the frequency of class interval 4.5 to 5.5 and standard deviation of binomial distribution will correspond to mean and standard deviation of normal distribution;

$$z = \frac{4.5 - np}{\sqrt{npq}} = \frac{4.5 - 10}{3} = -1.83.$$

Standard normal variate correspondent to 5.5 is

$$z = \frac{x - \mu}{\sigma} = \frac{5.5 - np}{\sqrt{npq}} = \frac{5.5 - 10}{3} = -1.50.$$

The area under normal curve between $z = -1.83$ and $z = -1.50$ is $0.668 - 0.336 = 0.0332$. Hence, approximate frequency of exactly 5 successes in 100 trials is $0.0332 \times 100 = 3.32$.

Example 8: *When an aptitude test for selecting officers in a bank was conducted on 1,000 candidates, the average score is 42 and the standard deviation of scores is 24. Assuming normal distribution for the scores, find (a) number of candidates whose score exceeds 58, and (b) number of candidates whose scores lie between 30 and 66.*

Solution: (a) Number of candidates whose score exceeds 58:

$$z = \frac{x - \mu}{\sigma} = \frac{58 - 42}{24} = 0.667.$$

NOTES

Area under normal curve for $z = 0.667$ is $(0.5 - 0.2476) = 0.2524$. Thus the number of candidates whose score exceeds 58 is $1000 \times 0.2524 = 252.4$ or 252.

(b) Number of candidates whose scores lie between 30 and 66.

Standard normal variate corresponding to 30 is:

$$z_1 = \frac{x - \mu}{\sigma} = \frac{30 - 42}{24} = -0.5.$$

Standard normal variate corresponding to 66 is:

$$z_2 = \frac{x - \mu}{\sigma} = \frac{60 - 42}{24} = 1.$$

The area under normal curve between $z_1 = -0.5$ and $z_2 = 1$ is .1915 + 0.3413 = 0.5328. Hence, the number of candidates whose scores lie between 30 and 66 is $1000 \times 0.5328 = 532.8$ or 533.

Example 9: *There are 600 business students in the post-graduate of a university, and the probability for any student to need a copy of particular textbook from the university library on any day is 0.05. How many copies of the book should be kept in the university library so that the probability may be greater than 0.90 that none of the students needing a copy from the library has to come back disappointed.*

Solution: Let n be the number of students and p be the probability for any student to need a copy of a particular textbook from the university library. Then

Mean, $\mu = np = 600 \times 0.05 = 30$, and

Standard deviation $(\sigma) = \sqrt{npq} = \sqrt{600 \times 0.05 \times 0.95} = 5.34.$

Let x_1 be the number of copies of a textbook required on any day. Then

$$P(x_1) \geq 90\% = 0.9, \textit{ i.e. } P\left[z = \frac{x_1 - \mu}{\sigma}\right] = P\left[z = \frac{x_1 - 30}{5.34}\right] \geq 1.28$$

$$x_1 - 30 \geq 6.835, \textit{ i.e. } x_1 \geq 36.835 = 37.$$

[Since area under normal curve for $z_1 \geq 0.90 = 1.28$.]

12.6 SUMMARY

- Let S be the sample space of a given random experiment. A real valued function 'x' defined on the sample space S is called a **random variable.**
- A random variable is called a **discrete random variable** if it can take only finitely many values.
- A random variable is called a **continuous random variable** if it can take any value between certain limits.

- The mean of random variable x is called the **expected value** of x and is denoted by E(x). The mean and variance of a random variable are also referred to as the mean and variance of the corresponding P.D.
- Unlike discrete random variables, continuous random variables do not have probability distribution functions specifying the exact probabilities of their specified values. Instead, probability distribution is created by distributing one unit of probability along the real line. Such a distribution (also called *probability density function*) determines probabilities that the random variables falls into a specified interval of values.

12.7 GLOSSARY

- ***Discrete Random Variable:*** A random variable is called a discrete random variable
- ***Expected Value:*** The mean of random variable x is called the expected value of x and is denoted by E(x).
- ***Probability Density Function:*** Probability distribution is created by distributing one unit of probability along the real line.
- ***Gaussian Distribution:*** German mathematician Karl Friedrich Gausss developed the concept of normal distribution (also known as *Gaussian distribution*).

12.8 ANSWERS TO CHECK YOUR PROGRESS

1. sample space S
2. continuous random variable
3. expected value
4. finitely many values
5. variance
6. False
7. True
8. False
9. False
10. True

NOTES

12.9 TERMINAL AND MODEL QUESTIONS

1. Find the probability distribution of the random variable "no. of sixes" in two tosses of a die.
2. A pair of unbiased dice is tossed. If the random variable x is the sum of two numbers obtained on two dice, find the probability distribution of x.
3. From a well shuffled pack of 52 cards, 3 cards are drawn one-by-one with replacement. Find the probability distribution of number of queens.
4. Find the mean of the following probability distributions:

(i)

x	1	2	3
$P(x)$	0.4	0.2	0.4

(ii)

x	0	1	2	3	4
$P(x)$	0.1	0.2	0.3	0.2	0.2

5. Find the mean and the variance of the following probability distributions:

(i)

x	2	3	4	5
$P(x)$	0	0.4	0.1	0.5

(ii)

x	1	2	3	4	5	6
$P(x)$	0.1	0.1	0.2	0.3	0.1	0.2

6. Two urns contain 5 black, 4 white balls and 4 black, 5 white balls. One ball is drawn from each urn. Find the mean and variance of the probability distribution of the random variable "no. of white balls drawn".
7. Find the S.D. of probability distribution of the number of aces drawn, when 2 cards are drawn one by one from a pack of playing cards with replacement.
8. Find the S.D. of probability distribution of the number of aces drawn, when 2 cards are drawn one by one from a pack of playing cards without replacement.
9. Three defective pencils are mixed with ten good ones. Two pencils are drawn simultaneously at random. Find the mean number of defective pencils drawn.
10. A draw of 3 balls in made from an urn containing 7 red and 6 black balls. Find the S.D. of the probability distribution of the discrete random variable "number of black balls drawn".
11. Of a large group of men, 4 % are under 60 inches in height and 40 % are between 60 and 45 inches. Assuming a normal distribution, find the mean height and standard deviation.
12. The customer accounts at a certain departmental store have an average balance of ₹ 480 and a standard deviation of ₹ 160. Assuming that the account balances are normally distributed.

(a) What proportion of the accounts is over ₹ 600?

(b) What proportion of the accounts is between ₹ 240 and ₹ 360?

13. 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days.

(a) How many bulbs will expire in less than 90 days?

(b) If it is decided to replace all the bulbs together, what interval should be allowed between replacements if not more than 10 per cent should expire before replacement?

14. The lifetimes of certain kinds of electronic devices have a mean of 300 hours and standard deviation of 25 hours. Assuming that the distribution of these lifetimes, which are measured to the nearest hour, can be approximated closely with a normal curve:

(a) Find the probability that any one of these electronic devices will have a lifetime of more than 350 hours.

(b) What percentage will have lifetimes of 300 hours or less?

(c) What percentage will have lifetimes from 220 or 260 hours?

15. In a certain examination, the percentage of passes and distinctions were 46 and 9, respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75, respectively (assume the distribution of marks to be normal).

Also determine what would have been the minimum qualifying marks for admission to a re-examination of the failed candidates, had it been desired that the best 25 per cent of them should be given another opportunity of being examined.

12.10 REFERENCES

1. Cheema, Col. D.S. (2011), *Operations Research,* University Science Press (An imprint of Laxmi Publications), Delhi.

2. Sharma Dr. J.K. (2013), *Operations Research Theory and Applications,* Trinity Press (An imprint of Laxmi Publications), Delhi.

UNIT 13: BINOMIAL DISTRIBUTION

NOTES

Structure

13.0 INTRODUCTION

In the previous unit you have learnt various types of probability distributions like discrete and continuous distributions. We know that a real valued function defined on the sample space of a random experiment is called random variable. A random variable is either discrete or continuous.

The binomial distribution is a particular type of probability distribution. This was discovered by **James Bernoulli** (1654–1705) in the year 1700. This distribution mainly deals with attributes. An attribute is either present or absent with respect to elements of a population. For example, if a coin is tossed, we get either *head* or *tail*. The workers of a factory may be classified as *skilled* and *unskilled.*

In this unit you will study binomial probability distribution and its properties. You will learn how to apply binomial distribution in various situations in real life.

13.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- List various conditions for applicability of Binomial distribution
- Define Binomial variable and Binomial probability function

- Define and explain Binomial frequency distribution
- Describe various properties of Binomial distribution

13.2 CONDITIONS FOR APPLICABILITY OF BINOMIAL DISTRIBUTION

The following conditions are essential for the applicability of Binomial Distribution:

(i) ***The random experiment is performed for a finite and fixed number of trials:*** If in an experiment, a coin is tossed repeatedly or a ball is drawn from an urn repeatedly, then each toss or draw is called a **trial.** For example, if a coin is tossed 6 times, then this experiment has 6 trials. The number of trials is an experiment is generally denoted by '*n*'.

(ii) ***The trials are independent:*** By this we mean that the result of a particular trial should not affect the result of any other trial. For example, if a coin is tossed or a die is thrown, then the trials would be independent. If from a pack of playing cards, some draws of one card are made without replacing the cards, then the trials would not be independent. But, if the card drawn is replaced before the next draw, the trials would be independent.

(iii) ***Each trial must result in either "success" or "failure":*** In other words, in every trial, there should be only two possible outcomes *i.e., success* or *failure*. For example, if a coin is tossed, then either *head or tail* is observed. Similarly, if an item is examined, it is either *defective* or *non-defective*.

(iv) ***The probability of success in each trial is same:*** In other words, this condition requires that the probability of *success* should not change in different trials. For example, if a sample of two items is drawn, then the probability of exactly one being defective will be constant in different trials provided the items are replaced before the next draw.

13.3 BINOMIAL VARIABLE AND BINOMIAL PROBABILITY FUNCTION

Binomial Variable

A random variable which counts the number of successes in a random experiment with trials satisfying above four conditions is called a **binomial variable.**

For example, if a coin is tossed 5 times and the event of getting head is *success,* then the possible values of the binomial variable are 0, 1, 2, 3, 4, 5. This is so, because, the minimum number of successes is 0 and the maximum number of successes is 5.

Binomial Probability Function

When a fair coin is tossed, we have only two possibilities: head and tail. Let us call the occurrence of head as 'success'. Therefore, the occurrence of tail would be a 'failure'. Let this coin be tossed 5 times. Suppose we are interested in finding the probability of getting 4 heads and 1 tail *i.e.,* of getting 4 successes. If S and F denote 'success' and 'failure' in a trial respectively, then there are ${}^5C_4 = 5$ ways of having 4 successes.

These are: SSSSF, SSSFS, SSFSS, SFSSS, FSSSS.

The probability of getting 4 successes in each case is $\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)$, because the trials are independent.

$\therefore$ By using *addition theorem*, the required probability of having 4 successes is ${}^5C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)$, which is equal to $\frac{5}{32}$. Now we shall generalise this method of finding the probabilities for different values of a *binomial variable*.

Let a random experiment satisfying the conditions of **binomial distribution** be performed. Let the number of trials in the experiment be n. Let p denotes the probability of *success* in any trial.

$\therefore$ Probability of failure, $q = 1 - p$.

Let x denotes the binomial variable corresponding to this experiment.

$\therefore$ The possible values of x are 0, 1, 2, ..., n.

If there are r successes in n trials, then there would be $n - r$ failures. One of the ways in which r successes may occur is

SS.....S FF.....F

(r times) ($n - r$ times)

where S and F denote success and failure in the trials.

Now, P(SS ..... SFF ..... F) = P(S)P(S) ..... P(S)P(F)P(F) ..... P(F)

($\because$ the trials are independent)

$= p.p$ ..... $p.q.q$ ..... $q = p^r\, q^{n-r}$.

We know that ${}^nC_r$ is the number of combinations of n things taking r at a time. Therefore, the number of ways in which r successes can occur in n trials is equal to the number of ways of choosing r trials (for successes) out of total n trials *i.e.,* it is ${}^nC_r$. Therefore, there are ${}^nC_r$ ways in which we get r successes and $n - r$ failures and the probability of occurrence of each of these ways is $p^r q^{n-r}$. Hence the probability of r successes in n trials in any order is

$P(x = r) = p^r\, q^{n-r} + p^r\, q^{n-r} + ...$ ${}^nC_r$ terms (By addition theorem)

or $\mathbf{P(x = r) = {}^nC_r p^r q^{n-r},\ 0 \le r \le n.}$

This is called the **binomial probability function.** The corresponding **binomial distribution** is

x	0	1	2 ..... n
$P(x)$	${}^nC_0 p^0 q^n$	${}^nC_1 p^1 q^{n-1}$	${}^nC_2 p^2 q^{n-2}$ ..... ${}^nC_n p^n q^0$

The probabilities of 0 success, 1 success, 2 successes, ..., n successes are respectively the 1st, 2nd, 3rd, ..., $(n + 1)$th terms in binomial expansion of $(q + p)^n$. This is why, it is called **binomial distribution.**

Check Your Progress

Fill in the blanks:

1. There are only two possible outcomes of each trial either ..... or .....
2. Binomial variable counts the number of ..... in a random experiment with trials satisfying 4 conditions of binomial distribution.
3. Binomial distribution mainly deals with .....
4. In binomial distribution, the probabilities of 0 success, 1 success, 2 successes, ... n successes are the 1st, 2nd, 3rd ... $(n + 1)$th terms in ..... of $(q + p)^n$.
5. Binomial distribution can be applied only when number of trials are ..... and wol .....

13.4 BINOMIAL FREQUENCY DISTRIBUTION

If a random experiment, satisfying the requirements of Binomial distribution, is repeated N times, then the expected frequency of getting $r(0 \le r \le n)$ successes is given by

$$\mathbf{N.\ P(x = r) = N.\ {}^nC_r\ p^r q^{n-r},\ 0 \le r \le n.}$$

The *frequencies* of getting 0 success, 1 success, 2 successes, ..., n successes are respectively the 1st, 2nd, 3rd, ..., $(n + 1)$th terms in the expansion of $N(q + p)^n$.

Histogram of Binomial Distribution

We know the method of drawing histogram of a frequency distribution. The method of drawing histogram of a binomial distribution is analogous to the procedure of drawing histogram of a frequency distribution. In case of a binomial distribution, we mark all the values of the random variable on the horizontal axis and their respective probabilities on the vertical axis. Rectangles of uniform width are

NOTES

constructed with values of the variable at centre and heights equal to their corresponding probabilities.

Working Rules for Solving Problems

I. Make sure that the trials in the random experiment are independent and each trial result in either 'success' or 'failure'.

II. Define the binomial variable and find the values of n and p from the given data. Also find q by using: $q = 1 - p$.

III. Put the values of n, p and q in the formula:

$$P(r \text{ successes}) = {}^{n}C_{r}\, p^{r}\, q^{n-r},\ r = 0, 1, 2, \ldots\ldots, n \qquad \ldots(1)$$

IV. Express the event, whose probability is desired, in terms of values of the binomial variable x. Use (1) to find the required probability.

Example 1: *An unbiased coin is tossed 10 times. Find, by using binomial distribution, the probability of getting at least 3 heads.*

Solution: Let p be the probability of success, *i.e.*, of getting head in the toss of the coin.

$$\therefore \quad n = 10, \quad p = \frac{1}{2} \quad \text{and} \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

Let x be the binomial variable, "*no. of successes*".

By **Binomial distribution,** $P(x = r) = {}^{n}C_{r}\, p^{r}\, q^{n-r}$, $0 \le r \le n$.

$$\therefore \qquad P(x = r) = {}^{10}C_{r} \left(\frac{1}{2}\right)^{r} \left(\frac{1}{2}\right)^{10-r} = {}^{10}C_{r} \left(\frac{1}{2}\right)^{10}$$

$$= {}^{10}C_{r}\, \frac{1}{1024},\ 0 \le r \le 10.$$

Now, P(at least 3 heads) = $P(x \ge 3) = 1 - P(x < 3)$

$$= 1 - [P(x = 0 \text{ or } x = 1 \text{ or } x = 2)]$$

$$= 1 - [P(x = 0) + P(x = 1) + P(x = 2)]$$

$$= 1 - \left[{}^{10}C_{0}\, \frac{1}{1024} + {}^{10}C_{1}\, \frac{1}{1024} + {}^{10}C_{2}\, \frac{1}{1024}\right]$$

$$= 1 - \frac{1}{1024}\, [{}^{10}C_{0} + {}^{10}C_{1} + {}^{10}C_{2}]$$

$$= 1 - \frac{1}{1024}\, [1 + 10 + 45] = \frac{1024 - 56}{1024}$$

$$= \frac{968}{1024} = \frac{121}{128}.$$

Example 2: *An unbiased coin is tossed six times. Find the probability of obtaining:*

(*i*) *exactly 4 heads*

(*ii*) *less than 3 heads*

(*iii*) *more than 4 heads*

(*iv*) *more than 4 heads and less than 6 heads*

(*v*) *more than 6 heads*

(*vi*) *at least 4 heads*

(*vii*) *at most 4 heads*

(*viii*) *2 heads*

(*ix*) *at least 2 heads.*

Solution: Let p be the probability of success, *i.e.,* of getting head in the toss of the coin.

$$\therefore \quad n = 6, \quad p = \frac{1}{2} \quad \text{and} \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

Let x be the binomial variable, "*no. of successes*".

By **Binomial distribution,** $P(x = r) = {}^{n}C_r \, p^r \, q^{n-r}$, $0 \le r \le n$.

$$\therefore \qquad P(x = r) = {}^{6}C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{6-r} = {}^{6}C_r \left(\frac{1}{2}\right)^6 = {}^{6}C_r \frac{1}{64}, \; 0 \le r \le 6.$$

(i) P(exactly 4 heads) $= P(x = 4) = {}^{6}C_4 \frac{1}{64} = \frac{15}{64}$.

(ii) P(less than 3 heads) $= P(x < 3) = P(x = 0 \text{ or } 1 \text{ or } 2)$

$$= P(x = 0) + P(x = 1) + P(x = 2)$$

$$= {}^{6}C_0 \frac{1}{64} + {}^{6}C_1 \frac{1}{64} + {}^{6}C_2 \frac{1}{64}$$

$$= (1 + 6 + 15) \frac{1}{64} = \frac{22}{64} = \frac{11}{32}.$$

(iii) P(more than 4 heads) $= P(x > 4) = P(x = 5 \text{ or } 6) = P(x = 5) + P(x = 6)$

$$= {}^{6}C_5 \frac{1}{64} + {}^{6}C_6 \frac{1}{64} = (6 + 1) \frac{1}{64} = \frac{7}{64}.$$

(iv) P(more than 4 heads and less than 6 heads)

$$= P(4 < x < 6) = P(x = 5) = {}^{6}C_5 \frac{1}{64} = \frac{6}{64} = \frac{3}{32}.$$

(v) P(more than 6 heads) $= P(x > 6) = 0$. ($\because$ The event is impossible)

(vi) P(at least 4 heads) = P($x \geq 4$)

$$= P(x = 4 \text{ or } 5 \text{ or } 6) = P(x = 4) + P(x = 5) + P(x = 6)$$

$$= {}^6C_4 \frac{1}{64} + {}^6C_5 \frac{1}{64} + {}^6C_6 \frac{1}{64} = (15 + 6 + 1) \frac{1}{64}$$

$$= \frac{22}{64} = \frac{11}{32}.$$

(vii) P(at most 4 heads) = P($x \leq 4$)

$$= P(x = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3 \text{ or } 4)$$

$$= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$$

$$= {}^6C_0 \frac{1}{64} + {}^6C_1 \frac{1}{64} + {}^6C_2 \frac{1}{64} + {}^6C_3 \frac{1}{64} + {}^6C_4 \frac{1}{64}$$

$$= (1 + 6 + 15 + 20 + 15) \frac{1}{64} = \frac{57}{64}.$$

Alternative Method

P(at most 4 heads) = P($x \leq 4$) = 1 – P($x > 4$)

$$= 1 - P(x = 5 \text{ or } 6) = 1 - [P(x = 5) + P(x = 6)]$$

$$= 1 - \left[{}^6C_5 \frac{1}{64} + {}^6C_6 \frac{1}{64} \right] = 1 - (6 + 1) \frac{1}{64}$$

$$= 1 - \frac{7}{64} = \frac{57}{64}.$$

(viii) P(2 heads) = P($x = 2$) = ${}^6C_2 \frac{1}{64} = \frac{15}{64}$.

(ix) P(at least 2 heads) = P($x \geq 2$)

$$= 1 - P(x < 2) = 1 - P(x = 0 \text{ or } 1) = 1 - [P(x = 0) + P(x = 1)]$$

$$= 1 - \left[{}^6C_0 \frac{1}{64} + {}^6C_1 \frac{1}{64} \right] = 1 - \left[\frac{1}{64} + \frac{6}{64} \right] = \frac{57}{64}.$$

Example 3: *A coin is tossed 7 times. What is the probability that head appears an odd number of times.*

Solution: Let p be the probability of success, *i.e.,* of getting a head.

$$\therefore \quad n = 7, \quad p = \frac{1}{2} \quad \text{and} \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

Let x be the Binomial variable "*no. of successes*".

By **Binomial distribution,** $P(x = r) = {}^nC_r\, p^r\, q^{n-r}$, $0 \le r \le n$.

$$\therefore \quad P(x = r) = {}^7C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{7-r} = {}^7C_r \left(\frac{1}{2}\right)^7$$

$$= {}^7C_r \left(\frac{1}{128}\right), \; 0 \le r \le 7.$$

Required probability = P(head appearing an odd number of times)

$$= P(x = 1 \text{ or } 3 \text{ or } 5 \text{ or } 7) = P(x = 1) + P(x = 3) + P(x = 5) + P(x = 7)$$

$$= {}^7C_1 \left(\frac{1}{128}\right) + {}^7C_3 \left(\frac{1}{128}\right) + {}^7C_5 \left(\frac{1}{128}\right) + {}^7C_7 \left(\frac{1}{128}\right)$$

$$= (7 + 35 + 21 + 1)\left(\frac{1}{128}\right) = \frac{64}{128} = \frac{1}{2}.$$

Example 4: *Draw a histogram for the binomial probability distribution of the number of heads in 5 tosses of coin.*

Solution: Let p be the probability of success, *i.e.,* of getting a head.

$$\therefore \quad n = 5, \quad p = \frac{1}{2} \quad \text{and} \quad q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}.$$

Let x be the Binomial variable "no. of successes".

$$\therefore \quad x = 0, 1, 2, \ldots\ldots, 5.$$

By **Binomial distribution,**

$$P(x = r) = {}^nC_r\, p^r\, q^{n-r}, \; 0 \le r \le n.$$

$$\therefore \quad P(x = r) = {}^5C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{5-r} = {}^5C_r \left(\frac{1}{2}\right)^5 = {}^5C_r \left(\frac{1}{32}\right), \; 0 \le r \le 5.$$

$$\therefore \quad P(x = 0) = {}^5C_0 \left(\frac{1}{32}\right) = \frac{1}{32}, \qquad P(x = 1) = {}^5C_1 \left(\frac{1}{32}\right) = \frac{5}{32}$$

$$P(x = 2) = {}^5C_2 \left(\frac{1}{32}\right) = \frac{10}{32}, \qquad P(x = 3) = {}^5C_3 \left(\frac{1}{32}\right) = \frac{10}{32}.$$

$$P(x = 4) = {}^5C_4 \left(\frac{1}{32}\right) = \frac{5}{32}, \qquad P(x = 5) = {}^5C_5 \left(\frac{1}{32}\right) = \frac{1}{32}.$$

$\therefore$ The required probability distribution is

x	0	1	2	3	4	5
$P(x)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$

The histogram of the Binomial Probability Distribution is shown in the figure:

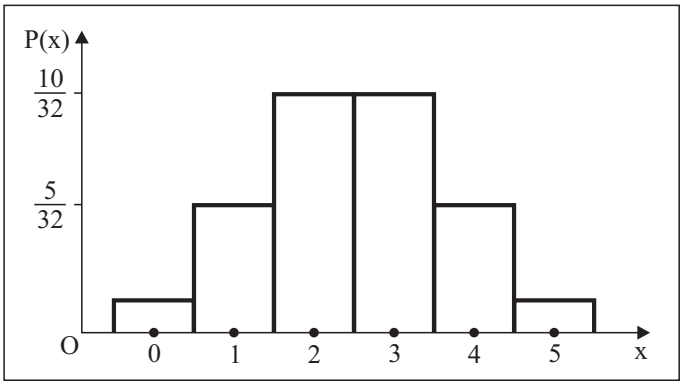


Fig. 13.1: Histogram

13.5 PROPERTIES OF BINOMIAL DISTRIBUTION

Shape of Binomial Distribution

The shape of the binomial distribution depends upon the probability of success (p) and the number of trials in the experiment. If $p = q = \frac{1}{2}$, then the distribution will be symmetrical for every value of n. If $p \neq q$, then the distribution would be asymmetrical, *i.e.*, skewed. The magnitude of skewness varies as the difference between p and q.

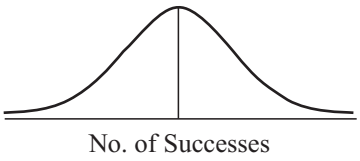


Fig. 13.1

The probabilities in binomial distribution depends upon n and p. These are called the **parameters** of the distribution.

Limiting Case of Binomial Distribution

As number of trials (n) in the binomial distribution increases, the number of successes also increases. If neither p nor q is very small, then as n approaches infinity, the skewness in the distribution disappears and it becomes continuous. Such a continuous, bell shaped distribution is called *normal distribution*. Thus, the normal distribution is limiting case of binomial distribution as n approaches infinity.

Mean of Binomial Distribution

Let x be a binomial variable and $P(x = r) = {}^nC_r p^r q^{n-r}$, $0 \leq r \leq n$.

Here n is the number of trials and p, the probability of success in a trial.

The *mean* of x is the average number of successes.

$$\therefore \quad \text{Mean, } \mu = \sum_{r=0}^{n} r.\mathrm{P}(x = r) = \sum_{r=0}^{n} r.{}^{n}\mathrm{C}_r\, p^r q^{n-r}$$

$$= 0 \,.\, {}^{n}\mathrm{C}_0\, p^0 q^n + 1.{}^{n}\mathrm{C}_1\, p^1 q^{n-1} + 2.{}^{n}\mathrm{C}_2\, p^2 q^{n-2} + \ldots\ldots + n.{}^{n}\mathrm{C}_n\, p^n \,.\, q^0$$

$$= 0 + n \,.\, pq^{n-1} + \frac{n(n-1)}{1.2}\, p^2 q^{n-2} + \ldots\ldots + n.1.p^n$$

$$= np\left\{q^{n-1} + \frac{n-1}{1}\, pq^{n-2} + \ldots\ldots\ldots + p^{n-1}\right\}$$

$$= np\,\{{}^{n-1}\mathrm{C}_0\, p^0 q^{n-1} + {}^{n-1}\mathrm{C}_1\, p^1 q^{n-2} + \ldots\ldots + {}^{n-1}C_{n-1}\, p^{n-1}\, q^0\}$$

$$= np(q+p)^{n-1} = np\,(1)^{n-1} = np.$$

$\therefore$ **Mean (μ) of x = np.**

Variance and S.D. of Binomial Distribution

Let x be a binomial variable and $\mathrm{P}(x = r) = {}^{n}\mathrm{C}_r\, p^r q^{n-r}$, $0 \le r \le n$.

Here n is the number of trials and p, the probability of success in a trial.

The variance and standard deviation of x measures the dispersion of the binomial distribution and are given by

$$\text{Variance} = \sum_{r=0}^{n} r^2 .\mathrm{P}(x = r) - \mu^2$$

and

$$\text{S.D.} = \sqrt{\sum_{r=0}^{n} x^2.\mathrm{P}\,(x = r) - \mu^2}\ .$$

Now,

$$\sum_{r=0}^{n} r^2 .\mathrm{P}(x = r) = \sum_{r=0}^{n} r^2 .{}^{n}\mathrm{C}_r\, p^r q^{n-r}$$

$$= 0.{}^{n}\mathrm{C}_0\, p^0 q^n + 1^2.{}^{n}\mathrm{C}_1\, p^1 q^{n-1} + 2^2.{}^{n}\mathrm{C}_2\, p^2 q^{n-2}$$

$$+ \ 3^2.{}^{n}\mathrm{C}_3\, p^n q^{n-3} + \ldots\ldots + n^2 \,.\, {}^{n}\mathrm{C}_n\, p^n q^0$$

$$= 0 + 1 \,.\, \frac{n}{1}\, pq^{n-1} + 2^2 \,.\, \frac{n(n-1)}{1.2}\, p^2 q^{n-2} + \frac{3^2 \,.\, n(n-1)(n-2)}{1.2.3}\, p^3 q^{n-3} + \ldots\ldots$$

$$+ n^2 \,.\, 1 \,.\, p^n \,.\, 1$$

$$= np\left\{q^{n-1} + \frac{2(n-1)}{1}\, pq^{n-2} + \frac{3(n-1)(n-2)}{1\times 2}\, p^2 q^{n-3} + \ldots. + np^{n-1}\right\}$$

$$= np\left\{\left(q^{n-1} + \frac{(n-1)}{1}\, pq^{n-2} + \frac{(n-1)(n-2)}{1\times 2}\, p^2 q^{n-3} + \ldots. + p^{n-1}\right)\right.$$

$$\left. + \left(\frac{(n-1)}{1}\, pq^{n-2} + \frac{2(n-1)(n-2)}{1\times 2}\, p^2 q^{n-3} + \ldots. + (n-1)\, p^{n-1}\right)\right\}$$

NOTES

$$= np\ \{(q+p)^{n-1} + (n-1)\ p(q^{n-2} + (n-2)\ pq^{n-3} + ... + p^{n-2})\}$$

$$= np\ \{1 + (n-1)\ p(q+p)^{n-2}\} = np\ \{1 + (n\ -1)\ p.1\}$$

$$= np\ \{1 + np - p\} = np + n^2p^2 - np^2.$$

$$\therefore \quad \text{Variance} = \sum_{r=0}^{n} x^2\ .\mathrm{P}(x = r) - \mu^2 = (np + n^2p^2 - np^2) - n^2p^2$$

$$= np - np^2 = np(1-p) = \mathbf{npq.}$$

Also, $\mathbf{S.D.} = \sqrt{\mathbf{Variance}} = \sqrt{\mathbf{npq}}$.

γ_1 and γ_2 of Binomial Distribution

The values of γ_1 and γ_2 for the binomial probability function

$$\mathrm{P}(x = r) = {}^{n}\mathrm{C}_r\ p^r q^{n-r},\ 0 \le r \le n$$

are given by $\gamma_1 = \dfrac{\mathbf{q - p}}{\sqrt{\mathbf{npq}}}$ and $\gamma_2 = \dfrac{\mathbf{1 - 6pq}}{\mathbf{npq}}$.

Recurrence Formula for Binomial Distribution

Let x be a binomial variable and $\mathrm{P}(x = r) = {}^{n}\mathrm{C}_r p^r q^{n-r}$, $0 \le r \le n$.

For $0 \le k < n$, $\quad \mathrm{P}(k) = {}^{n}\mathrm{C}_k\ p^k\ q^{n-k}$

and $\quad \mathrm{P}(k+1) = {}^{n}\mathrm{C}_{k+1}\ p^{k+1}\ q^{n-(k+1)}$.

Dividing, we get

$$\frac{\mathrm{P}(k+1)}{\mathrm{P}(k)} = \frac{{}^{n}\mathrm{C}_{k+1}\ p^{k+1}\ q^{n-k-1}}{{}^{n}\mathrm{C}_k p^k q^{n-k}}$$

$$= \frac{n!}{(k+1)!\,(n-(k+1))!} \cdot \frac{k!\,(n-k)!}{n!} \cdot \frac{p}{q}$$

$$= \frac{n-k}{k+1} \cdot \frac{p}{q}.$$

$$\therefore \quad \mathbf{P(k+1) = \frac{n-k}{k+1} \cdot \frac{p}{q}\ P(k) \quad for\ 0 \le k < n.}$$

This is the required **recurrence formula**.

Applications of Binomial Distribution

This distribution is applied to problems concerning:

1. The number of defectives items in a sample.
2. The estimation of reliability of systems.
3. Number of rounds fired from a gun hitting a target.
4. Radar detection.

Check Your Progress

State whether the following statements are True or False:

6. If from a pack of playing cards, some draws of one card are made without replacing the cards, then trials would be independent.
7. In case of binomial distribution, we mark all the values of random variable on the horizontal axis and their respective probabilities on the vertical axis.
8. Normal distribution is elliptical shaped distribution.
9. Mean of binomial distribution is np where n is number of trials and p is probability of success.
10. Recurrence formula for binomial distribution is

$$P(k+1) = \frac{n-k}{k+1}, \frac{p}{q} P(k) \qquad \text{for } 0 \le k \le n$$

Example 5: *Find the expectation of the number of heads in 15 tosses of a coin.*

Solution: Here $n = 15$. Let p be the probability of getting a head in a trial, *i.e.,* in a loss.

$$\therefore \qquad p = \frac{1}{2}.$$

Let x be the Binomial variable "*no. of heads*".

$$\therefore \quad \text{Expectation of } x = E(x) = \text{mean} = np = 15 \times \frac{1}{2} = 7.5.$$

Example 6: *Obtain the binomial distribution whose mean is 10 and standard deviation is* $2\sqrt{2}$.

Solution: Let number of trials = n and probability of success = p.

$\therefore$ P(r successes) = ${}^nC_r\, p^r\, q^{n-r}$, $0 \le r \le n$.

We have mean = $np = 10$ and S.D. = $\sqrt{npq} = 2\sqrt{2}$.

$$\therefore \qquad \sqrt{10q} = \sqrt{8} \quad \Rightarrow \quad q = \frac{8}{10} = \frac{4}{5}$$

$$\therefore \qquad p = 1 - q = 1 - \frac{4}{5} = \frac{1}{5}$$

$$\therefore \qquad np = 10 \quad \Rightarrow \quad n\left(\frac{1}{5}\right) = 10 \quad \Rightarrow \quad n = 50$$

$$\therefore \quad \textbf{P(r successes)} = {}^{50}\mathbf{C_r} \left(\frac{\mathbf{1}}{\mathbf{5}}\right)^{\mathbf{r}} \left(\frac{\mathbf{4}}{\mathbf{5}}\right)^{\mathbf{50-r}}, \; 0 \le r \le 50.$$

Example 7: *A discrete random variable x has mean score equal to '6' and variance equal to '2'. Assuming that the underlying distribution of x is binomial, what is the probability when* $5 \le x \le 6$.

Solution: We have mean = $np = 6$ ...(1)

and variance = $npq = 2$ ...(2)

(1) and (2) $\Rightarrow$ $6 \times q = 2 \Rightarrow q = \frac{2}{6} = \frac{1}{3}$.

$\therefore$ $p = 1 - q = 1 - \frac{1}{3} = \frac{2}{3}$

(1) $\Rightarrow$ $n\left(\frac{2}{3}\right) = 6 \Rightarrow n = 9$.

$\therefore$ $P(r \text{ successes}) = {}^{n}C_{r}\, p^{r}\, q^{n-r},\ 0 \le r \le n$

$$= {}^{9}C_{r}\left(\frac{2}{3}\right)^{r}\left(\frac{1}{3}\right)^{9-r},\ 0 \le r \le 9.$$

$\therefore$ $P(5 \le x \le 6) = P(x = 5 \text{ or } x = 6) = P(x = 5) + P(x = 6)$

$$= {}^{9}C_{5}\left(\frac{2}{3}\right)^{5}\left(\frac{1}{3}\right)^{4} + {}^{9}C_{6}\left(\frac{2}{3}\right)^{6}\left(\frac{1}{3}\right)^{3}$$

$$= \frac{1}{3^{9}}\,[126 \times 32 + 84 \times 64] = \frac{9408}{3^{9}}.$$

Example 8: *If the sum of the mean and the variance of a binomial distribution of 5 trials is 1.8, then find the binomial distribution.*

Solution: Let the binomial distribution be $P(x = r) = {}^{n}C_{r}\, p^{r}\, q^{n-r},\ 0 \le r \le n$.

$\therefore$ Mean = np and variance = npq

By the given condition, $np + npq = 1.8$ and $n = 5$.

$\Rightarrow$ $5p + 5p(1 - p) = \frac{9}{5} \Rightarrow 25p^{2} - 50p + 9 = 0$

$\therefore$ $p = \frac{1}{5}$ and $q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}$.

$\therefore$ The binomial distribution is $P(x = r) = {}^{5}C_{r}\left(\frac{1}{5}\right)^{r}\left(\frac{4}{5}\right)^{5-r},\ 0 \le r \le 5.$

Check Your Progress

Choose the correct option for the following statements:

11. The probability of success in each trial should be ..... for binomial distribution.

    (a) same (b) different

    (c) 0 (d) 1

12. In every trial there should be only ..... possible outcomes.

    (a) 1 (b) 2

    (c) 3 (d) 4

13. If from a pack of playing cards, the card drawn is replaced before the next draw, the trial would be

    (a) independent (b) dependent

    (c) successful (d) unsuccessful

14. If $p = q = \frac{1}{2}$, then the distribution will be ..... .

    (a) symmetrical (b) asymmetrical

    (c) continuous (d) discontinuous

15. Variance of binomial distribution is given by:

    (a) np (b) npq

    (c) $\sqrt{npq}$ (d) $(npq)^2$

13.6 SUMMARY

- This ditribution mainly deals with attributes. An attributes is either present or absent with respect to elements of a population.
- The random experiment is performed for a finite and fixed number of trials.
- Each trial must result in either "success" or "failure".
- The probability of success in each trial is same.
- The method of drawing histogram of a binomial distribution is analogous to the procedure of drawing histogram of a frequency distribution.
- The shape of the binomial distribution depends upon the probability of success (p) and the number of trials in the experiment. If $p = q = \frac{1}{2}$, then the distribution will be symmetrical for every value of n. If $p \neq q$, then the distribution would be asymmetrical, *i.e.*, skewed.

- As number of trials (n) in the binomial distribution increases, the number of successes also increases.
- The variance and standard deviation of x measures the dispersion of the binomial distribution and are given by

$$\text{Variance} = \sum_{r=0}^{n} r^2 .\text{P}(x = r) - \mu^2$$

and $$\text{S.D.} = \sqrt{\sum_{r=0}^{n} x^2 .\text{P}(x = r) - \mu^2} .$$

13.7 GLOSSARY

- ***Random Variable:*** A real valued function defined on the sample space of a random experiment is called random variable. A random variable can be either discrete or continuous.
- ***Trial:*** If in an experiment, a coin is tossed repeatedly or a ball is drawn from an urn repeatedly, then each toss or drawn is called trial.
- ***Binomial Variable:*** A random variable which counts the number of successes in a random experiment with trials satisfying above 4 conditions is called binomial variable.
- ***Normal Distribution:*** A continuous, bell shaped distribution which is a limiting case of binomial distribution is called normal distribution.
- ***Binomial Probability Function:*** The probability function which specifies the number of times (x) that an event occurs in n independent trials where p is the probability of the event occurring in a single trial.
- ***Parameters:*** If n is the number of trials in an experiment and p is the probability of success then n and p are called parameters of the distribution.

13.8 ANSWERS TO CHECK YOUR PROGRESS

1. success, failure
2. successes
3. attributes
4. binomial expansion
5. finite and fixed

6. False
7. True
8. False
9. True
10. True
11. (a)
12. (b)
13. (a)
14. (a)
15. (c)

13.9 TERMINAL AND MODEL QUESTIONS

1. A unbiased coin is tossed 32 times. Find the probability of obtaining 2 heads.
2. An unbiased coin is tossed 8 times. Find by using binomial distribution, the probability of getting at least 3 heads.
3. A box contains 100 tickets each bearing one of the numbers from 1 to 100. If 5 tickets are drawn successively with replacement from the box, find the probability that all the tickets bear number divisible by 10.
4. If the chance that any of 5 telephone lines is busy at any instance is 0.01, what is the probability that all the lines are busy? What is the probability that more than 3 lines are busy?
5. Assume that on an average one telephone number out of 15 called between 2 P.M. and 3 P.M. on week days is busy. What is the probability that if six randomly selected telephone numbers are called, at least three of them will be busy?
6. A bag contains 10 balls each marked with one of the digits 0 to 9. If four balls are drawn successively with replacement from the bag, what is the probability that none is marked with the digit '0'?
7. In a box containing 100 bulbs, 10 are defective. What is the probability that out of a sample of 5 bulbs

   (i) none is defective? (ii) exactly two are defective?
8. A bag contains 25 items of which 5 are defective. A random sample of two is drawn (without replacement). What is the probability that (i) of both being good (ii) of both being bad (iii) at least one being good.

NOTES

9. Five men in a group of 20 are graduates. If 3 men are picked out of 20 at random, then what is the probability that (i) all are graduates (ii) at least one is graduate.

10. (a) The sum of mean and variance of a binomial variance is 15 and the sum of their squares is 117. Find the distribution.

    (b) The sum and the product of the mean and variance of a binomial distribution are 24 and 128 respectively. Find the distribution.

    (c) If the probability of a defective bulb is 0.1, find the mean and the standard deviation of defective bulbs in a total of 900.

11. If the probability of getting a defective bolt is 0.1, find the mean and standard deviation for the distribution of defective bolts in a total of 500 bolts.

12. (i) Determine the binomial distribution whose mean is 5 and S.D. is $\sqrt{2.5}$.

    (ii) The mean and variance of a binomial distribution are 12 and 3 respectively. Find the probability distribution.

    (iii) Test the validity of the following statement "The mean of the binomial distribution and the standard deviation is 3".

    (iv) Examine the validity of the following statement "The mean of the binomial distribution is 10 and variance is 16".

    (v) Is there any fallacy in the statement "the mean of a Binomial distribution is 4 and its variance is 6".

13. Determine the probability of 3 successes in a binomial distribution whose mean and variance are respectively 2 and 1.5.

14. A die is thrown 6 times. What is the probability that there will be (i) no ace (ii) not more than one ace (iii) more than 4 aces? Find also the mean and the variance of the number of aces (Note that 'ace' means the number 1 (or one dot) on the die).

15. Calculate $P(x$ successes) for $x = 1, 2, 3, 4$ and 5, taking $n = 5$ and $p = \frac{1}{6}$ with the help of the recurrence formula of the binomial distribution. Hence, draw a histogram for this distribution.

16. Suppose that a radio tube inserted into a certain type of set has a probability 0.2 of functioning more than 500 hours. If we test 4 tubes, what is the probability that exactly k of these function for more than 500 hours, where $k = 0, 1, 2, 3$ and 4? Also draw a histogram for this distribution.

17. The following data are the number of seeds germinating out of 10 on damp filter paper for 80 sets of seeds. Fit a binomial distribution to these data.

x	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
f	*6*	*20*	*28*	*12*	*8*	*6*	*0*	*0*	*0*	*0*	*0*

18. There are 20% chances for a worker of an industry to suffer from an occupational disease. 50 workers were selected at random and examined for the occupational disease. Find the probability that (i) only one worker is found suffering from the disease; (ii) more than 3 are suffering from the disease; (iii) none is suffering from the disease.

19. Assuming that half the population is vegetarian so that the chance of an individual being vegetarian is $\frac{1}{2}$ and assuming that 100 investigators can take samples of 10 individuals to see whether they are vegetarian, how many investigators would you expect to report that three people or less were vegetarian.

20. Out of 800 families with 3 children each, how many would you expect to have

    (i) all boys (ii) all girls

    (iii) 2 boys and 1 girl (iv) at least one boy

    (v) at the most 2 girls

    Assume equal probabilities for boys and girls.

13.10 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitative Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 14: NORMAL AND POISSON DISTRIBUTION

NOTES

Structure

14.0 INTRODUCTION

In the previous unit you have learnt 'Binomial Distribution' and its various properties. In this unit, you will learn about Poisson Distribution and Normal Distribution and various properties of both. Poisson distribution is also a discrete probability distribution and Normal distribution is a limiting case of Binomial distribution.

The Poisson distribution was discovered by French mathematician **Simon Denis Poisson** (1781–1840) in the year 1837. This distribution deals with the evaluation of probabilities of *rare* events such as "number of car accidents on road", "number of earthquakes in a year", "number of misprints in a book" etc.

The normal distribution is a limiting case of the Binomial distribution under the following conditions:

(1) When n, the number of trials is very large and

(2) p, the probability of a success, is close to $\frac{1}{2}$.

Remark: (i) The normal distribution was first discovered by De'Moivre, in 1733, a French mathematician.

(ii) The normal distribution is a continuous distribution.

14.2 UNIT OBJECTIVES

After reading this unit you will be able to:

- Define the meaning of Poisson's Distribution
- Explain the conditions for applicability of Poisson distribution
- Describe Poisson variable and Poisson probability function
- Explain Poisson's frequency distribution
- Learn various working rules for solving problems related to Poisson distribution
- Explain various properties of Poisson distribution
- Define the term 'Normal Distribution'
- Explain the term 'Normal Variate'
- Define and explain Normal curve and its properties
- List basic properties of Normal distribution
- Explain area property of Normal distribution
- Formulate moments of Normal distribution
- Define 'Error function and Probable error'
- Apply normal distribution in various area of real life

14.2 CONDITIONS FOR APPLICABILITY OF POISSON DISTRIBUTION

The Poisson distribution is derived as a limiting case of the binomial distribution. So, the conditions for the applicability of the Poisson distribution are same as those for the applicability of Binomial distribution. Here the additional requirement is that the probability of 'success' is quite near to zero.

NOTES

14.3 POISSON VARIABLE AND POISSON PROBABILITY FUNCTION

Poisson Variable

A random variable which counts the number of successes in a random experiment with trials satisfying above conditions is called a **Poisson variable.** If the probability of an article being defective is 1/500 and the event of getting a defective article is *success* and samples of 10 articles are checked for defective articles, then the possible values of Poisson variable are 0, 1, 2, ..., 10.

Poisson Probability Function

Let a random experiment satisfying the conditions for Poisson Distribution be performed. Let the number of trials in the experiment be n, which is indefinitely large. Let p denotes the probability of *success* in any trial. We assume that p is indefinitely small, *i.e.*, we are dealing with a rare event. Let x denotes the Poisson variable corresponding to this random experiment.

$\therefore$ The possible values of x are 0, 1, 2, ..., n.

The Poisson distribution is obtained as a limiting case of the corresponding binomial distribution of the experiment under the conditions:

(i) n, the number of trials is indefinitely large, *i.e.,* $n \to \infty$.

(ii) p, the probability of success in a trial is indefinitely small, *i.e.,* $p \to 0$.

(iii) The product np of n and p is constant.

By **Binomial distribution**, $P(x = r) = {}^{n}C_r\, p^r q^{n-r}$, $0 \le r \le n$, where $q = 1 - p$.

Let $np = m$. $\therefore$ $p = \dfrac{m}{n}$ and $q = 1 - p = 1 - \dfrac{m}{n}$.

$$\therefore \quad P(x = r) = \frac{n!}{r!\,(n-r)!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{n(n-1)(n-2)\ldots\ldots(n-(r-1))(n-r)!}{r!\,(n-r)!}\cdot\frac{m^r}{n^r}\left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{m^r}{r!}\cdot\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\ldots\ldots\frac{n-(r-1)}{n}\cdot\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-r}$$

$$= \frac{m^r}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\ldots\left(1-\frac{r-1}{n}\right)\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-r}$$

$$\therefore \quad \lim_{n\to\infty} P(x = r) = \frac{m^r}{r!}\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\lim_{n\to\infty}\left(1-\frac{2}{n}\right)\ldots\ldots\lim_{n\to\infty}\left(1-\frac{r-1}{n}\right)$$

$$\times \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{-r}$$

$$= \frac{m^r}{r!}(1-0)(1-0)\ldots\ldots(1-0)e^{-m}(1-0)^{-r}$$

$$\left[\because \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = e^{-m}\right]$$

$$= \frac{m^r}{r!} \cdot e^{-m} \cdot 1 = \frac{e^{-m}\, m^r}{r!} .$$

$\therefore$ When n is indefinitely large, we have $\mathbf{P(x = r) = \frac{e^{-m}\, m^r}{r!}}$, **r = 0, 1, 2, ...**

This is called the **Poisson probability function.** The corresponding **Poisson distribution** is

x	0	1	2	3 .....
$P(x)$	$\frac{e^{-m}m^0}{0!}$	$\frac{e^{-m}m^1}{1!}$	$\frac{e^{-m}m^2}{2!}$	$\frac{e^{-m}m^3}{3!}$ .....

The constant m is the product of n and p and is called the **parameter** of the Poisson distribution.

14.4 POISSON FREQUENCY DISTRIBUTION

If a random experiment, satisfying the requirements of Poisson distribution, is repeated N times, then the expected frequency of getting $r(0 \le r \le n)$ successes is given by

$$\mathbf{N \cdot P(x = r) = N \frac{e^{-m} r m^r}{r!}} \text{ , } \mathbf{r = 0, 1, 2, ...}$$

14.5 WORKING RULES FOR SOLVING PROBLEMS

I. Make sure that the trials in the random experiment are independent and the success is a rare event and each trial result in either success or failure.

II. Define the Poisson variable and find the value of n and p from the given data. Find $m = np$. Sometimes, the value of m is directly given.

III. Put the value of m in the formula:

$$\text{P}(r \text{ successes}) = \frac{e^{-m}\, m^r}{r!} \text{ , } r = 0, 1, 2, \ldots\ldots, n. \qquad \ldots(1)$$

IV. Express the event, whose probability is desired in terms of values of the Poisson variable x. Use (1) to find the required probability.

Remark 1: The distribution to be used in solving a problem is generally given is the problem. If it is not given, then the student should make use of Poisson distribution only when the event in the problem is of rare nature, *i.e.,* the probability of happening of event is quite near to zero.

Remark 2: The value of e^{-m} required in any particular problem is generally given with the problem itself. Otherwise, the value of e^{-m} can be found out by using the table given in this unit. In the examination hall, generally the table of e^{-m} is available for students. If at all the value of e^{-m} is neither given with the problem nor the table of e^{-m} is supplied in the examination hall, then the students are advised to retain their final result in terms of e^{-m}.

Check Your Progress

Fill in the blanks:

1. Normal distribution is a ..... distribution.
2. Additional requirement for Poisson's distribution is that the probability of 'success' should be nearly .....
3. Poisson's variable counts the number of ..... in a random experiment with trials satisfied.
4. Poisson's distribution deals with the evaluation of probabilities of .....
5. For normal distribution, number of trials should be .....

Example 1: *Out of 100 bulbs sample, the probability of a bulb to be defective is 0.03. Using Poisson distribution, obtain the Probability that in a sample of 100 bulbs, none is defective.* *[Given $e^{-3} = 0.04979$]*

Solution: Let x be the Poisson variable, "*no. of defective bulbs in a sample of 100 bulbs*".

By **Poisson distribution,** $P(x = r) = \dfrac{e^{-m}\, m^r}{r!}$, $r = 0, 1, 2, ...$

Here $n = 100$, $p = 0.03$. $\therefore$ $m = np = 100 \times 0.03 = 3$.

$\therefore$ $P(x = r) = \dfrac{e^{-3}\,(3)^r}{r!}$, $r = 0, 1, 2, ..., 100$.

$\therefore$ P(none is defective) $= P(x = 0) = \dfrac{e^{-3}\,(3)^0}{0!} = \dfrac{0.04979 \times 1}{1} = 0.04979$.

Example 2: *2% of the bolts manufactured by a factory are found to defective. Find the probability that in a packet of 200 bolts not more than 3 bolts will come out to be defective.* *[Take $e^{-4} = 0.0183$]*

Solution: Let x be the Poisson variable, "*no. of defective bolts in a packet of 200 bolts*".

By **Poisson distribution,**

$$P(x = r) = \frac{e^{-m}\, m^r}{r\,!}\ ,\ r = 0, 1, 2, ...$$

Here $n = 2000$, $\quad p = 2\% = \frac{2}{100} \quad \therefore \quad m = np = 200 \times \frac{2}{100} = 4.$

$$\therefore \quad P(x = r) = \frac{e^{-4}\,(4)^r}{r\,!}\ ,\ r = 0, 1, 2, ..., 200.$$

$\therefore$ P(not more than 3 defective) = P($x \le 3$)

$$= P(x = 0 \text{ or } x = 1 \text{ or } x = 2 \text{ or } x = 3)$$

$$= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$= \frac{e^{-4}\,(4)^0}{0\,!} + \frac{e^{-4}\,(4)^1}{1\,!} + \frac{e^{-4}\,(4)^2}{2\,!} + \frac{e^{-4}\,(4)^3}{3\,!}$$

$$= e^{-4}\left[1 + 4 + \frac{16}{2} + \frac{64}{6}\right] = 0.0183\left(\frac{71}{3}\right) = 0.4331.$$

Example 3: *There are 50 telephone lines in an exchange. The probability that any one of them will be busy is 0.1. What is the probability that all the lines are busy?*

Solution: Let x be the Poisson variable, "*no of busy lines in the exchange*".

By **Poisson distribution,** $\quad P(x = r) = \frac{e^{-m}\, m^r}{r\,!}\ ,\ r = 0, 1, 2, ...$

Here $\quad n = 50, p = 0.1. \quad \therefore \quad m = np = 50 \times 0.1 = 5.$

$$\therefore \quad P(x = r) = \frac{e^{-5}\,(5)^r}{r\,!}\ ,\ r = 0, 1, 2, ..., 50.$$

$$\therefore \quad \text{P(all lines are busy)} = P(x = 50) = \frac{e^{-5}\,(5)^{50}}{50\,!}.$$

Example 4: *Eight percent of the bolts produced in a certain factory turns out to be defective. Find the probability, using Poisson distribution, that in a sample of 25 bolts chosen at random, (i) exactly 3 (ii) more than 3, will be defective.*

[Take $e^{-2} = 0.135$]

Solution: Let x be the Poisson variable, "*no. of defective bolts in a sample of 25 bolts*".

By **Poisson distribution,** $\quad P(x = r) = \frac{e^{-m}\, m^r}{r\,!}\ ,\ r = 0, 1, 2, ...$

Here $n = 25, \qquad p = \frac{8}{100} = \frac{2}{25}. \quad \therefore \quad m = np = 25 \times \frac{2}{25} = 2.$

NOTES

$$\therefore \qquad P(x = r) = \frac{e^{-2}(2)^r}{r!},\ r = 0, 1, 2, ..., 25.$$

(i) P(exactly 3 defectives)

$$= \frac{e^{-2}(2)^3}{3!} = \frac{0.135 \times 8}{6} = 0.18. \qquad \text{(Using } e^{-2} = 0.135\text{)}$$

(ii) P(more than 3 defectives) = $P(x > 3) = 1 - P(x \le 3)$

$$= 1 - P(x = 0 \quad \text{or} \quad x = 1 \quad \text{or} \quad x = 2 \quad \text{or} \quad x = 3)$$

$$= 1 - [P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)]$$

$$= 1 - \left[\frac{e^{-2}(2)^0}{0!} + \frac{e^{-2}(2)^1}{1!} + \frac{e^{-2}(2)^2}{2!} + \frac{e^{-2}(2)^3}{3!}\right]$$

$$= 1 - e^{-2}\left[1 + 2 + \frac{4}{2} + \frac{8}{6}\right] = 1 - 0.135\left(1 + 2 + 2 + \frac{4}{3}\right)$$

$$= 1 - 0.135 \times \frac{19}{3} = 0.145.$$

14.6 PROPERTIES OF POISSON DISTRIBUTION

Shape of Poisson Distribution

The shape of the Poisson distribution depends upon the parameter *m*, the average number of successes per unit. As value of *m* increases, the graph of Poisson distribution would get closer to a symmetrical continuous curve.

Special Usefulness of Poisson Distribution

The Poisson distribution is specially used when there are events which do not occur as outcomes of a definite number of trials in an experiment, rather occur randomly in nature. This distribution is used when the event under consideration is rare and casual. In finding probabilities by *Poisson distribution*, we require only the measure of average chance of occurrence (*m*) based on past experience or a small sample drawn for the purpose.

Mean of Poisson Distribution

Let *x* be a Poisson variable and $P(x = r) = \frac{e^{-m}m^r}{r!}$, $r = 0, 1, 2, ...$

The *mean of x* is the average numbers of successes.

$$\therefore \quad \text{Mean,} \qquad \mu = \sum_{r=0} r.P(x = r) = \sum_{r=0} r \cdot \frac{e^{-m}m^r}{r!}$$

$$= 0 \cdot \frac{e^{-m}m^0}{0!} + 1 \cdot \frac{e^{-m}m^1}{1!} + 2 \cdot \frac{e^{-m}m^2}{2!} + 3 \cdot \frac{e^{-m}m^3}{3!} + ...$$

$$= 0 + me^{-m}\left(\frac{1}{1!} + \frac{2m}{2!} + \frac{3m^2}{3!} + ...\right)$$

$$= me^{-m}\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + ...\right)$$

$$= me^{-m} \cdot e^m = me^0 = m \cdot 1 = m.$$

$\therefore$ **Mean (μ) of x = m.**

Variance and S.D. of Poisson Distribution

Let x be a Poisson variable and $P(x = r) = \frac{e^{-m}m^r}{r!}$, $r = 0, 1, 2, ...$

The variance and standard deviation of x measures the dispersion of the Poisson distribution and are given by

$$\text{Variance} = \sum_{r=0} r^2.P(x = r) - \mu^2 \quad \text{and} \quad \text{S.D.} = \sqrt{\sum_{r=0} r^2.P(x = r) - \mu^2}.$$

$$\therefore \quad \sum_{r=0} r^2.P(x = r) = \sum_{r=0} r^2.\frac{e^{-m}m^r}{r!}$$

$$= 0^2 \cdot \frac{e^{-m}m^0}{0!} + 1^2 \frac{e^{-m}m^1}{1!} + 2^2 \frac{e^{-m}m^2}{2!} + 3^2 \frac{e^{-m}m^3}{3!} + 4^2 \frac{e^{-m}m^4}{4!} + ...$$

$$= 0 + me^{-m}\left(\frac{1}{1!} + \frac{2m}{1!} + \frac{3m^2}{2!} + \frac{4m^3}{3!} + ...\right)$$

$$= me^{-m}\left\{\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + ...\right) + \left(\frac{m}{1!} + \frac{2m^2}{2!} + \frac{3m^3}{3!} + ...\right)\right\}$$

$$= me^{-m}\left\{e^m + m\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + ...\right)\right\}$$

$$= me^{-m}\{e^m + me^m\} = me^{-m}e^m(1 + m) = me^0(1 + m)$$

$$= m(1 + m) = m + m^2.$$

$\therefore$ Variance $= \sum_{r=0} r^2.P(x = r) - \mu^2 = (m + m^2) - m^2 =$ **m.**

Also, **S.D. = $\sqrt{\text{Variance}}$ = $\sqrt{\mathbf{m}}$.**

NOTES

γ_1 and γ_2 of Poisson Distribution

The values of γ_1 and γ_2 for the Poisson probability function

$$P(x = r) = \frac{e^{-m}\, m^r}{r\,!}, \; r = 0, 1, 2, ...$$

are given by $\gamma_1 = \dfrac{1}{\sqrt{m}}$ and $\gamma_2 = \dfrac{1}{m}$.

Recurrence Formula for Poisson Distribution

Let x be a Poisson variable and $P(x = r) = \dfrac{e^{-m}\, m^r}{r\,!}$, $r = 0, 1, 2, ...$

For $k \geq 0$, $P(k) = \dfrac{e^{-m}\, m^k}{k\,!}$ and $P(k+1) = \dfrac{e^{-m}\, m^{k+1}}{(k+1)\,!}$.

Dividing, we get $\dfrac{P(k+1)}{P(k)} = \dfrac{e^{-m}\, m^{k+1}}{(k+1)\,!} \cdot \dfrac{k\,!}{e^{-m}\, m^k} = \dfrac{m}{k+1}$.

$$\therefore \quad \mathbf{P(k+1) = \frac{m}{k+1}\, P(k), \quad k = 0, 1, 2, ...}$$

This is the required **recurrence formula.**

Applications of Poisson Distribution

This distribution is applied to problems concerning.

1. The number of persons born blind per year in a country.
2. The number of deaths by horse kick in an army corps.
3. The number of fragments from a shell hitting a target.
4. Demand pattern for certain spare parts.

Example 5: *A pair of dice is thrown 200 times. If getting a sum of 9 is considered as success, using Poisson distribution, find the mean and variance of the number of successes.*

Solution: Let p be the probability of getting sum 9 in a throw of pair of dice. Out of total 36 outcomes, the favourable outcomes are (3, 6), (4, 5), (5, 4) and (6, 3).

$$\therefore \quad p = \frac{4}{36} = \frac{1}{9}.$$

Also, $n = 200$.

$$\therefore \quad m = np = 200 \times \frac{1}{9} = \frac{200}{9} = 22.22 \qquad | \because \text{ Mean = Variance}$$

$\therefore$ Mean = m = 22.22 and variance = m = 22.22.

Example 6: *If a Poisson variate x is such that P(x = 1) = 2. P(x = 2), find the mean and variance of the distribution.*

Solution: Let $P(x = r) = \frac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, ...$

where m is the average number of successes.

We have $P(x = 1) = 2. P(x = 2)$

$$\Rightarrow \quad \frac{e^{-m} m^1}{1!} = 2 . \frac{e^{-m} m^2}{2!}$$

$$\Rightarrow \quad m = m^2 \quad \Rightarrow \quad m = 1 \qquad (\because \quad m \neq 0)$$

$\therefore$ Mean = $m = 1$ and variance = $m = 1$.

Example 7: (*i*) *For a Poisson distribution, it is given that, P(X = 1) = P(X = 2). Find the value of mean of the distribution. Hence find P (X = 0) and P(X = 4).*

(*ii*) *A random variable X follows a Poisson distribution with Parameter 4. Find the Probability that X assumes the values less than 2.*

Solution: (i) Let $P(X = r) = \frac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, ...$ …(1)

where m is the average number of successes.

We have $P(X = 1) = P(X = 2)$

$$\Rightarrow \quad \frac{e^{-m} m^1}{1!} = \frac{e^{-m} m^2}{2!}$$

$$\Rightarrow \quad m = \frac{m^2}{2} \quad \Rightarrow \quad m = 2 \qquad (\because \quad m \neq 0)$$

$\therefore$ Mean of the distribution, $m = 2$.

Using (1), we have $P(X = 0) = \frac{e^{-m} m^0}{0!} = e^{-2}$

and $P(X = 4) = \frac{e^{-m} m^4}{4!} = \frac{e^{-2}(2)^4}{24} = \frac{2}{3} e^{-2}$.

(ii) Here $m = 4$. By using Poisson distribution

we know

$$P(X = r) = \frac{e^{-m} m^r}{r!}, \; r = 0, 1, 2, ... \qquad ...(1)$$

Required Probability $= P(X < 2) = P(X \leq 1)$

$= P(X = 0) + P(X = 1) = e^{-m} + e^{-m} . m$ | Using (1)

$= e^{-4} (1 + 4) = 5e^{-4} = 5 \times 0.0183 = 0.09157.$

Example 8: *A telephone exchange receives on an average 4 calls per minute. Find the probabilities on the basis of Poisson distribution (m = 4), of:*

(*i*) *2 or less calls per minute* (*ii*) *upto 4 calls per minute*

(*iii*) *more than 4 calls per minute.*

Solution: Let x be the Poisson variable "*no. of calls per minute*".

By **Poisson distribution,** $P(x = r) = \frac{e^{-m} m^r}{r!}$, $r = 0, 1, 2, ...$

Here m = Average number of successes *i.e.*, calls per minute = 4

$$\therefore \quad P(x = r) = \frac{e^{-4}(4)^r}{r!}, \; r = 0, 1, 2, ...$$

(*i*) P(2 or less calls per minute) = $P(x \le 2) = P(x = 0) + P(x = 1) + P(x = 2)$

$$= \frac{e^{-4} \cdot 4^0}{0!} + \frac{e^{-4} \cdot 4^1}{1!} + \frac{e^{-4} \cdot 4^2}{2!} = e^{-4} \{1 + 4 + 8\}$$

$$= 0.01832 \times 13 = 0.2382.$$

(*ii*) P(upto 4 calls per minute) = $P(x \le 4) = P(x = 0)$

$$+ P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$$

$$= \frac{e^{-4} \cdot 4^0}{0!} + \frac{e^{-4} \cdot 4^1}{1!} + \frac{e^{-4} \cdot 4^2}{2!} + \frac{e^{-4} \cdot 4^3}{3!} + \frac{e^{-4} \cdot 4^4}{4!}$$

$$= e^{-4} \left\{1 + 4 + 8 + \frac{64}{6} + \frac{256}{24}\right\} = 0.01832 \times 34.3333 = 0.6289.$$

(*iii*) P(more than 4 calls per minute)

$$= P(x > 4) = 1 - P(x \le 4) = 1 - 0.6289 = 0.3711.$$

Check Your Progress

Choose the correct option for the following statements:

6. The Poisson distribution is a limiting case of .....
   (a) Binomial distribution (b) Normal distribution
   (c) Discrete distribution (d) Continuous distribution
7. Parameter of Poisson distribution is
   (a) Number of trials (n) (b) Probability of success (p)
   (c) Product of n and p (d) Probability of failure (q)
8. Poisson's distribution is used when the event under consideration is
   (a) rare (b) casual
   (c) frequent (d) both (a) and (b)

9. Standard deviation of Poisson's distribution is given by

(a) $\sqrt{m}$ (b) m^2

(c) m (d) $\frac{1}{m}$

10. For Poisson's distribution, probability of success should be

(a) indefinitely large (b) indefinitely small

(c) 0 (d) 1

14.7 NORMAL DISTRIBUTION

Introduction

In Binomial and Poission distributions, we considered the probabilities of discrete random variables. Now we shall consider random variables which may take non-countably infinitely many possible values. Such a random varaibale is called a **continuous random variable.** The random variables corresponding to the measurement of height, weight etc. are continuous. We have already discussed probability distributions of discrete random variables. We shall also be considering the probability function of very important continuous random variable, namely *normal variable.*

Probability Function of Continuous Random Variable

In discrete probability distributions, the probability is defined for each and every value of the variable and the sum of all these probabilities is one. On the other hand, continuous random variables are defined over intervals of real numbers which contains non-countably infinitely many numbers. Let x be a contiunous random variable. The probability of x to take any particular value is generally zero. For example, if an individual is selected at random from a large group of males then the probability that his weight (x) is exactly 56 kg (*i.e.*, 56.000 ..... kg) would be zero. On the other hand, the probability that weight (x) lying between 55.600 ..... kg and 56.200 ..... kg need not be zero. Thus, we cannot define a probability function for a continuous random variable as we did in the case of a discrete random variable. In case of a continuous random variable (x), the probability of x taking any particular value is generally zero and practically does not make any sense whereas the probability of x taking values between any two different values is meaningful.

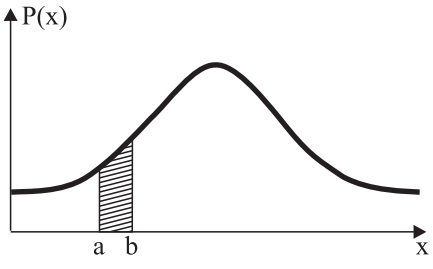


Fig. 14.1

NOTES

For a continuous random variable, x, a function $P(x)$ is called a **Probability function** if:

(i) $P(x) \geq 0$ and (ii) $\int_{-\infty}^{\infty} P(x)\,dx = 1.$

If $P(x)$ is a *probability function* of x, then we define:

$$P(a < x < b) = \int_{a}^{b} P(x)\,dx .$$

Thus, if $P(x)$ is a *probability function* of x, then:

(i) $P(x)$ is non-negative

(ii) area bounded by the curve and x-axis is equal to one

(iii) area bounded by the curve, x-axis and ordinates $x = a, x = b$ gives the measure of the probability that x lies between a and b.

Remark: Since the probability of x taking any particular value is generally zero, we have

$$P(a < x < b) = P(a \leq x < b) = P(a < x \leq b) = P(a \leq x \leq b).$$

14.8 NORMAL CURVE AND ITS PROPERTIES

The graph of the normal distribution is called the **normal curve.**

Properties:

(a) The graph of the normal distribution is bell-shaped and symmetrical about the line $x = u$.

(*i.e.,* If we fold the normal curve about the line $x = \mu$, the two halves coincide).

(b) The normal curve is unimodal.

(c) The line $x = \mu$ divides the area under the normal curves above x-axis into two equal Parts (fig.).

(d) The area under the normal curve between any two given ordinates $x = x_1$ and $x = x_2$, represents the probability of values falling into the given interval.

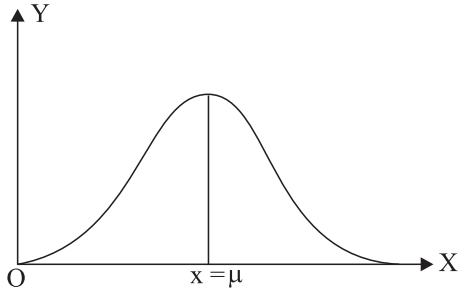


Fig. 14.2

(e) The total area under the normal curve above the x-axis is 1.

14.9 BASIC PROPERTIES OF THE NORMAL DISTRIBUTION

The probability density function (p.d.f) of the normal variate X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

Then the curve $y = f(x)$, known as normal probability curve and satisfies the following properties.

(a) The Normal distribution is symmetrical about the line $x = \mu$

(b) It is unimodal

(c) For a Normal distribution, mean = median = mode

(d) The area bounded by the curve $y = f(x)$, and x-axis is 1 unit, *i.e.*,

$$\int_{-\infty}^{\infty} f(x)\, dx = 1. \text{ Also } f(x) \geq 0$$

(e) The points of inflexion of the Normal curve (can be obtained by putting $\frac{d^2y}{dx^2} = 0$ and verifying that at these points, $\frac{d^3y}{dx^3} \neq 0$) are given by $x = \mu \pm \sigma$.

i.e., these points are equidistant from the mean on either side.

Standard Normal Distribution:

Let x be a normal variable with probability function:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty,$$

where μ and σ are the mean and standard deviation of the distribution respectively.

We define $z = \frac{x-\mu}{\sigma}$.

It can be proved mathematically that z is also a normal variable with mean zero and variance one. A normal with mean zero and variance one is called a **standard normal variable (S.N.V.)**.

In terms of z, the probability function of x reduced to

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, -\infty < x < \infty.$$

Remark: Thus, $x \sim N(\mu, \sigma^2)$ and z = S.N.V. of $x = \frac{x-\mu}{\sigma}$, then $z \sim N(0, 1)$.

14.10 AREA PROPERTY OF NORMAL DISTRIBUTION

1. The area under the normal curve between the ordinates $x = \mu - \sigma$ and $x = \mu + \sigma$, is 68.26%.

2. The area under the normal curve between the ordinates $x = \mu - 2\sigma$ and $x = \mu + 2\sigma$ is 95.44%.

3. The area under the normal curves between the ordinates $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$ is 99.73%.

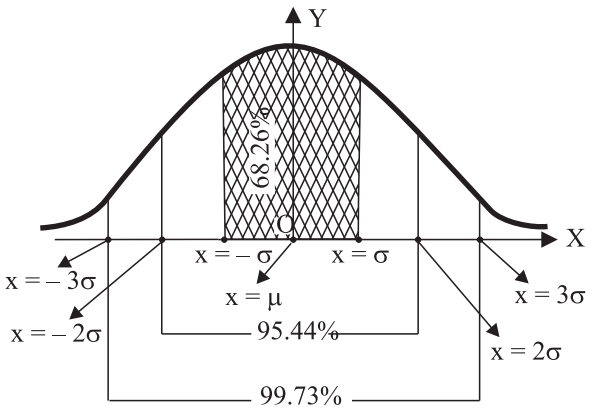


Fig. 14.3: Area Property of Normal Curve.

Table of Area Under Standard Normal Curve

The table titled *area under standard normal curve* is given at the end. Let z^* be any arbitrary but fixed value of the variable z. The first column of the table provides for z values with one decimal digit and the second column gives areas bounded between z-curve and ordinates $z = 0$ and $z = z^*$, which is equal to $P(0 \le z \le z^*)$.

For example, from the table $P(0 \le z \le 1.4) = 0.4192$. The first row of the table provides for the second decimal digit of z^*. For example, $P(0 \le z \le 1.43) = 0.4236$.

Remark: Sometimes, the table for the probabilities $P(-\infty < z \le z^*)$ is given in the examination hall. In such a case, the students should find the value of $P(0 \le z \le z^*)$ by using the following formula:

$P(0 \le z \le z^*) = P(-\infty < z \le z^*) - 0.5.$

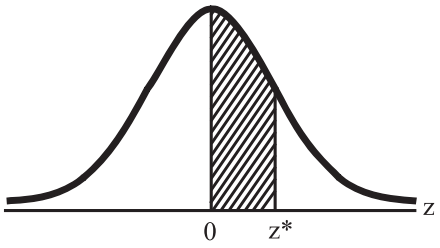


Fig. 14.4

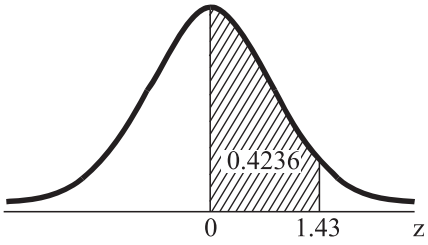


Fig. 14.5

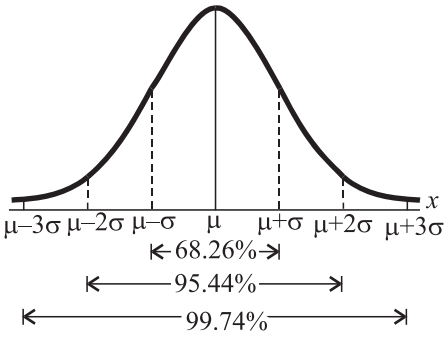


Fig. 14.6

14.11 MOMENTS OF NORMAL DISTRIBUTION

NOTES

Consider $\mu_{2n+1} = \int_{-\infty}^{\infty} (x-\mu)^{2n+1} \cdot f(x)\, dx$

$$= \int_{-\infty}^{\infty} (x-\mu)^{2n+1} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \int_{-\infty}^{\infty} (z\,\sigma)^{2n+1} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} \sigma\, dz$$

$$\left| \text{Put } z = \frac{x-\mu}{\sigma} \Rightarrow dz = \frac{1}{\sigma} dx \right.$$

$$= \sigma^{2n+1} \int_{-\infty}^{\infty} z^{2n+1} e^{-\frac{z^2}{2}} dz$$

$= 0$, since $z^{2n+1} e^{-\frac{z^2}{2}}$ is an odd function.

i.e., **all odd order moments about the mean vanish.**

Further, $\mu_{2n} = \int_{-\infty}^{\infty} (x-\mu)^{2n} \cdot f(x)\, dx$

$$= \int_{-\infty}^{\infty} (x-\mu)^{2n} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} = \int_{-\infty}^{\infty} z^{2n} \cdot e^{-\frac{z^2}{2}} dz$$

$$\left| \text{Put } \frac{x-\mu}{\sigma} = z \Rightarrow dx = \sigma\, dz \right.$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n-1} \left(e^{-\frac{z^2}{2}} z \right) dz \quad \text{Integrating by parts,}$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \left[\left| z^{2n-1} \left(-e^{-\frac{z^2}{2}} \right) \right|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} (2n-1)\, z^{2n-2} e^{-\frac{z^2}{2}} dz \right]$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \left[(0-0) + (2n-1) \int_{-\infty}^{\infty} z^{2n-2} e^{-\frac{z^2}{2}} dz \right]$$

$$= \frac{(2n-1) \cdot \sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n-2} e^{-\frac{z^2}{2}} dz$$

NOTES

$$= (2n-1)\,\sigma^2 \cdot \frac{\sigma^{2n-2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n-2} \cdot e^{-\frac{z^2}{2}}\, dz$$

Hence $\mu_{2n} = (2n-1)\,\sigma^2 \cdot \mu_{2n-4}$ ...(1)

Changing n to $n-1$ in (1), we get

$$\mu_{2n-2} = (2n-3)\,\sigma^2\,\mu_{2n-4} \qquad ...(2)$$

Using (2) in (1) , we get

$$\mu_{2n} = (2n-1)\,(2n-3)\,\sigma^4 .\, \mu_{2n-4}$$

$$.....$$

$$= (2n-1)\,(2n-3)\,.....\,3.1\,.\,\sigma^{2n}\,\mu_{2n-2n}$$

$$= (2n-1)\,(2n-3)\,.....\,3.1\,.\,\sigma^{2n} \qquad | \ \mu_0 = 1$$

Cor. If we put $n = 1, 2,$ in above, we get

$$\mu_2 = \sigma^2,\ \mu_4 = 3.1\,.\,\sigma^4 = 3\,\sigma^4$$

$$\therefore \qquad \beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}} = 0$$

| As odd order moments about the mean vanish

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{3\,\sigma^4}{\sigma^4} = 3$$

14.12 PROBABILITY INTEGRAL OR ERROR FUNCTION

The integral $P(z) = \frac{1}{\sqrt{2\pi}} \int_0^2 e^{-\frac{z^2}{2}}\, dz$, is known as probability integral or error function.

Probable Error (λ)

It is defined as the deviation on either side of the arithmetic mean, the probability of occurrence of which is equal to 0.5

i.e., It is the value of λ, satisfying,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-\lambda}^{\mu+\lambda} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 0.5$$

$$\Rightarrow \quad \frac{1}{\sigma\sqrt{2\pi}} \int_{-\frac{\lambda}{\sigma}}^{\frac{\lambda}{\sigma}} e^{-\frac{z^2}{2}} \cdot \sigma \, dz$$

Put $\frac{x-\mu}{\sigma} = z \quad \Rightarrow \quad dx = \sigma \, dz$

When $x = \mu + \lambda$, $z = \frac{\lambda}{\sigma}$,

When $x = \mu - \lambda$, $z = \frac{-\lambda}{\sigma}$

$$\Rightarrow \quad \frac{2}{\sqrt{2\pi}} \int_{0}^{\frac{\lambda}{\sigma}} e^{-\frac{z^2}{2}} \, dz = 0.5$$

$e^{-\frac{z^2}{2}}$ is an even function

$$\Rightarrow \quad \frac{1}{\sqrt{2\pi}} \int_{0}^{\frac{\lambda}{\sigma}} e^{-\frac{z^2}{2}} \, dz = 0.25$$

Using the table, $\frac{\lambda}{\sigma} = 0.67 \quad \Rightarrow \quad \lambda = 0.67\,\sigma \cong \frac{2}{3}\sigma$

Hence the Probable error $\lambda \cong \frac{2}{3}\,\sigma$.

14.13 APPLICATIONS OF NORMAL DISTRIBUTION

This distribution is applied to Problems concerning:

1. calculation of hit probability of a shot.
2. statistical inference in most branches of science.
3. calculation of errors made by chance in experimental measurements.

Check Your Progress

State whether the following statements are True or False:

11. The points of inflexion of the Normal curve are equidistant from the mean on either side.
12. Normal curve is bimodal.
13. The total area under the normal curve above *x*-axis is 1.
14. Probable error is the deviation on both side of the arithmetic mean.
15. Error function is also called probability integral.

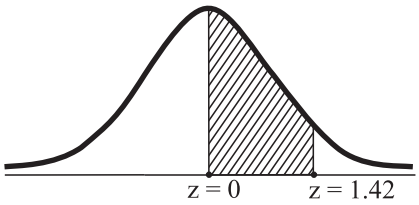
Example 9: *Let z be a standard normal variate, then find*

(i) P(0 ≤ z ≤ 1.42) *(ii) P(z ≥ – 1.28)*

(iii) P(|z| ≤ 0.5) *(iv) P (– 0.73 ≤ z ≤ 0)*

(v) P(0.81 ≤ z ≤ 1.94) *(vi) P (|z| ≥ 10.5)*

(vii) P(– .75 ≤ z ≤ 0).



Solution: (i) We know that

P(0 ≤ z ≤ 1.42) = Area under the standard normal curve between the ordinates z = 0 and z = 1.42

= 0.4222

(ii) P (z ≥ – 1.28) = Area under the standard normal curve to the right of z = – 1.28

= (Area between z = – 1.28 and z = 0)

+ (Area to the right of z = 0)

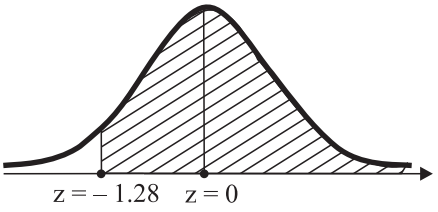
= P (– 1.28 ≤ z ≤ 0) + P (z ≥ 0)

= P (0 ≤ z ≤ 1.28) + P (z ≥ 0)

(due to Symmetry.)

= 0.3997 + 0.5

= 0.89997



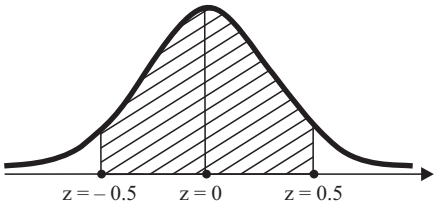
(iii) P(| z| ≤ 0.5) = P(0.5 ≤ z ≤ 0.5)

= Area between z = – 0.5 and z = 0.5

= 2 (Area between z = 0 and z = 0.5)

= 2P(0 ≤ z ≤ 0.5)

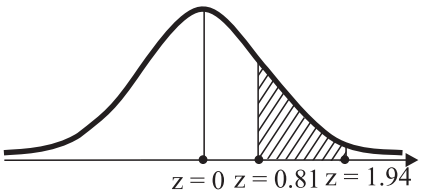
= 2(0.1915) = 0.3830.



(iv) P(– 0.73 ≤ z ≤ 0)

= P(0 ≤ z ≤ 0.73) | By Symmetry

= 0.2673



(v) P(0.81 ≤ z ≤ 1.94)

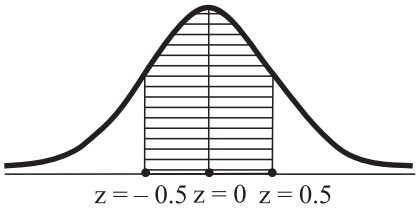
= Area under the normal variate

between z = 0.81 and z = 1.94

= (Area between z = 0 and z = 1.94)

– (Area between z = 0 and z = 0.81)

= P(0 ≤ z ≤ 1.94) – P(0 ≤ z ≤ 0.81)



= 0.4738 – 0.2910

= 0.1828.

(vi) P($|z| \geq 0.5$) = P($z \geq 0.5$ or $z \leq 0.5$) | $|a| \geq b \Rightarrow a \geq b$ or $a \leq -b$

= P ($z \geq 0.5$) + P($z \leq -0.5$)

= (Area to the right of $z = 0.5$) + (Area to the left of $z = -0.5$)

= 2(Area to the right of $z = 0.5$)

= 2 [(Area to the right of $z = 0$) – (Area between $z = 0$ and $z = 0.5$)]

= 2[0.5 – P($0 \leq z \leq 0.5$)

= 2(0.5 – 0.1915) = 2 (0.3085)

= 0.6170.

(vii) P($-.75 \leq z \leq 0$) = P($0 \leq z \leq 0.75$) = 0.2734. | due to symmetry

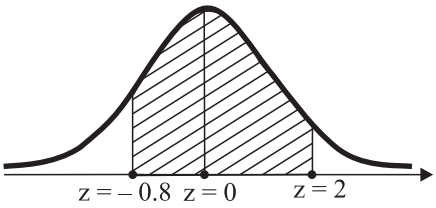
Example 10: *If x is a normal variate with mean 30 and S.D 5. find*

(i) P($26 \leq x \leq 40$) *(ii) P ($x \geq 45$)* *(iii) P($|x - 30| > 5$)*

(iv) If X is normally distributed with mean 5 and standard deviation 2, find P($X > 8$).

Solution: Given x is a normal variate with mean $\mu = 30$ and S.D. $\sigma = 5$.

Let z be the standard normal variate,



then $$z = \frac{x - \mu}{\sigma} = \frac{x - 30}{5}$$

(i) When $x = 26$, $$z = \frac{26 - 30}{5} = -\frac{4}{5} = -0.8$$

When $x = 40$, $$z = \frac{40 - 30}{5} = \frac{10}{5} = 2$$

$\therefore$ P($26 \leq x \leq 40$) = P($-0.8 \leq z \leq 2$)

= Area under the normal variate between $z = -0.8$ and $z = 2$

= (Area between $z = -0.8$ and $z = 0$) + (Area between $z = 0$ and $z = 2$)

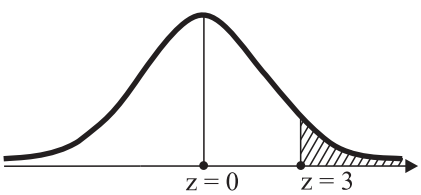
= P($-0.8 \leq z \leq 0$) + P ($0 \leq z \leq 2$)

= P($0 \leq z \leq 0.8$) + P ($0 \leq z \leq 2$)

= 0.2881 + 0.4772

= 0.7563.

(ii) When $x = 45$, $z = \frac{45 - 30}{5} = 3$



$\therefore$ P ($x \geq 45$) = P($z \geq 3$)

= Area under standard normal variate to the right of $z = 3$

= (Area to the right of $z = 0$) – (Area between $z = 0$ and $z = 3$)

= $P(z \geq 0) - P(0 \leq z \leq 3)$

= 0.5 – 0.49865

= 0.00135.

(iii) $P(|x - 30| > 5)$

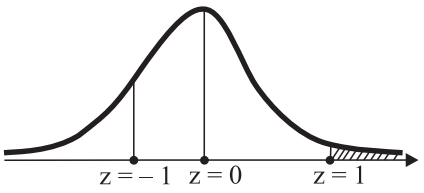
= $(1 - P|x - 30| \leq 5)$ | Note

= $1 - P(30 - 5 \leq x \leq 30 + 5)$

= $1 - P(25 \leq x \leq 35)$ | $(a - b) \leq c \Rightarrow b - c \leq a \leq b + c$

When $x = 25$, $z = \frac{25 - 30}{5} = -1$

When $x = 35$, $z = \frac{35 - 30}{5} = 1$



$P(|x - 30| > 5) = 1 - P(-1 \leq z \leq 1)$

$= 1 - 2\,P(0 \leq z \leq 1)$

$= 1 - 2\,(0.3413)$

$= 1 - 0.6826$

$= 0.3174.$

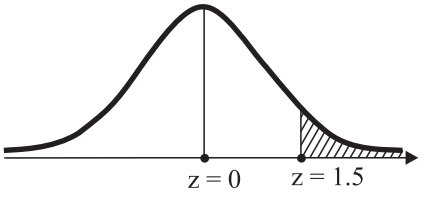
(iv) Given x is a normal variate with mean = $\mu = 5$,

S.D. = $\sigma = 2$

Let z be the standard normal variate,

then $z = \frac{x - \mu}{\sigma} = \frac{x - 5}{2}$

When $x = 8$, $z = \frac{3}{2} = 1.5$



Required probability = $P(x > 8)$

= $P(z > 1.5)$

= Area under standard normal variate to the right of z

= 1.5

= $P(z \geq 0) - P(0 < z < 1.5)$

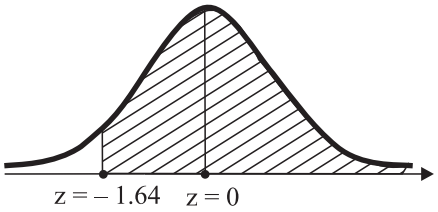
= 0.5 – 0.4332

= 0.0668.

Example 3: *The income of a group of 10,000 persons was found to be normally distributed with mean = ₹750 p.m. and standard deviation = ₹50. Show that of this group about 95% had income exceeding ₹668 and only 5% had income exceeding ₹832. What was the lowest income among the richest?*

Solution: Let x denote the income then, given, x is a normal variate with mean $\mu = 750$ and S.D $\sigma = 50$. Let z be the standard normal variate, then

$$z = \frac{x - \mu}{\sigma} = \frac{x - 750}{50}$$



(i) When $x = 668$, $\quad z = \dfrac{668 - 750}{50}$

$$= \frac{-82}{50} = -1.64$$

$$P(x > 668) = P(z > -1.64)$$

$$= \text{Area to the right of } z = -1.64$$

$$= (\text{Area between } z = -1.64 \text{ and } z = 0) + (\text{Area to the right of } z = 0)$$

$$= P(-1.64 \le z \le 0) + P(z \ge 0)$$

$$= P(0 \le z \le 1.64) + P(z \ge 0)$$

$$= 0.4495 + 0.5 = 0.9495$$

Hence required % of persons having income greater than ₹ 668

$$= 94.95\% \cong 95\%$$

(ii) When $x = 832$, $z = \dfrac{832 - 750}{50} = 1.64$

$\therefore \quad P(x > 832) = P(z > 1.64)$

$$= \text{Area to the right of } z = 1.64$$

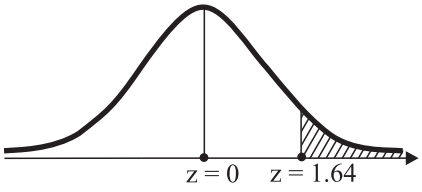
$$= (\text{Area to the right of } z = 0) - (\text{Area between } z = 0 \text{ and } z = 1.64)$$

$$= P(z \ge 0) - P(0 \le z \le 1.64)$$

$$= 0.5 - 0.4495 = 0.0505$$

Hence required % of persons having income greater than ₹ 832 = 5.

Lastly, to find the lowest income among the richest 100, we need to find the value of r such that $P(x \ge r) = 0.01$



NOTES

when $x = r$, $$z = \frac{x - \mu}{\sigma} = \frac{r - 750}{50} = z_1, \text{ say}$$

Now $P(x \geq r) = 0.01$

$\Rightarrow$ $P(z \geq z_1) = 0.01$

$\Rightarrow$ $0.5 - P(0 \leq z \leq z_1) = 0.01$

$\Rightarrow$ $P(0 \leq z \leq z_1) = 0.5 - 0.01 = 0.49$

$\Rightarrow$ $z_1 = 2.33$

$\Rightarrow$ $\frac{r - 750}{50} = 2.33 = r = 750 + 50\ (2.33) = 866.5$

Hence the lowest income among the richest 100 = ₹ 866.50.

14.14 SUMMARY

- Poisson distribution is also a discrete probability distribution and Normal distribution is a limiting case of Binomial distribution.
- The Poisson distribution is derived as a limiting case of the binomial distribution. So, the conditions for the applicability of the Poisson distribution are same as those for the applicability of Binomial distribution.
- The shape of the Poisson distribution depends upon the parameter *m*, the average number of successes per unit.
- The Poisson distribution is specially used when there are events which do not occur as outcomes of a definite number of trials in an experiment, rather occur randomly in nature.
- The random variables corresponding to the measurement of height, weight etc. are continuous.
- In discrete probability distributions, the probability is defined for each and every value of the variable and the sum of all these probabilities is one. On the other hand, continuous random variables are defined over intervals of real numbers which contains non-countably infinitely many numbers.
- The graph of the normal distribution is bell-shaped and symmetrical about the line $x = u$.
- The normal curve is unimodal.
- The total area under the normal curve above the *x*-axis is 1.
- The area under the normal curve between the ordinates $x = \mu - \sigma$ and $x = \mu + \sigma$, is 68.26%.
- The area under the normal curve between the ordinates $x = \mu - 2\sigma$ and $x = \mu + 2\sigma$ is 95.44%.

- The area under the normal curves between the ordinates $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$ is 99.73%.

14.15 GLOSSARY

- ***Poisson Variable:*** A random variable which counts the number of successes on a random experiment with trials satisfying above conditions is called Poisson variable.
- ***Continuous Random Variable:*** Random variables which may take non-countably infinitely many possible values are called continuous random variable.
- ***Poisson Distribution:*** A discrete probability distribution which deals with the evaluations of probabilities of rare events.
- ***Normal Distribution:*** A continuous distribution which is a limiting case of the Binomial distribution under following conditions:

  (a) When n, the number of trials is very large and

  (b) p, the probability of a success, is close to $\frac{1}{2}$.
- ***Standard Normal Variable:*** A normal variable with mean zero and variance one is called standard normal variable.
- ***Normal Probability Curve:*** The probability density function (p.d.f.) of the normal variate X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

$$-\infty < \mu < \infty$$

$$\sigma > 0$$

  Then the curve $y = f(x)$ is known as normal probability curve.

- ***Error Function:*** The integral $P(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{z^2}{2}} dz$, is known as error function. It is also known as probability integral.
- ***Probable Error:*** It is defined as deviation on either side of the arithmetic mean, the probability of occurrence of which is equal to 0.5.

14.16 ANSWERS TO CHECK YOUR PROGRESS

1. continuous
2. near to zero
3. successes
4. rare events
5. very large
6. (a)
7. (c)
8. (d)
9. (a)
10. (b)
11. True
12. False
13. True
14. False
15. True

14.17 TERMINAL AND MODEL QUESTIONS

1. Comment on the following statement: "The mean and variance of a Poisson distribution are equal only if the average occurrence of the Poisson variance is ≤ 4".
2. The standard deviation of a Poisson distribution is 3. Find the probability of getting 3 successes.
3. A telephone exchange receives on an average 3 calls per minute. Find the probability on the basis of the Poisson distribution ($m = 3$), of ;

   (i) exactly 1 call per minute (ii) exactly 3 calls per minute

   (iii) less than 3 calls per minute (iv) more than 1 call per minute.
4. If a random variable x follows Poisson distribution such that $P(x = 2) = 9P(x = 4) + 90P(x = 6)$. Find the mean and variance of x.
5. Suppose that a manufactured product has 2 defects per unit of product inspected. Using Poisson distribution, calculate the probabilities of finding a product without any defect, 3 defects and 4 defects. (Given $e^{-2} = 0.135$)

6. A car-hire firm has two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused.

   ($e^{-1.5} = 0.2231$)

7. The number of accidents in a year attributed to taxi drivers in a city follows Poisson distribution with mean 3. Out of 1,000 taxi drivers, find approximately the number of drivers with (*i*) no accident in a year (*ii*) more than 3 accidents in a year.

   (Given $e^{-1} = 0.3679$, $e^{-2} = 0.1353$, $e^{-3} = 0.0498$)

8. If the probability of a bad reaction from a certain injection is 0.001. Determine the chance that out of 2000 individuals, more than two will get a bad reaction.
9. The incidence of occupational disease in an industry is such that the workmen have a 10% chance of suffering from it. What is the probability that in a group of 7, five or more will suffer from it.

14.18 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitative Techniques*, University Science Press, New Delhi.
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi.

TABLE VALUES OF e^{-m}

$(0 < m < 1)$

m	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0.0	1.0000	0.9900	0.9802	0.9704	0.9608	0.9512	0.9418	0.9324	0.9231	0.9139
0.1	0.9048	0.8958	0.8869	0.8781	0.8694	0.8607	0.8521	0.8437	0.8353	0.8270
0.2	0.8187	0.8106	0.8025	0.7945	0.7866	0.7788	0.7711	0.7634	0.7558	0.7483
0.3	0.7408	0.7334	0.7261	0.7189	0.7118	0.7047	0.6977	0.6907	0.6839	0.6771
0.4	0.6703	0.6636	0.6570	0.6505	0.6440	0.6376	0.6313	0.6250	0.6188	0.6126
0.5	0.6065	0.6005	0.5945	0.5886	0.5827	0.5770	0.5712	0.5655	0.5599	0.5543
0.6	0.5488	0.5434	0.5379	0.5326	0.5273	0.5220	0.5169	0.5117	0.5066	0.5016
0.7	0.4966	0.4916	0.4868	0.4819	0.4771	0.4724	0.4677	0.4630	0.4584	0.4538
0.8	0.4493	0.4449	0.4404	0.4360	0.4317	0.4274	0.4332	0.4190	0.4148	0.4107
0.9	0.4066	0.4025	0.3985	0.3946	0.3906	0.3867	0.3829	0.3791	0.3753	0.3716

NOTES

(**m = 1, 2, 3, ..., 10**)

m	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
e^{-m}	0.36788	0.13534	0.04979	0.01832	0.006738	0.002479	0.000912	0.000335	0.000123	0.000045

Illustrations:

1. $e^{-4.37} = e^{-4} \cdot e^{-.37} = (0.01832)(0.6907) = 0.012654$
2. $e^{-6.05} = e^{-6} \cdot e^{-.05} = (0.002479)(0.9512) = 0.002358.$

10. Students of a class were given a mechanical aptitude test. This marks were found to be normally distributed with mean 60 and standard deviation 5. What per cent of students scored

    (i) more than 60 marks? (ii) less than 56 marks?

    (iii) between 45 and 65 marks?

11. In an examination taken by 500 candidates, the average and the standard deviation of marks obtained (normally distributed) are 40% and 10%. Find approximately

    (i) how many will pass, if 50% is fixed as a minimum?

    (ii) what should be the minimum if 350 candidates are to pass?

    (iii) how many have scored marks above 60%?

12. In a certain examination, the percentage of passes and distinction were 46 and 9 respectively. Estimate the average marks obtained by the candidate, the minimum pass and distinction marks, being 40 and 75 respectively. (Assume the distribution of marks to be normal).
13. Suppose the waist measurements X of 800 girls are normally distributed with mean 66 cm and standard deviation 5 cm. Find the number of girls with waist

    (i) between 65 cm and 70 cm.

    (ii) greater than or equal to 72 cm.

14. Write short note on Normal distribution.
15. Distinguish between Binomial and Normal distribution
16. Distinguish between Poisson and Normal distribution
17. The weekly wages of 1,000 workmen are normally distributed with a mean of ₹ 70 and a standard deviation of ₹ 5. Estimate the number of workers whose weekly wages will be between ₹ 69 and 72.
18. If the salary of workers in a factory is assumed to follow a normal distribution with a mean of ₹ 500 and a S.D. of ₹ 100. Find the number of workers whose salary vary between ₹ 400 and ₹ 650, given the number of workers in the factory is 15000.

MS 104

Quantitative Techniques in Management

Block - IV
Block Title : Operation Reasearch

UTTARAKHAND OPEN UNIVERSITY
SCHOOL OF MANAGEMENT STUDIES AND COMMERCE
University Road, Teenpani By pass, Behind Transport Nagar, Haldwani- 263 139
Phone No: (05946)-261122, 261123, 286055
Toll Free No.: 1800 180 4025
Fax No.: (05946)-264232, e-mail: info@uou.ac.in, som@uou.ac.in
http://www.uou.ac.in
www.blogsomcuou.wordpress.com

Board of Studies

Programme Coordinator

Dr. Manjari Agarwal
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

Units Written By		Unit No.
Text material developed by	Devashish Dutta	
Typeset by	Goswami Associate, Delhi	

Editor(s)

Dr. Hitesh Kumar Pant
Assistant Professor
Department of Management Studies
Kumaun University
Bhimtal Campus

Dr. Manjari Agarwal
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

Er. Sumit Prasad
Assistant Professor
Department of Management Studies
Uttarakhand Open University
Haldwani

ISBN : 978-93-85740-10-7

Edition : 2016 (Restricted Circulation)
Published by : Uttarakhand Open University, Haldwani, Nainital - 263 139
Printed at : Laxmi Publications (P) Ltd., New Delhi
DUO-8159-134.52-QUAN TECH MGMT B-VI

CONTENTS

UNIT 15: LINEAR PROGRAMMING

Structure

15.0 INTRODUCTION

In this unit you will learn about linear programming and its applications in various areas and formulation of Linear Programming problems:

Linear Programming (LP) is a mathematical technique, which is used for allocating limited resources to a number of demands in an optimal manner. When a set of alternatives is available and one wants to select the best, this technique is very helpful. Management wants to make the best use of organizational resources. Human resources, which may be skilled, semi-skilled or unskilled must be put to optimal

NOTES

use. Similarly, the material resources like machines must be used in an effective manner. Time is very important resource and any job must be completed in allotted time. Application of LP requires that the following conditions must be met:

(a) There must be a well-defined objective of the organization such as:

(i) Maximizing profit

(ii) Minimizing cost.

(b) This objective function must be expressed as a linear function of variables involved in decision-making.

(c) There must be a constraint on availability of resources for the objective functions, *i.e.,* for achieving maximum profit or for reducing the cost to a minimum.

LP technique establishes a linear relationship between two or more variables involved in management decisions described above. Linear means it is directly proportional, *i.e.,* if 5 per cent increase in manpower results in 5 per cent increase in output, it is a linear relationship.

(d) Alternative course of action must be available to select the best, for example, if a company is producing four different types of products and wants to cut down one product, which one should stop manufacturing. The problem gives rise to a number of alternatives and so LP can be used.

(e) Objective function must be expressed mathematically, *i.e.,* we must be able to develop a linear mathematical relationship between the objective and its limitation. Linear equations are of first degree, *i.e.,* if we want x and y as the variables the equation $5x + 10y = 20$ is a linear equation in which x and y can assume different values. However, an equation like $5x^2 + 10y^2 = 200$ is not a linear equation, because of the variable x and y are squared, this is a typical second degree equation.

15.1 UNIT OBJECTIVES

After reading this unit you will be able to:

- Define and explain linear programming
- Explain the concept of linear programming problems (LPP)
- Explain various steps involved in graphical method
- Explain various steps involved in simplex method
- State and derive Duality theorems
- Explain duality of simplex method and solve related problems
- Explain dual bounded variables
- Formulate linear programming problems
- Explain the concept of integer programming
- List the various limitations of integer linear programming
- Explain the concept of goal programming

15.2 INTRODUCTION TO LINEAR PROGRAMMING PROBLEMS (LPP)

I. When a problem is identified then the attempt is to make an mathematical model. In decision making all the decisions are taken through some variables which are known as decision variables. In engineering design, these variables are known as design vectors. So in the formation of mathematical model the following **three phases** are carried out:

(i) Identify the decision variables.

(ii) Identify the objective using the decision variables and

(iii) Identify the constraints or restrictions using the decision variables.

Let there be n decision variables $x_1, x_2, ..., x_n$ and the general form of the mathematical model which is called as Mathematical programming problem under decision-making can be stated as follows:

Maximize/Minimize $z = f(x_1, x_2, ..., x_n)$

Subject to, $g_i(x_1, x_2, ..., x_n) \{\leq, \geq \text{or} =\} b_i$

$i = 1, 2, ..., m.$

and the type of the decisions *i.e.,* $x_j \geq 0$

or, $x_j \leq 0$ or x_j's are unrestricted

or combination types decisions.

In the above, if the functions f and g_i $(i = 1, 2, ..., m)$ are all linear, then the model is called "*Linear Programming Problem (LPP)*". If any one function is non-linear then the model is called "*Non-linear Programming Problem (NLPP)*".

II. We define some basic aspects of LPP in the following:

(a) ***Convex set:*** A set X is said to be convex if

$x_1, x_2 \in X$, then for $0 \leq \lambda \leq 1$,

$x_3 = \lambda x_1 + (1 - \lambda)x_2 \in X$

Some examples of convex sets are:

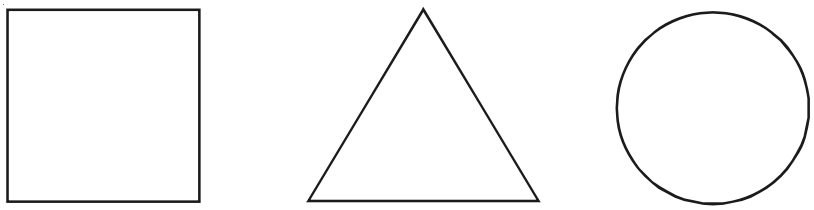


Fig. 15.1: Convex Sets

Some examples of non-convex sets are:

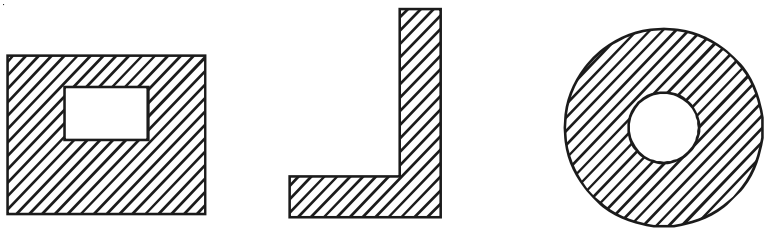


Fig. 15.2: Non-convex Sets

Basically if all the points on a line segment forming by two points lies inside the set/ geometric figure then it is called convex.

(b) ***Extreme point or vertex or corner point of a convex set:*** It is a point in the convex set which can not be expressed as $\lambda x_1 + (1 - \lambda)x_2$ where x_1 and x_2 are any two points in the convex set.

For a triangle, there are three vertices, for a rectangle there are four vertices and for a circle there are infinite number of vertices.

(c) Let $Ax = b$ be the constraints of an LPP. The set $X = \{x \mid Ax = b, x \geq 0\}$ is a convex set.

Feasible Solution: A solution which satisfies all the constraints in LPP is called feasible solution.

Basic Solution: Let m = no. of constraints and n = no. of variables and $m < n$. Then the solution from the system $Ax = b$ is called basic solution. In this system there are ${}^{n}c_{m}$ number of basic solutions. By setting $(n - m)$ variables to zero at a time, the basic solutions are obtained. The variables which is set to zero are known as 'non-basic' variables. Other variables are called basic variables.

Basic Feasible Solution (BFS): A solution which is basic as well as feasible is called basic feasible solution.

Degenerate BFS: If a basic variable takes the value zero in a BFS, then the solution is said to be degenerate.

Optimal BFS: The BFS which optimizes the objective function is called optimal BFS.

Check Your Progress

Fill in the blanks:

1. LP technique establishes ..... between two or more variables involved in management decisions.
2. Linear Programming is a ..... which is used for allocating limited resources to a number of demands in an optimal manner.
3. In engineering design, decision variables are know as .....
4. If a basic variable takes the value zero in a Basic Feasible Solution (BFS), then the solution is said to be ..... .
5. A solution which satisfy all the constraints in LPP is called .....

15.3 GRAPHICAL METHOD

Let us consider the constraint $x_1 + x_2 = 1$. The feasible region of this constraint comprises the set of points on the straight line $x_1 + x_2 = 1$.

If the constraint is $x_1 + x_2 \geq 1$, then the feasible region comprises not only the set of points on the straight line $x_1 + x_2 = 1$ but also the points above the line. Here above means away from origin.

If the constraint is $x_1 + x_2 \leq 1$, then the feasible region comprises not only the set of points on the straight line $x_1 + x_2 = 1$ but also the points below the line. Here below means towards the origin.

The above three cases depicted as follows:

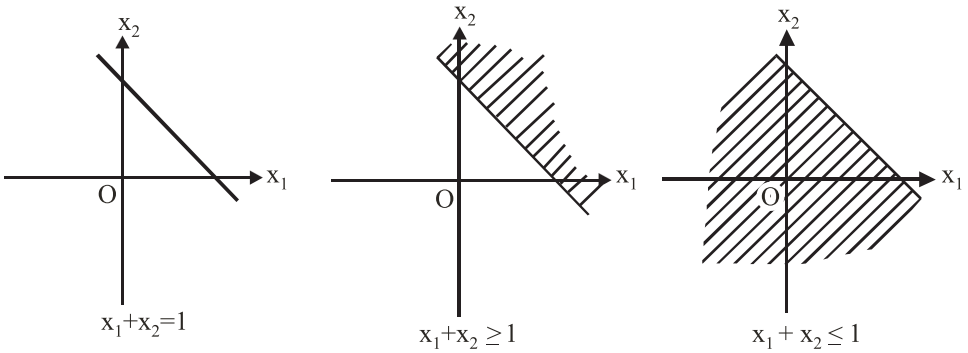


Fig. 15.3

For the constraints $x_1 \geq 1$, $x_1 \leq 1$, $x_2 \geq 1$, $x_2 \leq 1$ the feasible regions are depicted below:

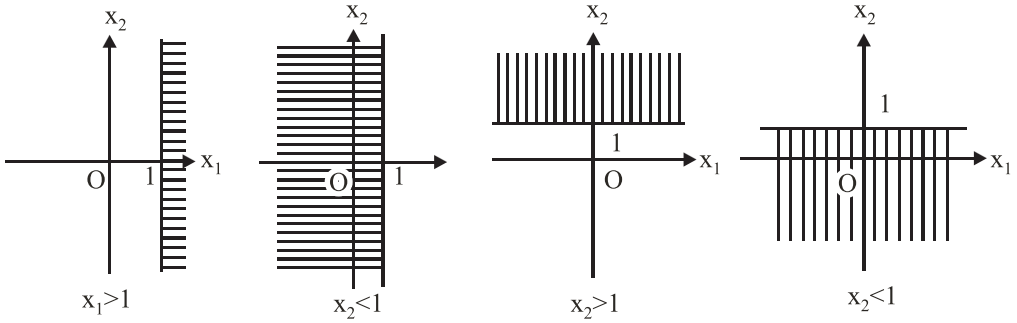


Fig. 15.4

For the constraints $x_1 - x_2 = 0$, $x_1 - x_2 \geq 0$ and $x_1 - x_2 \leq 0$ the feasible regions are depicted in Fig. 15.5.

The *steps of graphical method* can be stated as follows:

(i) Plot all the constraints and identify the individual feasible regions.

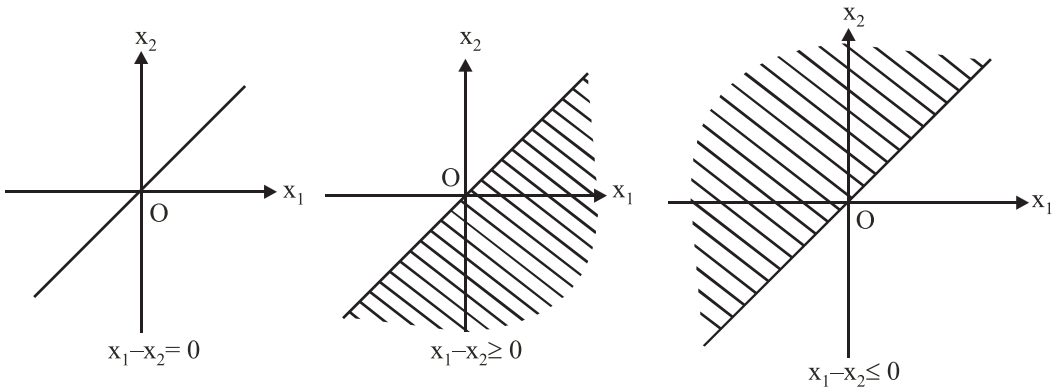


Fig. 15.5

(ii) Identify the common feasible region and identify the corner points *i.e.,* vertices of the common feasible region.

(iii) Identify the optimal solution at the corner points if exists.

Example 1: *Using graphical method solve the following LPP:*

$$\text{Maximize } z = 5x_1 + 3x_2$$

$$\text{Subject to, } 2x_1 + 5x_2 \le 10,$$

$$5x_1 + 2x_2 \le 10,$$

$$2x_1 + 3x_2 \ge 6,$$

$$x_1 \ge 0,\ x_2 \ge 0.$$

Solution: Let us present all the constraints in intercept form *i.e.,*

$$\frac{x_1}{5} + \frac{x_2}{2} \le 1 \qquad \text{...(I)}$$

$$\frac{x_1}{2} + \frac{x_2}{5} \le 1 \qquad \text{...(II)}$$

$$\frac{x_1}{3} + \frac{x_2}{2} \ge 1 \qquad \text{...(III)}$$

The common feasible region ABC is shown in Fig. 1.6 and the individual regions are indicated by arrows. (Due to non-negativity constraints *i.e.,* $x_1 \ge 0$, $x_2 \ge 0$, the common feasible region is obtained in the first quadrant).

The corner points are $A\left(\frac{18}{11}, \frac{10}{11}\right)$, $B\left(\frac{10}{7}, \frac{10}{7}\right)$ and C (0, 2). The value of the objective function at the corner points are $z_A = \frac{120}{11} = 10.91$, $z_B = \frac{80}{7} = 11.43$ and $z_C = 6$.

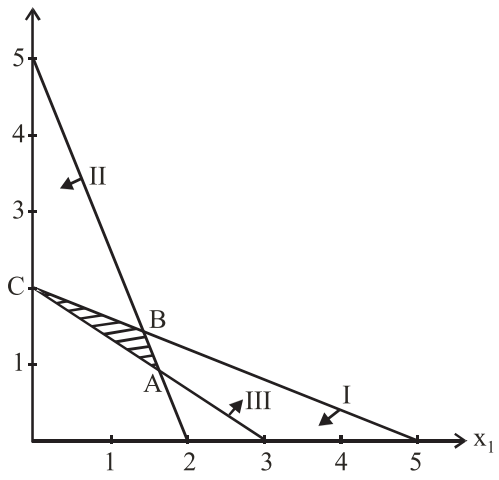


Fig. 15.6

Here the common feasible region is bounded and the maximum has occurred at the corner point B. Hence the optimal solution is

$$x_1^* = \frac{10}{7},\ x_2^* = \frac{10}{7} \text{ and } z^* = \frac{80}{7} = 11.43.$$

Example 2: *Using graphical method solve the following LPP:*

$$\text{Minimize } z = 3x_1 + 10x_2$$

$$\text{Subject to, } 3x_1 + 2x_2 \geq 6,$$

$$4x_1 + x_2 \geq 4,$$

$$2x_1 + 3x_2 \geq 6,$$

$$x_1 \geq 0,\ x_2 \geq 0.$$

Solution: Let us present all the constraints in intercept form *i.e.,*

$$\frac{x_1}{2} + \frac{x_2}{3} \geq 1 \qquad \text{...(I)}$$

$$\frac{x_1}{1} + \frac{x_2}{4} \geq 1 \qquad \text{...(II)}$$

$$\frac{x_1}{3} + \frac{x_2}{2} \geq 1 \qquad \text{...(III)}$$

Due to the non-negativity constraints *i.e.*, $x_1 \geq 0$ and $x_2 \geq 0$ the feasible region will be in the first quadrant.

The common feasible region is shown in Fig. 15.7 where the individual feasible regions are shown by arrows. Here the common feasible region is unbounded.

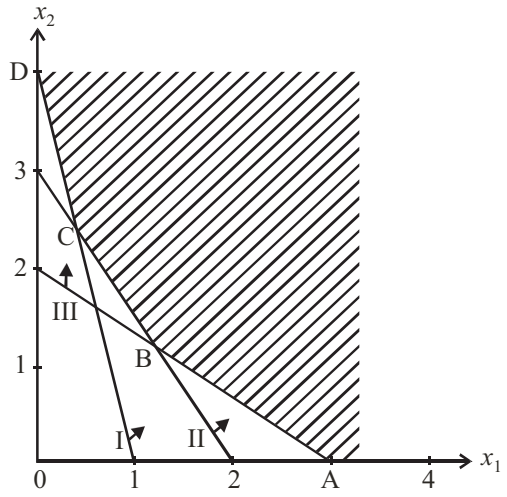


Fig. 15.7

i.e., open with the corner points A(3, 0), $B\left(\frac{3}{5}, \frac{8}{5}\right)$, $C\left(\frac{2}{5}, \frac{12}{5}\right)$ and D (0, 4). The value of the objective function at the corner points are $z_A = 9$, $z_B = \frac{89}{5} = 17.8$, $z_C = \frac{126}{5} = 25.2$, and $z_D = 40$.

Here the minimum has occurred at A and there is no other point in the feasible region at which the objective function value is lower than 9. Hence the optimal solution is

$$x_1^* = 3,\ x_2^* = 0 \text{ and } z^* = 9$$

Exceptional Cases in Graphical Method

There are three cases may arise. When the value of the objective function is maximum/ minimum at more than one corner points then '*multiple optimal*' solutions are obtained.

Sometimes the optimum solution is obtained at infinity, then the solution is called '*unbounded solution*'. Generally, this type of solution is obtained when the common feasible region is unbounded and the type of the objective function leads to unbounded solution.

When there does not exist any common feasible region, then there does not exist any solution. Then the given LPP is called *infeasible i.e.,* having *no solution*. For example, consider the LPP which is infeasible

$$\text{Maximize } z = 5x_1 + 10x_2$$
$$\text{Subject to, } x_1 + x_2 \le 2,$$
$$x_1 + x_2 \ge 3,$$
$$x_1, x_2 \ge 0.$$

Example 3: *Solve the following LPP using graphical method:*

Maximize $$z = x_1 + \frac{3}{5}x_2$$

Subject to, $$5x_1 + 3x_2 \leq 15,$$

$$3x_1 + 4x_2 \leq 12,$$

$$x_1, x_2 \geq 0.$$

Solution: Let us present all the constraints in intercept forms *i.e.,*

$$\frac{x_1}{3} + \frac{x_2}{5} \leq 1 \qquad \text{...(I)}$$

$$\frac{x_1}{4} + \frac{x_2}{3} \leq 1 \qquad \text{...(II)}$$

Due to non-negativity constraints *i.e.,* $x_1 \geq 0$, $x_2 \geq 0$ the common feasible region is obtained in the first quadrant as shown in Fig. 15.8 and the individual feasible regions are shown by arrows.

The corner points are O(0, 0), A (3, 0), $B\left(\frac{24}{11}, \frac{15}{11}\right)$ and C(0, 3). The values of the objective function at the corner points are obtained as

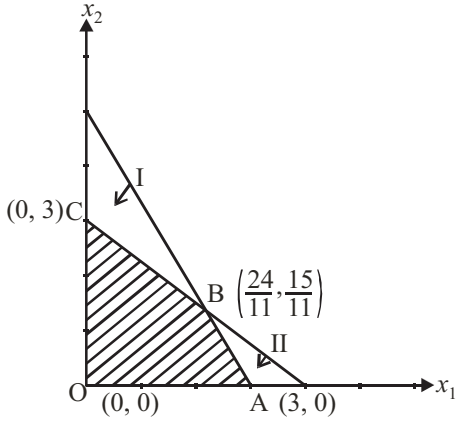


Fig. 15.8

$$z_O = 0,\ z_A = 3,\ z_B = 3,\ z_C = \frac{9}{4}.$$

Since the common feasible region is bounded and the maximum has occurred at two corner points *i.e.,* at A and B respectively, these solutions are called multiple optima. So the solutions are

$$x_1^* = 3,\ x_2^* = 0 \qquad \text{and} \qquad x_1^* = \frac{15}{11},\ x_2^* = \frac{24}{11} \text{ and } z^* = 3.$$

NOTES

15.4 SIMPLEX METHOD

Simplex method is an algebraic procedure in which a series of repetitive operations are used and we progressively approach the optimal solution. Thus, this procedure has a number of steps to find the solution to any problem, consisting of any number of variables and constraints, however problems with more than 4 variables cannot be solved manually and require the used of computer for solving them.

This method developed by the American mathematician G.B. Dantizg, can be used to solve any problem, which has a solution. The process of reaching the optimal solution through different stages is also called iterative, because the same computational steps are repeated a number of times before the optimum solution is reached.

The algorithm is discussed below with the help of a numerical example *i.e.,* consider

$$\text{Maximize } z = 4x_1 + 8x_2 + 5x_3$$
$$\text{Subject to, } x_1 + 2x_2 + 3x_3 \leq 18,$$
$$2x_1 + 6x_2 + 4x_3 \leq 15,$$
$$x_1 + 4x_2 + x_3 \leq 6,$$
$$x_1, x_2, x_3 \geq 0.$$

Step 1. If the problem is in minimization, then convert it to maximization as

Min z = – Max $(-z)$.

Step 2. All the right side constants must be positive. Multiply by – 1 both sides for negative constants. All the variables must be non-negative.

Step 3. Make standard form by adding slack variables for '≤' type constraints, surplus variables for '≥' type constraints and incorporate these variables in the objective function with zero coefficients.

For example,
$$\text{Maximum } z = 4x_1 + 8x_2 + 5x_3 + 0, s_1 + 0.s_2 + 0.s_3$$
$$\text{Subject to, } x_1 + 2x_2 + 3x_3 + s_1 = 18,$$
$$2x_1 + 6x_2 + 4x_3 + s_2 = 15,$$
$$x_1 + 4x_2 + x_3 + s_3 = 6,$$
$$x_1, x_2, x_3 \geq 0, s_1, s_2, s_3 \geq 0$$

Note that an unit matrix due to s_1, s_2 and s_3 variables is present in the coefficient matrix which is the key requirement for simplex method.

Step 4. Simplex method is an iterative method. Calculations are done in a table which is called simplex table. For each constraint there will be a row and for each variable there will be a column. Objective function coefficients c_j are kept on the top of the table. x_B stands for basis column in which the variables are called 'basic variables'. Solution column gives the solution, but in iteration 1, the right side constants are kept. At the bottom $z_j - c_j$ row is called 'net evaluation' row.

In each iteration one variable departs from the basis and is called departing variable and in that place one variable enter which is called entering variable to improve the value of the objective function.

Minimum ratio column determines the departing variable.

Iteration 1

Table 15.1

		c_j	*4*	*8*	*5*	*0*	*0*	*0*	**Min. ratio**
c_B	x_B	soln.	x_1	x_2	x_3	s_1	s_2	s_3	
0	s_1	18	1	2	3	1	0	0	
0	s_2	15	2	6	4	0	1	0	
0	s_3	6	1	4	1	0	0	1	
	$z_j - c_j$								

Note: Variables which are forming the columns of the unit matrix enter into the basis column. In this table the solution is $s_1 = 18$, $s_2 = 15$, $s_3 = 6$, $x_1 = 0$, $x_2 = 0$, $x_3 = 0$ and $z = 0$.

To test optimality we have to calculate $z_j - c_j$ for each column as follows:

$$z_j - c_j = c_B^T . [x_j] - c_j$$

For first column, $(0, 0, 0)\begin{pmatrix}1\\2\\1\end{pmatrix} - 4 = -4$

For second column, $(0, 0, 0)\begin{pmatrix}2\\6\\4\end{pmatrix} - 8 = -8$ and so on.

These are displayed in the following table:

Table 15.2

		c_j	*4*	*8*	*5*	*0*	*0*	*0*	**Min. ratio**
c_B	x_B	soln.	x_1	x_2	x_3	s_1	s_2	s_3	
0	s_1	18	1	2	3	1	0	0	
0	s_2	15	2	6	4	0	1	0	
0	s_3	6	1	4	1	0	0	1	
	$z_j - c_j$		−4	−8	−5	0	0	0	

↑

Decisions: If all $z_j - c_j \geq 0$. Then the current solution is optimal and stop. Else, Select the negative most value from $z_j - c_j$ and the variable corresponding to this value will be the entering variable and that column is called '*key column*'. Indicate this column with an upward arrow symbol.

In the given problem '– 8' is the most negative and variable x_2 is the entering variable. If there is a tie in the most negative, break it arbitrarily.

To determine the *departing variable*, we have to use minimum ratio. Each ratio is calculated as $\frac{[\text{soln.}]}{[\text{key column}]}$, componentwise division only for positive elements (*i.e.,* > 0) of the key column. In this example,

$$\text{min.}\left\{\frac{18}{2}, \frac{15}{6}, \frac{6}{4}\right\} = \text{min.}\{9, 2.5, 1.5\} = 1.5$$

The element corresponding to the min. ratio *i.e.,* here s_3 will be the departing variable and the corresponding row is called 'key row' and indicate this row by an outward arrow symbol. The intersection element of the key row and key column is called key element. In the present example, 4 is the key element which is highlighted. This is the end of this iteration, the final table is displayed as follow:

Iteration 1

Table 15.3

		c_j	4	8	5	0	0	0	Min.
c_B	x_B	soln.	x_1	x_2	x_3	s_1	s_2	s_3	ratio
0	s_1	18	1	2	3	1	0	0	$\frac{18}{2} = 9$
0	s_2	15	2	6	4	0	1	0	$\frac{15}{6} = 2.5$ →
0	s_3	6	1	4	1	0	0	1	$\frac{6}{4} = 1.5$
	$z_j - c_j$		– 4	– 8	– 5	0	0	0	

↑

Step 5. For the construction of the next iteration (new) table the following calculations are to be made:

(a) Update the x_B column and the c_B column.

(b) Divide the key row by the key element.

(c) Other elements are obtained by the following formula:

$$\begin{pmatrix}\text{new}\\ \text{element}\end{pmatrix} = \begin{pmatrix}\text{old}\\ \text{element}\end{pmatrix} - \frac{\begin{pmatrix}\text{element}\\ \text{corresponding to}\\ \text{key row}\end{pmatrix}.\begin{pmatrix}\text{element}\\ \text{corresponding to}\\ \text{key column}\end{pmatrix}}{\text{key element}}$$

(d) Then go to step 4.

Iteration 2

Table 15.4

		c_j	*4*	*8*	*5*	*0*	*0*	*0*	**Min. ratio**
c_B	x_B	**soln.**	x_1	x_2	x_3	s_1	s_2	s_3	
0	s_1	15	$\frac{1}{2}$	0	$\frac{5}{2}$	1	0	$-\frac{1}{2}$	$15\times\frac{3}{5}=6$
0	s_2	6	$\frac{1}{2}$	0	$\frac{5}{2}$	0	1	$-\frac{3}{2}$	$6\times\frac{2}{5}=2.4$ →
8	x_2	$\frac{3}{2}$	$\frac{1}{4}$	1	$\frac{1}{4}$	0	0	$\frac{1}{4}$	$\frac{3}{4}\times 4=3$
	z_j-c_j		-2	0	-3	0	0	2	

↑

Iteration 3

Table 15.5

		c_j	*4*	*8*	*5*	*0*	*0*	*0*	**Min. ratio**
c_B	x_B	**soln.**	x_1	x_2	x_3	s_1	s_2	s_3	
0	s_1	9	0	0	0	1	-1	1	-
5	x_3	$\frac{12}{5}$	$\frac{1}{5}$	0	1	0	$\frac{2}{5}$	$-\frac{3}{5}$	$\frac{12}{5}\times\frac{5}{1}=12$
8	x_2	$\frac{9}{10}$	$\frac{1}{5}$	1	0	0	$-\frac{1}{10}$	$\frac{2}{5}$	$\frac{9}{10}\times\frac{5}{1}=4.5$ →
	z_j-c_j		$-\frac{7}{5}$	0	0	0	$\frac{6}{5}$	$\frac{1}{5}$	

↑

NOTES

Iteration 4

Table 15.6

		c_j	4	8	5	0	0	0	Min. ratio
c_B	x_B	soln.	x_1	x_2	x_3	s_1	s_2	s_3	
0	s_1	9	0	0	0	1	– 1	1	
5	x_3	$\frac{3}{2}$	0	– 1	1	0	$\frac{1}{2}$	– 1	
4	x_1	$\frac{9}{2}$	1	5	0	0	$-\frac{1}{2}$	2	
	$z_j - c_j$		0	7	0	0	$\frac{1}{2}$	3	

Since all $z_j - c_j \geq 0$, the current solution is optimal.

$$\therefore \qquad x_1^* = \frac{9}{2},\ x_2^* = 0,\ x_3^* = \frac{3}{2},\ z^* = \frac{51}{2}.$$

Note (exceptional cases)

(a) If in the key column, all the elements are non-positive *i.e.,* zero or negative, then min. ratio cannot be calculated and the problem is said to be unbounded.

(b) In the net evaluation of the optimal table all the basic variables will give the value zero. If any non-basic variable give zero net evaluation then it indicates that there is an alternative optimal solution. To obtain that solution, consider the corresponding column as key column and apply one simplex iteration.

(c) For negative variables, $x \leq 0$, set $x = -x'$, $x' \geq 0$.

For unrestricted variables set $x = x' - x''$ where $x', x'' \geq 0$.

Example 4: *Solve the following by simplex method:*

$$\textit{Maximize } z = x_1 + 3x_2$$

$$\textit{Subject to, } -x_1 + 2x_2 \leq 2,\ x_1 - 2x_2 \leq 2,\ x_1, x_2 \geq 0.$$

Solution: Standard form of the given LPP can be written as follows:

$$\text{Maximum } z = x_1 + 3x_2 + 0.s_1 + 0.s_2$$

$$\text{Subject to, } -x_1 + 2x_2 + s_1 = 2,\ x_1 - 2x_2 + s_2 = 2,$$

$$x_1, x_2 \geq 0,\ s_1, s_2 \text{ slacks} \geq 0.$$

Iteration 1

		c_j	1	3	0	0	Min.
c_B	x_B	soln.	x_1	x_2	s_1	s_2	ratio
0	s_1	2	– 1	2	1	0	$\frac{2}{2}=1$
0	s_2	2	1	– 2	0	1	–
	$z_j - c_j$		– 1	– 3	0	0	

→

↑

Iteration 2

		c_j	1	3	0	0	Min.
c_B	x_B	soln.	x_1	x_2	s_1	s_2	ratio
3	x_2	1	$-\frac{1}{2}$	1	$\frac{1}{2}$	0	
0	s_2	4	0	0	1	1	
	$z_j - c_j$		$-\frac{5}{2}$	0	$\frac{3}{2}$	0	

↑

Since all the elements in the key column are non-positive, we cannot calculate min. ratio. Hence the given LPP is said to be unbounded.

15.5 DUALITY THEOREMS

Theorem 1: (Weak Duality)

Consider the symmetric primal (max. type) and Dual (min. type). The value of the objective function of the (dual) minimum problem for any feasible solution is always greater than or equal to that of the maximum problem (primal) for any feasible solution.

Proof: Let x^0 be a feasible solution to the primal.

Then $Ax^0 \leq b$, $x^0 \geq 0$ and $z = cx^0$.

Let y^0 be a feasible solution to the dual.

NOTES

Then $A^T y^0 \geq c^T$, $y^0 \geq 0$ and $T = b^T y^0$.

Taking transpose on both sides, we have

$$c \leq (y^0)^T . A$$

$$\Rightarrow \quad cx^0 \leq (y^0)^T . Ax^0$$

$$\Rightarrow \quad cx^0 \leq (y^0)^T . b$$

$$\Rightarrow \quad cx^0 \leq b^T . y^0 \qquad (\because \quad (y^0)^T b = b^T y^0)$$

Hence proved.

Example 5: *Using the C.S.C. find the optimal solution of the following primal.*

$$\text{Minimize } z = 2x_1 + 3x_2 + 5x_3 + 3x_4 + 2x_5$$

$$\text{Subject to: } x_1 + x_2 + 2x_3 + 3x_4 + x_5 \geq 4,$$

$$2x_1 - 2x_2 + 3x_3 + x_4 + x_5 \geq 3,$$

$$x_1, x_2, x_3, x_4, x_5 \geq 0.$$

Solution: The dual is

$$\text{Maximize } T = 4y_1 + 3y_2$$

$$\text{Subject to: } y_1 + 2y_2 \leq 2$$

$$y_1 - 2y_2 \leq 3$$

$$2y_1 + 3y_2 \leq 5$$

$$3y_1 + y_2 \leq 3$$

$$y_1 + y_2 \leq 2$$

$$y_1, y_2 \geq 0.$$

The solution of this dual, by graphically is $y_1^* = \frac{4}{5}$, $y_2^* = \frac{3}{5}$, $T^* = 5$. Let u_1, u_2, u_3, u_4 and u_5 be the slack variables of the dual and v_1, v_2 be the surplus variables of the primal. Then by C.S.C., we have

$$x_1 u_1 = 0, x_2 u_2 = 0, x_3 u_3 = 0,$$

$$x_4 u_4 = 0, x_5 u_5 = 0, y_1 v_1 = 0, y_2 v_2 = 0.$$

Since y_1^* and y_2^* are non-zero $\Rightarrow \quad v_1 = v_2 = 0.$

It is also seen that at optimality, the two constraints $y_1 + 2y_2 \leq 2$ and $3y_1 + y_2 \leq 3$ are satisfying in equality sense which mean $u_1^* = 0$ and $u_4^* = 0$.

For the remaining constraints, u_2^*, u_3^* and u_5^* are non-zero *i.e.,* by C.S.C., $x_2^* = 0$, $x_3^* = 0$ and $x_5^* = 0$.

Then the primal constraints reduces to

$$x_1^* + 3x_4^* = 4$$

$$2x_1^* + x_4^* = 3.$$

Solving we get $x_1^* = 1$ and $x_4^* = 1.$

Hence the optimal solution of the primal is

$$x_1^* = 1, x_2^* = 0, x_3^* = 0, x_4^* = 1, x_5^* = 0 \text{ and } z^* = 5.$$

Results on Feasibility

		Primal (Max. z)	
		Feasible Solution	**Infeasible solution**
Dual (Min. T)	Feasible solution	Max. z = Min T	Dual unbounded (Min. T $\to -\infty$)
	Infeasible solution	Primal unbounded (Max. $z \to \infty$)	May occur.

Let the primal as: Minimize $z = -x_1 - x_2$

Subject to: $x_1 - x_2 = 3,$

$x_1 - x_2 = -3,$

$x_1 \geq 0, x_2 \geq 0.$

Then the dual can be written as

Maximize $T = 3y_1 - 3y_2$

Subject to: $y_1 + y_2 \leq -1,$

$-y_1 - y_2 \leq -1,$

y_1, y_2 unrestricted in sign.

Here both the primal and the dual are inconsistent and hence no feasible solutions.

15.6 DUALITY OF SIMPLEX METHOD

The fundamental theorem of duality helps to obtain the optimal solution of the dual from optimal table of the primal and vice-versa. Using C.S.C., the correspondence between the primal (dual) variables and slack and/or surplus variables of the dual (primal) to be identified. Then the optimal solution of the dual (primal) can be read off from the net evaluation row of the primal (dual) of the simplex table.

For example, if the primal variable corresponds to a slack variable of the dual, then the net evaluation of the slack variable in the optimal table will give the optimal solution of the primal variable.

Example 6: *Using the principle of duality solve the following problem:*

$$\text{Minimize } z = 4x_1 + 14x_2 + 3x_3$$

$$\text{Subject to: } -x_1 + 3x_2 + x_3 \geq 3,$$

$$2x_1 + 2x_2 - x_3 \geq 2,$$

$$x_1, x_2, x_3 \geq 0.$$

Solution: The dual problem is

$$\text{Maximize T} = 3y_1 + 2y_2$$

$$\text{Subject to: } -y_1 + 2y_2 \leq 4$$

$$3y_1 + 2y_2 \leq 14$$

$$y_1 - y_2 \leq 3$$

$$y_1, y_2 \geq 0$$

Standard form:

$$\text{Maximize T} = 3y_1 + 2y_2 + 0.u_1 + 0.u_2 + 0.u_3$$

$$\text{Subject to: } -y_1 + 2y_2 + u_1 = 4$$

$$3y_1 + 2y_2 + u_2 = 14$$

$$y_1 - y_2 + u_3 = 3$$

$y_1, y_2 \geq 0$, u_1, u_2, u_3 are slacks and ≥ 0.

Let the surplus variables of the dual v_1 and v_2.

Then by C.S.C., $y_1v_1 = 0, y_2v_2 = 0,$

$$x_1u_1 = 0, x_2u_2 = 0, x_3u_3 = 0.$$

Let us solve the dual by simplex method and the optimal table is given below (Iteration 3):

Table 15.7

		c_j	3	2	0	0	0
c_B	x_B	Soln.	y_1	y_2	u_1	u_2	u_3
0	u_1	6	0	0	1	$-\frac{1}{5}$	$\frac{3}{5}$
2	y_2	1	0	1	0	$\frac{1}{5}$	$-\frac{3}{5}$
3	y_1	4	1	0	0	$\frac{1}{5}$	$\frac{2}{5}$
$z_j - c_j$			0	0	0	1	0

The optimal solution of the dual is $y_1^* = 4, y_2^* = 1, \mathrm{T}^* = 14.$

The optimal solution of the primal can be read off from the $(z_j - c_j)$-row. Since x_1, x_2, x_3 corresponds to u_1, u_2, u_3 respectively, then

$$x_1^* = 0, x_2^* = 1, x_3^* = 0, \text{ and } z^* = 14.$$

Check Your Progress

State whether the following statements are True or False:

6. If the optimum solution is obtained at finite points, then the solution is called Unbounded solution.
7. Infeasible LPP has no solution.
8. In simplex method, single operation is used.
9. In simplex method, problem can be made in standard form by adding surplus variables for '≤' type constraints and slack variable for '≥' type constraints.
10. The value of the objective function of the minimum problem for any feasible solution is always greater than or equal to that of the maximum problem for any feasible solution.

15.7 BOUNDED VARIABLE

Introduction

In addition to the constraints in any LP problem, the value of some or all variables is restricted with lower and upper limits. In such cases the standard form of an LP problem appears as:

Optimize (Max or Min) $Z = \mathbf{cx}$

Subject to the constraints

$$\mathbf{Ax} = \mathbf{b} \quad \text{and} \quad l \leq \mathbf{x} \leq \mathbf{u}$$

where $l = (l_1, l_2, ..., l_n)$ and $\mathbf{u} = (u_1, u_2, ..., u_n)$ denote the lower and upper constraints bounds for variable x respectively. Other symbols have their usual meaning.

The inequality constraints $l \leq \mathbf{x} \leq \mathbf{u}$ in the LP model can be converted into equality constraints by introducing slack and/or surplus variables s' and s'' as follows.

$$\mathbf{x} \geq l \quad \text{or} \quad \mathbf{x} - \mathbf{s}'' = l, \qquad \mathbf{s}'' \geq 0$$

and $$\mathbf{x} \leq \mathbf{u} \quad \text{or} \quad \mathbf{x} + \mathbf{s}' = \mathbf{u}, \qquad \mathbf{s}' \geq 0$$

Thus, the given LP model contains $m + n$ constraints equations with $3n$ variables. However, this size can be reduced to simply $\mathbf{Ax} = \mathbf{b}$.

NOTES

The lower bound constraints $l \leq \mathbf{x}$ can also be written as: $\mathbf{x} = l + \mathbf{s''}$, $\mathbf{s''} \geq 0$, and therefore, with this substitution variable **x** can be eliminated from all the constraints.

The upper bound constraints $\mathbf{x} \leq \mathbf{u}$ can also be written as: $\mathbf{x} = \mathbf{u} - \mathbf{s'}$, $\mathbf{s'} \geq 0$. Such substitution, however, does not ensure non-negative value of **x**. It is in this context that a special technique known as *bounded variable simplex method* was developed in order to overcome this difficulty.

In bounded variable simplex method, the optimality condition for a solution is the same as the simplex method, discussed earlier. But the inclusion of constraints $\mathbf{x} + \mathbf{s'} = \mathbf{u}$ in the simplex table requires modification in the feasibility condition of the simplex method due to the following reasons:

(i) A basic variable should become a non-basic variable at its upper bound (in usual simplex method all non-basic variables are at zero level).

(ii) When a non-basic variable becomes a basic variable, its value should not exceed its upper bound and should also not disturb the non-negativity and upper bound conditions of all existing basic variables.

The Simplex Algorithm

Step 1. (i) If the objective function of a given LP problem is of minimization, then convert it into that of maximization by using the following relationship:

Minimize Z = – Maximize Z*;

(ii) Check whether all b_i ($i = 1, 2, \ldots, m$) are positive. If any one is negative, then multiply the corresponding constraint by – 1 in order to make it positive.

(iii) Express the mathematical model of the given LP problem in standard form by adding slack/or surplus variables.

Step 2. Obtain an initial basic feasible solution. If any of the basic variables is at a positive lower bound, then substitute it out at its lower bound.

Step 3. Calculate $c_j - z_j$ as usual for all non-basic feasible. Examine values of $c_j - z_j$.

(i) If all $c_j - z_j \leq 0$, then the current basic feasible solution is the optimal solution.

(ii) If at least one $c_j - z_j > 0$ and this column has at least one entry positive (*i.e.* $y_{ij} > 0$) for some row *i,* then this indicates that an improvement in the value of objective function, Z is possible.

Step 4. If Case (ii) of Step 3 holds true then select a non-basic variable to enter into the new solution according to the following criterion:

$$c_k - z_k = \min_i \{c_j - z_j : c_j - z_j > 0\}$$

Step 5. After identifying the column vector (non-basic variable) that will enter the basis matrix **B,** the vector to be removed from **B** is calculated, for this calculate the quantities:

$$\theta_1 = \underset{i}{\text{Min}} \left\{ \frac{x_{Bi}}{y_{ir}},\ y_{ir} > 0 \right\}; \quad \theta_2 = \text{Min} \left\{ \frac{u_r - x_{Bi}}{-y_{ir}},\ y_{ir} < 0 \right\}$$

and $\theta = \text{Min}\ \{\theta_1, \theta_2, u_r\}$

where u_r is the upper bound for the variable x_r in the current basic feasible solution. Obviously, if all $y_{ir} > 0$, $\theta_2 = \infty$.

(i) If $\theta = \theta_1$, then the basic variable x_k (column vector $\mathbf{a}_k$) is removed from the basis and is replaced by non-basic variable, say x_r (column vector $\mathbf{a}_r$), as usual, by applying row operations.

(ii) If $\theta = \theta_2$, then the basic variable x_k (column vector $\mathbf{a}_k$) is removed and replaced with a non-basic variable x_r (column vector $\mathbf{a}_r$). But at this stage *value of basic variable* $x_r = x_{Br}$ is not at upper bound. This must be substituted out by using the relationship:

$$(x_{Bk})_r = (x_{Bk})'_r - y_{kr}\, u_r; \quad 0 \le (x_{Bk})'_r \le u_r$$

where $(x_{Bk})'_r$ denotes the value of variables x_r.

The value of non-basic variable x_r is given at its upper bound value while the remaining non-basic variables are put at zero value by using the relationship:

$$x_r = u_r - x'_r; \quad 0 \le x'_r \le u_r$$

(iii) If $\theta = u_r$, the variable x_r is given its upper bound value while the remaining non-basic variables are put at zero value by the relationship:

$$x_r = u_r - x'_r; \quad 0 \le x'_r \le u_r$$

Step 6. Go to Step 4 and repeat the procedure until all θ entries in the $c_j - z_j$, row are either negative or zero.

Example 7: *Solve the following LP problem*:

Maximize $Z = 3x_1 + 2x_2$

subject to the constraints

(*i*) $x_1 - 3x_2 \le 3$, (*ii*) $x_1 - 2x_2 \le 4$, (*iii*) $2x_1 + x_2 \le 20$

(*iv*) $x_1 + 3x_2 \le 30$, (*v*) $-x_1 + x_2 \le 6$

and $0 \le x_1 \le 8$; $0 \le x_2 \le 6$

Solution: We first add non-negative slack variables. s_i ($i = 1, 2, ..., 5$) to convert inequality constrains to equations. The standard form of LP problem then becomes:

Maximize $Z = 3x_1 + 2x_2 + 0s_1 + 0s_2 + 0s_3 + 0s_4 + 0s_5$

subject to the constraints

(i) $x_1 - 3x_2 + s_1 = 3$, (ii) $x_1 - 2x_2 + s_2 = 4$,

(iii) $2x_1 + x_2 + s_3 = 20$ (iv) $x_1 + 3x_2 + s_4 = 30$,

(v) $-x_1 + x_2 + s_5 = 6$

and $x_1, x_2, s_1, s_2, ..., s_5 \geq 0$

The initial basic feasible solution to this problem is: $x_{B1} = s_1 = 3$, $x_{B2} = s_2 = 4$, $x_{B3} - s_3 = 20$, $x_{B4} = s_4 = 30$ and $x_{B5} = s_5 = 6$. Since there are no upper bounds specified for these basic variables, arbitrarily assume that all of them have upper bound at ∞, *i.e.*, $s_1 = s_2 = s_3 = s_4 = s_5 = \infty$. This solution can also be read from the initial simplex Table 15.8.

$u_i \rightarrow$	8	6	∞	∞	∞	∞	∞
$c_j \rightarrow$	3	2	0	0	0	0	0

Table 15.8

c_B	Basic variables B	Solution values $b(=x_B)$	x_1	x_2	s_1	s_2	s_3	s_4	s_5	$u_i - x_{Bi}$
0	s_1	3	1	– 3	1	0	0	0	0	$\infty - 3 = \infty$
0	s_2	4	1	– 2	0	1	0	0	0	$\infty - 4 = \infty$
0	s_3	20	2	1	0	0	1	0	0	$\infty - 20 = \infty$
0	s_4	30	1	3	0	0	0	1	0	$\infty - 30 = \infty$
0	s_5	6	– 1	1	0	0	0	0	1	$\infty - 6 = \infty$
Z = 0		z_j	0	0	0	0	0	0	0	
		$c_j - z_j$	3	2	0	0	0	0	0	

↑

Since $c_1 - z_1 = 3$ is largest positive, variable x_1 is eligible to enter into the basis. As none of the basic variables s_1 to s_5 are at their upper bound, thus, for deciding about the variable to leave the basis, we compute:

$$\theta_1 = \underset{i}{\text{Min}} \left\{ \frac{x_{Bi}}{y_{i1}}, y_{i1} > 0 \right\} = \text{Min} \left\{ \frac{3}{1}, \frac{4}{1}, \frac{20}{2}, \frac{30}{1} \right\} = 3 \text{ (corresponds to } x_1)$$

$$\theta_2 = \underset{i}{\text{Min}} \left\{ \frac{u_r - x_{Bi}}{-y_{i1}}, y_{i1} < 0 \right\} = \frac{\infty - 6}{-(-1)} = \infty \text{ (corresponds to } \theta_1)$$

and $u_1 = 8$.

Therefore $\theta = \text{Min} \{\theta_1, \theta_2, u_1) = \text{Min} \{3, \infty, 8\} = 3$ (corresponds to θ_1)

Thus, s_1 is eligible to leave the basis and therefore $y_{11} = 1$ becomes the key element.

Introduce x_1 into the basis and remove s_1 from the basis by applying row operations in the same manner as discussed earlier. The improved solution is shown in Table 15.9.

$u_i \rightarrow$	8	6	∞	∞	∞	∞	∞
$c_j \rightarrow$	3	2	0	0	0	0	0

Table 15.9

c_B	Basic variables B	Solution values $b(= x_B)$	x_1	x_2	s_1	s_2	s_3	s_4	s_5	$u_i - x_{Bi}$
3	x_1	3	1	– 3	1	0	0	0	0	$8 - 3 = \infty$
0	s_2	1	0	(1)	– 1	1	0	0	0	$\infty - 1 = \infty \rightarrow$
0	s_3	14	0	7	– 2	0	1	0	0	$\infty - 14 = \infty$
0	s_4	27	0	6	– 1	0	0	1	0	$\infty - 27 = \infty$
0	s_5	9	0	– 2	1	0	0	0	1	$\infty - 9 = \infty$
Z = 9		z_j	0	– 9	3	0	0	0	0	
		$c_j - z_j$	0	11	– 3	0	0	0	0	

↑

Since $c_2 - z_2 = 11$ is the largest positive, variable x_2 is eligible to enter into the basis. For deciding which variable should leave the basis, we compute:

$$\theta_1 = \underset{i}{\text{Min}} \left\{ \frac{x_{Bi}}{y_{i1}}, y_{i2} > 0 \right\} = \text{Min} \left\{ \frac{1}{1}, \frac{14}{7}, \frac{27}{6} \right\} = 1 \text{ (corresponds to } x_2)$$

$$\theta_2 = \underset{i}{\text{Min}} \left\{ \frac{u_i - x_{Bi}}{-y_{i2}}, y_{i2} < 0 \right\}$$

$$= \text{Min} \left\{ \frac{8-3}{-(-3)}, \frac{\infty}{-(-2)} \right\} = \frac{5}{3} \quad \text{(corresponds to } x_1)$$

and $u_2 = 6$.

Therefore $\theta = \text{Min} \{\theta_1, \theta_2, u_2\} = \text{Min} \{1, 5/3, 6\} = 1$ (corresponds to s_2)

Thus, s_2 will leave the basis and $y_{22} = 1$ becomes the key element.

Introduce x_2, into the basis and remove s_2 from the basis as usual. The improved solution is shown in Table 15.10. Since $c_3 - z_3$ is the largest positive, therefore variable s_1 is eligible to enter into the basis. We compute:

$$\theta_1 = \underset{i}{\text{Min}} \left\{ \frac{y_{Bi}}{y_{i3}}, y_{i3} > 0 \right\} = \text{Min} \left\{ \frac{7}{5}, \frac{21}{5} \right\} = \frac{7}{5} \quad \text{(corresponds to } s_3)$$

NOTES

$$\theta_2 = \underset{i}{\text{Min}} \left\{ \frac{u_i - x_{Bi}}{-y_{i1}},\ y_{i1} < 0 \right\} = \text{Min} \left\{ \frac{8-3}{-(-3)}, \frac{\infty}{-(-2)} \right\} = \frac{5}{3}$$

(corresponds to x_1)

and $u_2 = 6$

Therefore $\theta = \text{Min}\ (\theta_1, \theta_2, u_2) = \text{Min}\ \{5/3, 6\} = 1$ (corresponds to x_2)

Thus, x_1 will leave the basis and $y_{13} = -1$ becomes the key element.

$u_i \rightarrow$	8	6	∞	∞	∞	∞	∞
$c_j \rightarrow$	3	2	0	0	0	0	0

Introduce x_1 into the basis and remove s_1 from the basis by applying row operations in the same manner as discussed earlier the improved solution is shown in Table 15.10.

Table 15.10

c_B	Basic Variables B	Solution Values $b(= x_B)$	x_1	x_2	s_1	s_2	s_3	s_4	s_5	$u_i - x_{Bi}$
3	x_1	6	1	0	(−2)	3	0	0	0	8 − 6 = 2 →
2	x_2	1	0	1	−1	1	0	0	0	6 − 1 = 5
0	s_3	7	0	0	5	−7	1	0	0	∞ − 7 = ∞
0	s_4	21	0	0	5	−6	0	1	0	∞ − 21 = ∞
0	s_5	11	0	0	−1	2	0	0	1	∞ − 11 = ∞
Z = 20		z_j	3	2	−8	11	0	0	0	
		$c_j - z_j$	0	0	8	−11	0	0	0	

↑

Introduce s_1 into the basis and remove x_1 from the basis, as usual. The improved solution is shown in Table 15.11.

$u_i \rightarrow$	8	6	∞	∞	∞	∞	∞
$c_j \rightarrow$	3	2	0	0	0	0	0

Table 15.11

c_B	Basic Variables B	Solution Values $b(= x_B)$	x_1	x_2	s_1	s_2	s_3	s_4	s_5	$u_i - x_{Bi}$
0	s_1	−3	−1/2	0	1	−3/2	0	0	0	∞ − (−3) = ∞
2	x_2	−2	−1/2	1	0	−1/2	0	0	0	6 − (−2) = ∞
0	s_3	22	5/2	0	0	(1/2)	1	0	0	∞ − 22 = ∞
0	s_4	36	5/2	0	0	3/2	0	1	0	∞ − 36 = ∞
0	s_5	8	−1/2	0	0	1/2	0	0	1	∞ − 8 = ∞
Z = −4		z_j	−1	2	0	−1	0	0	0	
		$c_j - z_j$	4	0	0	1	0	0	0	

Since $c_1 - z_1 = 4$ is the largest positive, therefore variable x_1 is eligible to enter into the basis. Also the upper bound for variable x_1 is 8; therefore we update the value of basic variables by using relationship and data of Table 15.11, as follows:

$$x_{B_1} = s_1 = x'_{B_1} - y_{11}u_1 = -3 - (-1/2)8 = 1$$

$$x_{B_2} = x_2 = x'_{B_2} - y_{21}u_1 = -2 - (-1/2)8 = 2$$

$$x_{B_3} = s_3 = x'_{B_3} - y_{31}u_1 = 22 - (5/2)8 = 2$$

$$x_{B_4} = s_4 = x'_{B_4} - y_{41}u_1 = 36 - (5/2)8 = 16$$

$$x_{B_5} = s_5 = x'_{B_5} - y_{51}u_1 = 8 - (-1/2)8 = 12$$

Also one of the non-basic variables x_1 at its upper bound can be brought at zero level by using the substitution.

$$x_1 = u_1 - x'_1 = 8 - x'_1; \quad 0 \le x'_1 \le 8$$

The data of Table 15.11 can now be update by substituting new values of basic variables as well as non-basic variables, as shown in Table 15.12. Since $c_4 - z_4$ is the only positive value, s_2 will enter into the basis. For deciding which variable should leave the basis, we compute:

$$\theta_1 = \min_i \left\{ \frac{x_{Bi}}{y_{i4}}, y_{i4} > 0 \right\} = \text{Min} \left\{ \frac{2}{1/2}, \frac{16}{3/2}, \frac{12}{12} \right\} = \text{Min } \{4, 32/3, 4\} = 4$$

(corresponds to x_3)

$$\theta_2 = \min_i \left\{ \frac{u_i - x_{Bi}}{-y_{i4}}, y_{i4} < 0 \right\} = \text{Min} \left\{ \frac{\infty}{-(-3/2)}, \frac{6-2}{-(-1/2)} \right\} = 8$$

(corresponds to x_2)

and $u_4 = \infty$.

$u_i \rightarrow$	8	6	∞	∞	∞	∞	∞	
$c_j \rightarrow$	–3	2	0	0	0	0	0	

Table 15.12

c_B	Basic variables B	Solution values $b(= x_B)$	x'_1	x_2	s_1	s_2	s_3	s_4	s_5	$u_i - x_{Bi}$
0	s_1	1	1/2	0	1	– 3/2	0	0	0	$\infty - 1 = \infty$
2	x_2	2	(1/2)	1	0	– 1/2	0	0	0	$6 - 2 = 4$
0	s_3	2	– 5/2	0	0	1/2	1	0	0	$\infty - 2 = \infty \rightarrow$
0	s_4	16	– 5/2	0	0	3/2	0	1	0	$\infty - 16 = \infty$
0	s_5	12	1/2	0	0	1/2	0	0	1	$\infty - 12 = \infty$
Z = 4 + 24 = 28		z_j	1	2	0	– 1	0	0	0	
		$c_j - z_j$	– 4	0	0	1	0	0	0	

↑

Therefore, $\theta = \text{Min}\ \{\theta_1, \theta_2, u_4) = \text{Min}\ \{4, 8, \infty\} = 4$ (corresponds to s_3).

Thus, variable s_3 will leave the basis and $y_{34} = 1/2$ becomes the key element.

Introduce s_2 into the basis and remove s_3 from the basis as usual. The improved solution is shown in Table 15.13.

$u_i \rightarrow$	8	6	∞	∞	∞	∞	∞
$c_j \rightarrow$	– 3	2	0	0	0	0	0

Table 15.13

c_B	Basic variables B	Solution values $b(= x_B)$	x_1'	x_2	s_1	s_2	s_3	s_4	s_5	$u_i - x_{Bi}$
0	s_1	7	– 7	0	1	0	3	0	0	$\infty - 1 = \infty$
2	x_2	4	– 2	1	0	0	1	0	0	$6 - 2 = 4 \rightarrow$
0	s_2	4	– 5	0	0	1	2	0	0	$\infty - 2 = \infty$
0	s_4	12	5	0	0	0	– 3	1	0	$\infty - 16 = \infty$
0	s_5	10	3	0	0	0	– 1	0	1	$\infty - 12 = \infty$
Z = 8 + 24 = 32		z_j	– 4	2	0	0	2	0	0	
		$c_j - z_j$	1	0	0	0	– 2	0	0	

↑

Since $c_1 - z_1$ is the only positive value, variable x_1' will enter the basis. To decide which variable will leave the basis, we compute:

$$\theta_1 = \min_i \left\{ \frac{x_{Bi}}{y_{i1}},\ y_{i1} > 0 \right\} = \text{Min}\ \left\{ \frac{12}{5}, \frac{10}{3} \right\} = \frac{12}{5} \qquad \text{(corresponds to } s_4\text{)}$$

$$\theta_2 = \min_i \left\{ \frac{u_i - x_{Bi}}{-y_{i1}},\ y_{i1} < 0 \right\} = \text{Min}\ \left\{ \frac{\infty}{-(-7)}, \frac{6-4}{-(-2)}, \frac{\infty}{-(-5)} \right\} = 1$$

(corresponds to x_2)

and $u_1 = 8$.

Therefore $\theta = \text{Min}\ \{\theta_1, \theta_2, u_1\} = \text{Min}\ \{12/5, 1, 8\} = 1$ (corresponds to x_2)

Thus, variable x_2 will leave the basis and $y_{21} = -2$ becomes the key element.

Introduce x_1' into the basis and remove x_2 from the basis. The new solution is shown in Table 15.14.

$u_i \rightarrow$	8	6	∞	∞	∞	∞	∞
$c_j \rightarrow$	– 3	2	0	0	0	0	0

Table 15.14

c_B	Basic variables B	Solution values $b\ (= x_B)$	x_1'	x_2	s_1	s_2	s_3	s_4	s_5	$u_i - x_{Bi}$
0	s_1	– 7	0	– 7/2	1	0	– 1/2	0	0	$\infty + 7 = \infty$
– 3	x_1'	– 2	1	– 1/2	0	0	– 1/2	0	0	– 3 + 2 = – 1
0	s_2	– 6	0	– 5/2	0	1	– 1/2	0	0	$\infty + 6 = \infty$
0	s_4	22	0	5/2	0	0	– 1/2	1	0	$\infty - 22 = \infty$
0	s_5	16	0	3/2	0	0	1/2	0	1	$\infty - 16 = \infty$
Z = 24 + 6		z_j	– 3	3/2	0	0	3/2	0	0	
= 30		$c_j - z_j$	0	1/2	0	0	– 3/2	0	0	

Since the upper bound for variable x_2 is 6, we update the value of basic variables by using the following relationships and data of Table 15.14:

$$x_{B_1} = s_1 = x'_{B_1} - y_{12}u_2 = -7 - (-7/2)6 = 14$$

$$x_{B_2} = x'_1 = x'_{B_2} - y_{22}u_2 = -2 - (-1/2)6 = 1$$

$$x_{B_3} = s_2 = x'_{B_3} - y_{32}u_2 = -6 - (-5/2)6 = 9$$

$$x_{B_4} = s_4 = x'_{B_4} - y_{42}u_2 = 22 - (5/2)6 = 7$$

$$x_{B_5} = s_5 = x'_{B_5} - y_{52}u_2 = 16 - (3/2)6 = 7$$

The non-basic variable x_2 at its upper bound can be brought at zero level by using the substitution:

$$x_2 = u_2 - x'_2 = 6 - x'_2 = 6 - x'_2,\ \ 0 \le x'_2 \le 6$$

The data of Table 15.14 can now be updated by substituting new values of basic variables and non-basic variables as shown in Table 15.15.

$u_i \rightarrow$		8	6	∞	∞	∞	∞	∞	
$c_j \rightarrow$		– 3	– 2	0	0	0	0	0	
c_B	Basic variables B	Solution values $b\ (= x_B)$	x_1'	x_2'	s_1	s_2	s_3	s_4	s_5
0	s_1	14	0	7/2	1	0	– 1/2	0	0
– 3	x_1'	1	1	1/2	0	0	– 1/2	0	0
0	s_2	9	0	5/2	0	1	– 1/2	0	0
0	s_4	7	0	– 5/2	0	0	– 1/2	1	0
0	s_5	7	0	– 3/2	0	0	1/2	0	1
Z = 24 – 3		z_j	– 3	– 3/2	0	0	3/2	0	0
= 21		$c_j - z_j$	0	– 1/2	0	0	– 3/2	0	0

In Table 15.15 all $c_j - z_j \leq 0$, an optimal solution is arrived at with values of variables as: $x_1' = 1$ or $x_1 = u_1 - x_1' = 8 - 1$; $x_2 = u_2 - x_2' = 6 - 0 = 6$ and Max Z = 33.

NOTES

15.8 FORMULATION OF LPP

The formulation of linear programming requires the following steps:

(a) Identifying/defining the decision variables.

(b) Specifying/defining the objective function to be maximized or minimized.

(c) Identifying the constraint equations, which have to be expressed as equalities or inequalities.

(d) Using the equation either in graphical or simplex method to find out the value of decision variables to optimize the objective function.

Assumptions for Solving a Linear Programming Problem

The application of LP makes use of the following assumptions:

(a) *Linearity.* The objective function and each constraint is linear.

(b) *Certain and Constant.* It means that the number of resources available and production requirements are known exactly and remain constant.

(c) *Non-negative Variables.* The values of decision variables are non-negative and represent real life solutions. Negative values of physical goods or products are impossible. Production of minus 10 refrigerators is meaningless.

Example 8: *A manufacturing company is producing two products A and B. Each of the products A and B requires the use of two machines P and Q. A requires 4 hours of processing on machine P and 3 hours of processing on machine Q. Product B requires 3 hours of processing on machine P and 6 hours of processing on machine Q. The unit profits for products A and B are ₹ 20 and ₹ 30 respectively. The available time in a given quarter on machine P is 1000 hours and on machine Q is 1200 hours. The market survey has predicted that 250 units of products A and 300 units of product B can be consumed in a quarter. The company is interested in deciding the product mix to maximize the profits. Formulate this problem as LP model.*

Solution: Formulating the problem in mathematical equations

Let X_A = the quantity of product of type A manufactured in a quarter.

X_B = the quantity of products of type B manufactured in a quarter.

Z = the profit earned in a quarter.

(Objective function,, which is to be maximized).

Therefore, $Z = 20\ X_A + 30\ X_B$

Z is to be maximized under the following conditions:

$$4X_A + 3X_B \leq 1000 \text{ (Time constraint of machine P)}$$

$$3X_A + 6X_B \leq 1200 \text{ (Time constraint of machine Q)}$$

$$X_A \leq 250 \text{ (Selling constraint of product A)}$$

$$X_B \leq 300 \text{ (Selling constraint of product B)}$$

$$X_A \text{ and } X_B \geq 0 \text{ (Condition of non-negativity).}$$

Example 9: *M/s Steadfast Ltd. produces both the interior and exterior house paints for wholesale distribution Two types of raw materials A and B are used to manufacture the paints. The maximum availability of A is 10 tons a day and that of B is 15 tons a day Daily requirement of raw material per ton of inierior and exterior paints are as follows:*

Type of raw material	Requirement of Raw material per ton of paint		
	Interior	**Exterior**	**Max availability (tons)**
Raw material A	3	2	10
Raw material B	2	3	15

The market survey indicates that daily demand of interior paints cannot exceed that of exterior paints by 2 tons. The survey also shows that maximum demand of interior paints is only 3 tons daily.

The wholesale price per ton is ₹ 75000 for exterior paints and ₹ 50000 for interior paints.

Problem: *How much interior and exterior paints M/s Steadfast should produce to maximize its profits*?

Solution: Let X_E – tons of external paints to be produced daily.

X_I – tons of internal paints to be produced daily.

Z – Profit earned (objective function which is to be maximized).

Therefore,

$$Z = 75000\ X_E + 50000\ X_I$$

Z is to be maximized under the following constraints or conditions:

$$2X_E + 3X_I \leq 10 \text{ (Availability constraint of raw material A)}$$

$$3X_E + 2X_I \leq 15 \text{ (Availability constraint of raw material B)}$$

$X_I - X_E \leq 2$ (Demand constraint – Demand of interior paints daily cannot exceed more than 2 tons that of exterior paint)

$X_I \leq 3$ (Demand of interior paint cannot exceed 3 tons everyday)

Also,

$X_I \geq 0$ (Non-negativity constraint of interior paints)

$X_E \geq 0$ (Non-negativity constraint of exterior paints).

The complete mathematical model for M/s Steadfast Ltd. problem may be written as given below:

Determine the tons of interior and exterior paints, X_I and X_E to be produced in order to maximize $Z = 75000\ X_E + 50000\ X_I$ (objective function) under the constraints (conditions) of

$$2X_E + 3X_I \leq 10$$
$$3X_E + 2X_I \leq 15$$
$$X_I - X_E \leq 2$$
$$X_I \leq 3$$
$$X_E \geq 0$$
$$X_I \geq 0$$

Let us verify if the above model satisfies the conditions of linearity as we are using LP method to solve the problem. Linearity demands that **Proportionality** and **Additivity** must be satisfied.

1. ***Proportionality:*** It requires that usage of resources is directly proportional to the value of the variables, in this case X_E and X_I. Suppose M/s Steadfast have a promotion policy that they will sell external paint at ₹ 60000 per ton after sales of exterior paint is more than 2 tons per day, then the equation of the objective function will no longer be true as each ton of external paint produced does not bring a revenue of ₹ 75000/- Actually for $X_E \leq 2$ tons, the revenue will be ₹ 75000 per ton and ₹ 60000/- per ton for $X_E \geq 2$ tons. This situation does not satisfy the condition of direct proportionality with X_E.
2. ***Additivity:*** It demands that the objective function (Z) must be the direct sum of the individual contribution of different variables. For example, in case of two competing products like tea and coffee where an increase in sale of coffee adversely affects the sale of tea, these two products do not satisfy the conditions of additivity.

   Since the mathematical model written above, satisfies both the properties of proportionality and additivity, we can use LP method to solve the problem.

Example 10: *A chemical manufacturer produces two types of chemicals X and Y. Each type of chemical is manufactured by a two-step process that involves machines A and B. The processing time for each unit of the two products on machines A and B are as follows:*

Product	*Machine A (Hours/unit)*	*Machine B (Hours/unit)*
X	2	3
Y	4	2

For a period of one month, the availability of machine A is 160 hours and availability of machine B is 120 hours. The manufacturer has found out from the market that maximum sale price of chemical X can be ₹ 400 per unit and that of chemical Y can be ₹ 500 per unit. Also, maximum of 20 units of chemical X can be sold in the market per month and 25 units of chemical B.

Problem: *How much of chemical X and chemical Y should be produced so that the profits can be maximized*?

Solution: Let X–be the quantity of chemical X to be produced.

Y–be the quantity of chemical Y to be produced.

$$Z = 400X + 500Y$$

Z is to be maximized under the following constraints or conditions:

$2X + 4Y \leq 160$ (Availability constraint of machine A)

$3X + 2Y \leq 120$ (Availability constraint of machine B)

$X \leq 20$ (Marketing constraint of product X)

$Y \leq 25$ (Marketing constraint of product B)

$X \geq 0$ (Non-negativity constraint of product X)

$Y \geq 0$ (Non-negativity constraint of product Y)

The complete mathematical model for the problem may be written as follows:

Determine the quantity of product X and Y to be produced in order to maximize $Z = 400X + 500Y$ (objective function) under the constraint of

$2X + 4Y \leq 160$

$3X + 2Y \leq 120$

$X \leq 20$

$Y \leq 25$

$X \geq 0$

$Y \geq 0$

15.9 PARAMETRIC PROGRAMMING

Introduction

Once an LP model based on real-life data has been solved, the decision-maker desires to know how the solution will change if parameters, such as cost (or profit) c_j, availability of resources b_i and the technological coefficients a_{ij} are changed. We

have already discussed the need to perform a sensitivity analysis in order to consider the impact of *discrete changes* in its parameters on optimal solution of LP model. In this chapter, we will discuss another parameter variation analysis also called *parametric analysis* to find various basic feasible solutions of an LP model that become optimal one after the other, due to continuous variations in the parameters. Since LP model parameters change as a linear function of a single parameter, this technique is known as *linear parametric programming*.

The purpose of this analysis is to keep to a minimum the additional efforts required to take care of changes in the optimal solution due to variation in LP model parameters *over a range of variation.* In this chapter we will perform parametric analysis only for the following two parameters (evaluation of other parameters, over a range, is also possible but tend to be more complex).

(i) Variation in objective function coefficients, c_j

(ii) Variation in resources availability (Right-hand side values), b_i

Let λ be the unknown (positive or negative) scalar parameter with which coefficients in the LP model vary. We start the analysis at optimal solution obtained at $\lambda = 0$. Then, using the optimally and feasibility conditions of the simplex method we determine the range of λ for which the optimal solution at $\lambda = 0$ remains unchanged. Let λ lies between 0 and λ_1. This means $0 \le \lambda \le \lambda_1$ is the range of λ beyond which the current solution will become infeasible and/or non-optimal. Thus at $\lambda = \lambda_1$ a new solution is determined which remains optimal and feasible in other interval, say $\lambda_1 \le \lambda \le \lambda_2$. Again a new solution at $\lambda = \lambda_2$ is obtained. The process of determining the range of λ is repeated till a stage is reached beyond which the solution either does not change or exist.

Variation in the Objective Function Coefficients

Let us define the parametric linear programming model as follows:

Maximize $$Z = \sum_{j=1}^{n} (c_j + \lambda c'_j)\, x_j$$

subject to the constraints

$$\sum_{j=1}^{n} a_{ij}\, x_j = b_i; \quad i = 1, 2, ..., m \quad \text{and} \quad x_j \ge 0; \quad j = 1, 2, ..., n$$

where $\lambda c'_j$ represents predetermined variation in the parameter c and $\lambda \ge 0$ is a scalar parameter. Now the aim is to determine such consecutive values of λ at which the current optimal basic feasible solution tends to change with a change in the coefficients c_j. Such consecutive values of λ are called *critical* (*range*) *values of* λ and are measured from $\lambda = 0$. Thus, the given LP problem is initially solved by using simplex method at $\lambda = 0$. Since changes in cost coefficient c_j only affect the optimality of the current solution, therefore as λ changes only $c_j - z_j$ values are affected. Hence, for

the perturbed LP problem let us calculate $c_j - z_j$ values corresponding to non-basic variable columns in the optimal simplex table as follows:

$$c_j(\lambda) - z_j(\lambda) = c_j(\lambda) - \mathbf{c_B}(\lambda)\,\mathbf{B^{-1}}\,\mathbf{a_j} = (c_j + \lambda c_j') - (\mathbf{c_B} + \lambda \mathbf{c_B'})\mathbf{y_j}; \quad \mathbf{y_j} = \mathbf{B^{-1}}\,\mathbf{a_j}$$
$$= (c_j - \mathbf{c_B}\,\mathbf{y_j}) + \lambda(c_j' - \mathbf{c_B'}\mathbf{y_1}) = (c_j - z_j) + \lambda\,(c_j' - z_j'); \quad z_j = \mathbf{c_B}\,\mathbf{y_j}$$

For a solution to be optimal for all values of λ we must have $c_j(\lambda) - z_j(\lambda) \le 0$ (maximization case) and $c_j(\lambda) - z_j(\lambda) \ge 0$ (minimization case). These inequalities, for a given solution, are also used for determining the range $\lambda_1 \le \lambda \le \lambda_2$, within which the current solution remains optimal as follows:

$$\lambda = \text{Min}\left\{\frac{-(c_j - z_j)}{(c_j' - z_j')}\right\}$$

where $c_j' - z_j' > 0$ for maximization and $c_j' - z_j' < 0$ for minimization.

Example 11: *Consider the parametric linear programming problem:*
Maximize $Z = (3 - 6\lambda)\,x_1 + (2 - 2\lambda)\,x_2 + (5 + 5\lambda)\,x_3$
subject to the constraints

(*i*) $x_1 + 2x_2 + x_3 \le 430$, (*ii*) $3x_1 + 2x_3 \le 460$, (*iii*) $3x_1 + 4x_2 \le 420$,
and $x_1, x_2, x_3 \ge 0$

Perform the parametric analysis and identify all the critical values of the parameter λ.

Solution: The given parametric LP problem can be written in its standard form as:
Maximize $Z = (3 - 6\lambda)\,x_1 + (2 - 2\lambda)\,x_2 + (5 + 5\lambda)\,x_3 + 0s_1 + 0s_2 + 0s_3$
subject to the constraints

(i) $x_1 + 2x_2 + x_3 + s_1 = 430$ (ii) $3x_1 + 2x_3 + s_2 = 460$
(iii) $x_1 + 4x_2 + s_3 = 420$ and $x_1, x_2, x_3, s_1, s_2, s_3 \ge 0$.

According to the problem, we have:

$$c(\lambda) = c_j + \lambda c_j' = (3, 2, 5, 0, 0, 0) + \lambda(-6, -2, 5, 0, 0, 0)$$

Solving the given LP problem with $\lambda = 0$. The optimal solution at $\lambda = 0$ is shown in Table 15.16.

Table 15.16: Optimal Solution at $\lambda = 0$

	$c_j \rightarrow$		3	2	5	0	0	0
c_B	Basic variables B	Solution values $b\ (= x_B)$	x_1	x_2	x_3	s_1	s_2	s_3
2	x_2	100	– 1/4	1	0	1/2	– 1/4	0
5	x_3	230	3/2	0	1	0	1/2	0
0	s_3	20	2	0	0	– 2	1	1
Z = 1,350		z_j	7	2	5	1	2	0
		$c_j - z_j$	– 4	0	0	– 1	– 2	0

The optimal solution is: $x_1 = 0$, $x_2 = 100$, $x_3 = 230$ and Max Z = 1,350.

NOTES

In order to find the first critical (or range) value of k in which the solution shown in Table 15.16 remains optimal, we first find $c_j' - z_j'$ values corresponding to non-basic variables x_1, s_1 and s_2 columns as follows:

$$c_j' - z_j' = c_j' - \mathbf{c_B'}\, y_j = (-6, 0, 0) - (-2, 5, 0)\begin{bmatrix} -1/4 & 1/2 & -1/4 \\ 3/2 & 0 & 1/2 \\ 2 & -2 & 1 \end{bmatrix}; \quad j = 1, 4, 5$$

$$= (-6, 0, 0) - \left[\frac{1}{2} + \frac{15}{2}, -1, \frac{1}{2} + \frac{5}{2}\right]$$

$$= (-6, 0, 0) - (8, -1, 3) = (-14, 1, -3)$$

For a maximization LP problem, the current solution will remain optimal provided all $c_j(\lambda) - z_j(\lambda) \leq 0$. Since $c_4' - z_4' > 0$, the first critical value of k is given by:

$$\lambda_1 = \text{Min}\left\{\frac{-(c_j - z_j)}{(c_j' - z_j') > 0}\right\} = \frac{(c_4 - z_4)}{c_4' - z_4'} = -\frac{(-1)}{1} = 1$$

This means that for $\lambda_1 \in [0, 1]$, the solution given in Table 15.16 remains optimal. The objective function value in this interval is given by:

$$Z(\lambda) = Z + Z'(\lambda) = \mathbf{c_B x_B} + \lambda\, \mathbf{c_B'\, x_B} = 1{,}350 + 950\,\lambda$$

Now, for values of λ other than zero in the interval [0, 1], we compute $c_j(\lambda) - z_j(\lambda)$ values for none basic variables x_1, s_1 and s_2 as shown in Table 29.2.

$$c_1(\lambda) - z_1(\lambda) = (c_1 - z_1) + \lambda(c_1' - z_1') = -4 - 14\lambda \leq 0 \text{ or } \lambda \geq 2/7$$

$$c_4(\lambda) - z_4(\lambda) = (c_4 - z_4) + \lambda(c_4' - z_4') = -1 + \lambda \leq 0 \text{ or } \lambda \geq 1$$

$$c_5(\lambda) - z_5(\lambda) = (c_5 - z_5) + \lambda(c_5' - z_5') = -2 - 3\lambda \qquad \leq 0 \text{ or } \lambda \geq -2/3$$

The optimal solution for any value of λ in the interval [0, 1] is given in Table 15.17.

$c_j' \rightarrow$				-6	-2	5	0	0	0
$c_j \rightarrow$				3	2	5	0	0	0

Table 15.17

c_B'	c_B	Basic variables B	Solution values $b\ (= x_B)$	x_1	x_2	x_3	s_1	s_2	s_3
2	2	x_2	100	$-1/4$	1	0	1/2	$-1/4$	0
5	5	x_3	230	3/2	0	1	0	1/2	0
0	0	s_3	20	2	0	0	-2	1	1
$Z(\lambda) = 1{,}350 + 950\lambda$		$c_j - z_j$	-4	0	0	-1	-2	0	
			$c_j' - z_j'$	-14	0	0	1	-3	0
			$c_j(\lambda) - z_j(\lambda)$	$-4 - 14\lambda$	0	$0 - 1 + \lambda$		$-2 - 3\lambda$	0

At $\lambda = 1$, $c_4(\lambda) - z_4(\lambda) = 0$ in the 's_1' column. But for $\lambda > 1$, $c_4(\lambda) - z_4(\lambda) > 0$ for non-basic variable and hence the solution in Table 15.18 no longer remains optimal. We now enter variable s_1 in the solution to find new optimal solution. The new optimal solution shown in Table 15.18 is: $x_1 = 0$, $x_2 = 0$, $x_3 = 230$ and Max Z = 2,300.

$c_j' \rightarrow$	– 6	– 2	5	0	0	0
$c_j \rightarrow$	3	2	5	0	0	0

Table 15.18

c_B'	c_B	Basic variables B	Solution values $b\ (= x_B)$	x_1	x_2	x_3	s_1	s_2	s_3
0	0	s_1	200	– 1/2	2	0	1	– 1/2	0
5	5	x_3	230	3/2	0	1	0	1/2	0
0	0	s_3	420	1	4	0	0	0	1
$Z(\lambda) = 2{,}300$			$c_j - z_j$	– 9/2	2	0	0	– 5/2	0
			$c_j' - z_j'$	– 27/2	– 2	0	0	– 5/2	0
			$c_j(\lambda) - z_j(\lambda)$	– 18	0	0	0	– 5	0

The solution shown in Table 15.18 will be optimal if all $c_j(\lambda) - z_j(\lambda) \leq 0$, $j = 1, 2, 5$. The check the optimality we compute these values for the non-basic variables x_1, x_2 and s_2 as follows:

$$c_1(\lambda) - z_1(\lambda) = (c_1 - z_1) + \lambda(c_1' - z_1') = -\frac{9}{2} - \frac{27}{2}\lambda \leq 0 \text{ or } \lambda \geq -\frac{1}{3}$$

$$c_2(\lambda) - z_2(\lambda) = (c_2 - z_2) + \lambda(c_2' - z_2') = 2 - 2\lambda \leq 0 \text{ or } \lambda \geq 1, \text{ where is true}$$

$$c_5(\lambda) - z_5(\lambda) = (c_5 - z_5) + \lambda(c_5' - z_5') = -\frac{5}{2} - \frac{5}{2}\lambda \leq 0 \text{ or } \lambda \geq -1$$

Therefore, for $\lambda = 1$, the $c_j(\lambda) - z_j(\lambda) \leq 0$ for all non-basic variable columns and hence the solution in Table 15.18 is optimal: $x_1 = x_2 = 0$, $x_3 = 230$ and Max Z = 2,300.

For $\lambda \leq -2/3$, $c_j(\lambda) - z_j(\lambda)$ value for non-basic variable s_2 becomes positive and again solution shown in Table 15.18 no longer remains optimal. Entering variable s_2 in the basis to find new optimal solution. The variable s_2 will replace basic variable s_3 in the basis. The new optimal solution is shown in Table 15.19.

$c_j' \rightarrow$	– 6	– 2	5	0	0	0
$c_j \rightarrow$	6	2	5	0	0	0

Table 15.19

c_B'	c_B	Basic variables B	Solution values $b\ (= x_B)$	x_1	x_2	x_3	s_1	s_2	s_3
– 2	2	x_2	105	1/4	1	0	0	0	1/4
5	5	x_3	220	1/2	0	1	1	0	– 1/2
0	0	s_2	20	2	0	0	– 2	1	1
Z = 1,310 + 890 λ			$c_j - z_j$	0	0	0	– 5	0	2
			$c_j' - z_j'$	– 8	0	0	– 5	0	3

15.10 CONCEPT OF INTEGER PROGRAMMING

Introduction

In mathematical programming problems, sometimes the values of negative and fraction. In such cases the solution is not optimal. In Linear Programming it is assumed that the decision variables can take continuous values, *i.e.*, these could be fractions or in integer. Integer Programming deals with solutions in which some or all the variables can assume integers non-negative values only. In LPP, the result may recommend the use of 4.5 machines or employing 6.5 men, which has no meaning as fractional machines and men cannot be used. Hence, there is a need to have a programming system where the results are always only integers and not fractions. This need is met by the Integer Programming techniques. We could have

(a) ***Pure Integer Linear Programming:*** If all the variables take only integer values.

(b) ***Mixed Integer Linear Programming:*** If some of the variables are restricted to have only integer values while others could have fractional values as the case may be in real life applications of the problem.

Limitations of Integer Linear Programming

We have seen that in LP problems with large and complex data can be solved in a reasonable time, however the performance of integers algorithms has not been found to be uniformly efficient and useful. In integer programming rounding off in used to get a value approximately true or correct. Rounding off is done in such a manner that the closest possible or nearest number is taken. If the result is 22.3 men, obviously the approximation is 22 men and not 23 men. This intersects rounding off error. This type of error or approximation may be acceptable when we talk to discrete number

of variables like men, machines, etc., however when we are using this algorithm for solution of financial investments, it is not rational or logical to make use of integers only. Here, in fact, it is required that exact values of money are worked out for the best possible results.

Methods of Integer Programming

The following two integer programming methods are available.

Cutting Plane Method: In this method of Integer Linear Programming, certain 'secondary' conditions are added in such a manner that the ultimate result satisfies the conditions of only integer solutions. These 'secondary' conditions 'cut' or eliminate certain aspects of the solution which are not feasible integers. Thus, the name 'cutting methods'.

Search Methods: Here all the possible feasible integers only are considered as the solution. The best known search method is called the *Branch-and-Bound techniques*. A special case of each method is when all the integer variables are binary in nature.

Cutting-Plane Algorithm

This method was developed by RE Gomory for pure-integer problems as also for mixed integer problems. Fractional algorithm and mixed algorithms are applied to the two problems respectively. The following steps are involved in finding the solution.

Step I. Minimization problem is converted into maximization problem.

Step II. Solve this maximization problem without considering the condition of integer values.

Step III. If the optimal solution found in step II for the variables, does not have integer values, then moves to step IV as given below.

Step IV. Carry-out the test of integrality of the solution.

Determine the highest fraction value in solution value column of the solution. Select the row with the largest value. If there is a negative fraction, convert this into the sum of negative integer and a non-negative fraction. Then the row, which contained the largest fraction, is written in the form of an equation. Now, we obtain equations with fractional parts of all coefficients by ignoring integral parts and replacing the whole number by zeros.

Step V. Here the technical coefficient = fractional part of a resource availability + some integer. Hence it is equal to or greater than the fractional part of resource availability. So, fractional part is taken to the R.H.S. and the inequation is formed as greater than or equal to (> =) type. If this is to be converted into < = type, it is multiplied with – 1 and to make it as an inequality a slack is introduced.

Step VI. The new constraint is added to the optimum simplex table of the solution found in step II. Now, solve the problem by Dual Simplex Method.

Step VII. If the solution has all integer values, then this is the optimal solution. However, if there are some fractional values, go back to step III. The procedure is repeated till an optimum solution wilh all the integer values is obtained. The above method will be explained with the help of examples.

Integer Programming Formulation

Use the same mathematical notations as were used in the formulation of LPP, the integer programming can be mathematically written as

Maximize or optimize $Z = \sum_{j=1}^{n} C_j X_j$

Subject to the constraints

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_{ij} = 1, 2, 3, ..., m$$

$$x_j \geq 0 \; j = 1,2,3, ..., n$$

and x_j integer value $j = 1,2, ..., s$

The most common use of integer programming is found in the real world problems related to investment decisions, budgeting, Production Planning and Control (PPC) in manufacturing industry, travelling salesmen, etc. Some of these cases are discussed in succeeding examples.

Example 12: *An investment consultant has four projects with different investments and present value of expected returns. Funds available for investment during the three proposals are also available. The detailed information regarding the project is as follows:*

Project	***Investment during year***			***PV of expected return***
	1	***2***	***3***	
P – 1	*1000000*	*600000*	*500000*	*800000*
P – 2	*500000*	*200000*	*400000*	*700000*
P – 3	*300000*	*250000*	*350000*	*400000*
P – 4	*400000*	*300000*	*260000*	*300000*
Funds for investment	*1800000*	*1000000*	*800000*	

Formulate an integer programming model for the consultant to make a decision as to which project should be accepted in order to maximize present value of expected return.

Solution: *Let X_1, X_2, X_3 and X_4 be the investment on projects P – 1, P – 2,* P – 3 and P – 4 respectively.

Maximize Z = 800000 + 70000000 × 2 + 400000 × 3 + 300000 × 4

subject to the constraints

$1000000 \times 1 + 50000000 \times 2 + 300000 \times 3 + 400000 \times 4 \leq 1800000$

$600000 \times 1 + 20000000 \times 2 + 250000 \times 3 + 300000 \times 4 \leq 1000000$

$500000 \times 1 + 40000000 \times 2 + 350000 \times 3 + 260000 \times 4 \leq 800000$

where $X_1, X_2, X_3, X_4 \geq 0$ and are integers.

Example 13: *A Multinational Company (MNC) is planning to invest in four different projects in Business Process Outsourcing (BPO) industry in an important town of North. The details of the investment of MNC (in thousands of rupees) are provided below.*

Project	***Present value of expected returns***	***Capital requirement for three years***		
		1	*2*	*3*
A	*800*	*600*	*500*	*550*
B	*550*	*900*	*400*	—
C	*400*	*300*	*200*	*400*
D	*250*	*400*	*150*	*100*
Funds available for investment	*1500*	*1200*	*700*	*500*

It is also known that projects A and B are mutually exclusive. However, project D can only be accepted if project C is acceptable due to technology constraints. Which project should the MNC accept to maximize their present value of expected returns?

Solution: Let X_1, X_2, X_3 and X_4 be the investment in projects A, B, C and D respectively. Also, let $X_j = 1$ (if project *j* is accepted) and $X_j = 0$ (if project is rejected)

Maximize (PV of returns) $Z = 800 \times 1 + 550 \times 2 + 400 \times 3 + 250 \times 4 \leq 1500$

Subject to the constraints

$$600 \times 1 + 900 \times 2 + 300 \times 3 + 400 \times 4 \leq 1200$$

$$500 \times 1 + 400 \times 2 + 200 \times 3 + 150 \times 4 \leq 700$$

$$550 \times 1 + 400 \times 3 + 100 \times 4 \leq 500$$

$$X_1 + X_2 \geq 1$$

$$-X_3 + X_4 \leq 1$$

$$X_j = 0 \text{ or } 1$$

Branch and Bound Method

In certain type of problems of the variables of an Integer Programming Problem (IPP) have the constraint of a upper limit or a lower limit or both upper and lower bounds. The method used to solve such problem is called *Branch and Bound Method* and is applicable to pure as well mixed IPP.

The basic method involves dividing the feasible region into smaller sub-sets, each sub-set is considered sequentially until a feasible solution giving the optimal value of objective function is arrived at. The procedure is as given on the following steps:

Step I. Optimal solution of the Linear Programming problem is obtained without considering the restrictions of integer.

Step II. Test the integrality of the optimal solution obtained above.

(a) If the solution turns out to be in integers, then this is the optimum solution of the given IPP.

(b) If the solution is not in integers, then proceed to step III.

Step III. Consider the upper bound values of the objective function, determine the lower bound values by rounding of to the integer values of the decision variables.

Step IV. Sub-divide the given LPP into two problem as follows:

Sub-Problem I – Given LPP with an additional constraint $x_j \leq [x^*_j]$

Sub-Problem II – Given LPP with an additional constraint $x_j \geq [x_j^*] + 1$

where x_j^* is the optimum value of x_j (not an integer) and $[x_j^*]$ is the largest integer contained in x_j^*.

Step V. Solve the above two sub-problems. The following cases may arise:

(a) Optimum solution of the two sub-problems is in integers, then the solution obtained is the optimal solution.

(b) One-sub-problem-Integral

Second-sub-problem-No feasible solution.

In this case, the optimum solution is that of the integral solution of sub-problem one. Second sub-problem solution is ignored.

(c) One sub-problem-Integral

Second sub-problem-Non-integral

In this case, repeat the steps III and IV for the second sub-problem.

Step VI. Repeat steps III to V until we get all solutions with integral values.

Step VII. Out of the integral value solutions achieved, select the one, which gives the optimum value of Z.

Example 14: *Min.* $Z = -4x_1 + x_2 + 2x_3$

Subject to $2x_2 - 3x_2 + 2x_3 \leq 12$

$-5x_1 + 2x_2 + 3x_3 \geq 4$

$3x_1 - 2x_3 = -1$

$x_1, x_2, x_3 \geq 0$

Solution: In simple form, the problem reduces to

$$\text{Min. } Z = -4x_1 + x_2 + 2x_3$$

Subject to $2x_1 - 3x_2 + 2x_3 + x_4 = 12$

$-5x_1 + 2x_2 + 3x_3 - x_5 = 4$

$-3x_1 + 2x_3 = 1$

$x_1, x_2, x_3, x_4, x_5 \geq 0$

Since the above equations do not contain basic variables, artificial variables x_6 and x_7 are added to the problem. Then the problem is

$$\text{Min. } z = -4x_1 + x_2 + 2x_3 + 0x_4 + \text{M}\, x_6 + \text{M}\, x_7$$

Subject to $2x_1 - 3x_2 + 2x_3 + x_4 = 12$

$-5x_1 + 2x_2 + 3x_3 - x_5 + x_6 = 4$

$-3x_1 + 2x_3 + x_7 = 1$

Let S_1 to S_7 denote the column vectors corresponding to x_1 to x_7.

$$P_1 = \begin{bmatrix} 2 \\ -5 \\ -3 \end{bmatrix} \qquad P_2 = \begin{bmatrix} -3 \\ 2 \\ 0 \end{bmatrix} \qquad P_3 = \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix}$$

$$P_4 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad P_5 = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \qquad P_6 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$P_7 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad b = \begin{bmatrix} 12 \\ 4 \\ 1 \end{bmatrix}$$

As x_4, x_6 and x_7 from the initial basis, we have

$$B = \begin{bmatrix} P_4 & P_6 & P_7 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$$

$B^{-1} I = b^{-1} = B^{-1} b = b$

The initial table of the revised simplex is given below:

Basic variables	B^{-1}	Solution values	Entering Variable	Pivot Column
x_4	$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	12	x_3	2
x_6		4		3
x_7		1		2

The Simplex multipliers are:

$$X = (x_1\, x_2\, x_3) = e_i B^{-1}$$

$$= (0\ M\ M)\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} = (0, M, M)$$

$$\therefore \quad C_1 - XP_1 = -4 - (0, M, M)\begin{bmatrix}2\\-5\\3\end{bmatrix} = 8M - 4$$

$$C_2 - XP_2 = 1 - (0, M, M)\begin{bmatrix}-3\\2\\0\end{bmatrix} = 1 - 2M$$

$$C_3 - XP_3 = 2 - (0, M, M)\begin{bmatrix}2\\3\\2\end{bmatrix} = 2 - 5M$$

$$C_5 - XP_5 = 0 - (0, M, M)\begin{bmatrix}0\\-1\\0\end{bmatrix} = M$$

As $C_3 - XP_3$ is the most negative values X_3 will be the entering variable.

The first solution is $$P_3 = B^{-1} P_3 = \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}2\\3\\2\end{bmatrix} = \begin{bmatrix}2\\3\\2\end{bmatrix}$$

2 is the first or key element.

Applying the minimum ratio rule $\frac{12}{2} = 6,\ \frac{4}{3} = 1.3,\ \frac{1}{2} = 0.5$

Minimum ratio is of X_7 so it will be the outgoing variable.

15.11 GOAL PROGRAMMING

Introduction

Goal Programming is relatively a new concept, the work on which began only in early sixties and has been developed during the seventies by Charnes Cooper and Lee. When the multiple goals of an organisation are conflicting, goal programming is extremely helpful. It may be defined as

Goal Programming is a mathematical problem in which the constraints of linear programming problems are treated as goals in the objective function. Effort is made to come as close as possible to the achievement of the goals in order of priority set by the decision-makers.

We have seen in the previous chapters, that we have restricted ourselves to the goal of either maximizing profits or minimizing costs in the linear programming problems. An organisation can have many objectives and that too with conflicting interests. It is not possible to solve such real life problems with the help of mathematical models already developed as these can solve only one objective function. As mentioned earlier any organisation could have a number of objectives as a matter of fact, Peter F. Drucker, the management guru of the century has suggested eight objectives for the organizations. Some of them are, increasing market share, maximizing the returns of different types of stakeholders, social responsibility and so on. Such objectives are selected by the management based on their philosophy, mission and strategy that they want to follow.

Since the LPP can measure the objective function in one dimension only, *i.e.,* it can either maximize profit or minimize profit or minimize costs, a new mathematical technique has been developed to find solution to problems with multiple, often conflicting objectives. In this technique all goals of the management are considered in the objective function and only business environment constraints are treated as constraints. Also, goals are set to be satisfied to the best fit solution 'as close as possible' level and not for optimal or best fit solutions. A set of solution satisfying the business environment conditions/constraints are provided in order of priority and effort is made to minimize the deviation from the set goals. There have been very few applications for Goal Programming techniques in business and industry. First effort to use this technique was made by Charnes Cooper who used it for advertising planning and also for manpower planning problems. Though Goal Programming has lot of flexibility and can have applications as wide-ranging as that of Linear Programming, yet its potential has not been realized and not much work done in this field. It is useful in the practical problems and it is more realistic in its approach. Some of the area where Goal Programming may be used are:

1. ***Marketing Management:*** Marketing management is a very vast discipline in which the organization could have many objectives and that too conflicting, *i.e.,* it is possible to achieve one at the cost of the other. Goals could be
   (i) Maximizing market share
   (ii) Maximizing profit margin/item sold
   (iii) Minimize advertising costs
   (iv) Optimize brand image.
2. ***Production Planning and Control (PPC):*** There is lot of contradicting requirements in production, like
   (i) Minimize operation time
   (ii) Minimize cost
   (iii) Maximize quality of the product
   (iv) Optimize resource utilization.
3. ***Inventory Management:*** Conflicting goals could be
   (i) Minimize stock outs
   (ii) Minimize storage cost
   (iii) Minimize lead-time (Just in Time).

It must be noted that Goal Programming aims at satisfaction of the goals set by the management or decision-makers. Exact achievement of the objective is not aimed. The technique attempts to do so in order of priority of the objectives, again decided by the management. Often, it is a complex task for decision-maker to decide the priority and accept the solution as satisfactory.

Formulation of Goal Programming Mathematical Model

As seen earlier, the first and the most important step in solution of a problem is the ability of the management to convert the problem into a mathematical model, which represents the problem. Here a number of assumptions have to be made *perforce* as we are trying to convert the real life situation into a scientific model written on a piece of paper, unless the problem is very clearly conceptualized by the experts it will not represent the real world problems and solution will give misleading results. It is a complex process. Finding the solution, using the model is again a very time-consuming, complicated and cumbersome process, however the computers and their software can help the decision-maker a great deals in this.

Steps involved in formulation of the goal programming model are as follows:

Step I. *Identification of decision-variable and constraints*: This is the vital step in finding a solution to the problem. Clear identification of all the decision variables and environment conditions, which are the constraints in the equations on the RHS, has to be determined. RHS constraints are:

(i) Available resources as specified in the problem.
(ii) Goals specified by the decision-maker.

Step II. *Formulation of objectives or goals of the problem:* As discussed earlier an organisation could have more than one objectives. Some of these could be

(i) Maximize profits
(ii) Maximize gain of shareholders
(iii) Maximize Machine utilization
(iv) Maximize Manpower utilization
(v) Maximize Mean Time Between Failures (MTBF) of machines
(vi) Minimize operation costs
(vii) Minimize operational time of the machine
(viii) Minimize overall time of production of the product
(ix) Optimize use of raw material
(x) Satisfy social responsibilities
(xi) Maximize quality of the product
(xii) Satisfy many Government rules and other legal requirements.

Step III. *Formulation of the constraints*: The constraints of the problem must be formulated. A constraint represents relationship between different variables in a problem. It could be the relationship between the decision-variables and the goals or objectives selected to be satisfied in order of priority.

Step IV. *Identify least important and redundant goals*: This is done to remove these from the problem which helps in simplifying the problem to some extent. This is again based on the judgement of the management.

Step V. *Establishing the objective function*: Objective function has to be established based on the goals selected by the decision maker. Priority weightage factors have to be allotted to deviational variables.

The goal processing models can be mathematically put as

Minimize objective function $$Z = \sum_{i=1}^{m} W_i (d_i^+ + d_i^-)$$

Subject to the constraints

$$\sum_{i=1}^{n} a_{ij} X_j + d_i^+ + d_i^+ = b_i \quad \text{where } i = 1, 2, 3, ..., n \text{ and } x_j, d_i^-, d_i^+, i, j \geq 0$$

where j is the decision variable.

W_i = is the weightage of goal i

d_i^- = degree of underachievement of goal i

d_i^+ = degree of overachievement of goal i

As seen earlier, Goal Programming attempts at full or partial achievement of goals in order of priority. Low priority goals are considered only after the high priority goals have been considered. This is very difficult to decide, as contribution of a particular goal to the overall well-being of an organisation is very difficult to determine. The concept of underachievement of goals or overachievement of goals

NOTES

may be understood as the most important. Selected goal continues to remain in the problem unless and until the achievement of a lower priority goal would cause the management to fail to achieve a higher priority goal.

Example 15: *ABC Ltd. produces two types of products P-1 and P-2 using common production facilities which are considered a scarce resource by the company. The scarce production facilities are in the two departments of Machining and Assembling. The company is in a position to sell whatever number it produces as their brand enjoys the market confidence. However, the production capacity is limited because of the availability of the scarce resources.*

The company wants to set a goal of maximum daily profit, because of its other problems and constraints and would be satisfied with ₹ 2000 daily profit.

The details of processing time, capacities of each of the departments and unit profit combinations of products P-1 and P-2 are given in the table below.

Type of product	Time to process each product (Hours)		Profit contribution per unit
P-1	*3*	*1*	*200*
P-2	*2*	*1*	*300*
Time available (hours) per day	*100*	*50*	

The company wishes to know the product mix that would get them the desired profit of ₹ 2000 per day. Formulate the problem as goal programming model.

Solution: Let X_1 be the number of units of P-1 to be produced.

Let X_2 be the number of units of P-2 to be produced.

d_i^- = the amount by which actual profit will fall short of ₹ 2000/day.

d_i^+ = the amount by which actual profit will exceed the desired profit of ₹ 2000/ day.

Minimize $Z = d_i^- + d_i^+$

Subject to $3X_1 + 2X_2 \leq 100$ (Machine hours constraint)

$X_1 + X_2 \leq 50$ (Assembly hours constraint)

and $200\ X_1 + 300 \times 2 + d_i^- + d_i^+ = 2000.$ (Desired profitgoal constraint)

where $X_1, X_2, d_i^+, d_i^- \geq 0.$

Example 16: *The manufacturing plant of an electronic firm produces two types of television sets, both colour and black and white, according to the past experience, production of either a colour or a black and white set requires on an average of one hour in the plant. The plant has a normal production capacity of 40 hours a week.*

The marketing department reports that, because of the limited sales opportunity, the maximum numbers of colour and black and white sets that can be sold are 24 and 30 respectively for the week. The gross margin from the sale of a colour set is ₹80, whereas it is 40 from the black and white set.

The chairman of the company has set the following goals arranged in the order of their importance to the organisation.

(*i*) *First he wants to avoid an under utiliztion of normal production capacity (no lay-offs of production workers).*

(*ii*) *Second he wants to sell as many television sets as possible. Since the gross margin from the sale of colour television set is twice the amount from a black and white, he has twice as much desire to achieve sale for colour sets as for black and white sets.*

(*iii*) *Third the chairman wants to minimize the overtime operation of the plant as much as possible.*

Formulate this as a goal-programming problem and solve it.

Solution: Let X_1 and X_2 denotes the number of colour TV sets and number of black and white TV sets for production respectively.

(i) The production capacity of both types of TV sets is given by

$X_1 + X_2 + d_1^- - d_2^+ = 40$ d_1^+ and d_2^- are deviational variables.

respectively underutilization and overutilization (overtime) of normal operation of plant respectively.

(ii) The sale capacity of two types of TV sets is given by

$$X_1 + d_2^- - d_2^+ = 24$$

$$X_2 + d_3^- - d_3^+ = 30$$

where $d_2^- - d_2^+$ are the deviational variables representing under achievements of sales towards goals and $d_3^- - d_3^+$ represents deviational variables to represent overachievement of sales goals.

(iii) Let p_1 and p_2 be the priority of the goals, complete mathematical formulation of goal programming is

$$\text{Minimize} \qquad Z = p_1 d_1^- + 2p_1 d_2^- + p_2 d_3^- + p_2 d_1^+$$

Subject to the constraints

$$X_1 + X_2 + d_1^- - d_1^+ = 40,\ X_1 + d_2^- - d_2^+ = 24$$

$$X_3 + d_3^- - d_3^+ = 30 \text{ and } X_1, X_2, d_1^-, d_2^-, d_3^-, d_1^+, d_2^+, d_3^+ \geq 0$$

Graphical Method: The graphical method used in Goal Programming is quite similar to the one used in Linear programming problems. The only difference is that in LPP

only one objective function is achieved either maximization or minimization with only one goal. In Goal Programming, there are a number of goals and total deviation from these goals is required to be minimized. The minimization in deviations is done in order of priority. The following procedure is followed:

Step I. Formulation of Linear Goal Programming mathematical model.

Step II. Construction of graph of all the goals in relation with the decision-variables.

For each goal write an equation with positive and negative deviation variables and set the equation to zero. For all the goal equation two points are selected arbitrarily and joined with straight lines. Positive deviations are indicated with → arrow and negative deviation by ← arrow for each goal.

Step III. Determine the goal line of that goal which has the highest priority. Identify the feasible region (area) with respect to the goal with highest priority.

Step IV. Proceed to the next highest priority and determine the best solutions space with respect to these goals corresponding to this priority.

Step V. Determine the optimal solution.

Example 17: *A manufacturer produces two types of products A and B. The plant has production capacity of 500 hours a month and production of product A or B or an average requires one hour in the plant. The number of products A and B sold every month and the net profit from the sales of these products are given in the following table:*

Type of product	*Number sold in a month*	*Net profit*
A	*250*	
B	*300*	

The MD of the company has set the following goals, which are arranged in order of priority.

P_1 *No underutilization of plant production capacity.*

P_2 *Sell maximum possible numbers of products A and B. The MD has twice as much desire to sell product A as for product B, because the net profit from the sale of product A is twice the amount from that of product B.*

P_3 *Minimize overtime operation of the plant.*

Formulate the above as a goal-programming problem and solve it.

Solution: Let X_1 and X_2 be the number of products of A and B. Since overtime operations are not allowed.

$X_1 + X_2 + d_1^- - d_2^+ = 500$ (Plant capacity constraint)

where d_1^- = under utilization of production capacity variable

d_1^+ = overtime production operation capacity variable.

Since goal is the maximization of sales, hence positive deviation will not appear in constraints related with sales.

Then $X_1 + d_2^- = 250$

and $X_2 + d_3^- = 300$

where, d_2^- = under achievement of sales goal for product A

d_3^- = under achievement of sales goal for product B

Now, the goal programming mathematical model can be written as

Minimize $Z = p_1 d_1^- + 2p_2 d_2^- + p_2 d_3^- + p_3 d_1^+$

Subject to the constraints

$X_1 + X_2 + d_1^- - d_1^+ = 500$

$X_1 + d_2^- = 250$

Check Your Progress

Choose the correct option for the following statements:

6. Proportionality condition for LP method requires that usage of resources is ..... to the value of the variable.
   (a) inversely proportional (b) directly proportional
   (c) equal (d) reciprocal
7. Parametric programming is a type of .....
   (a) operations resarch (b) mathematical optimization
   (c) linear programming problem (d) integer programming
8. If all the variables take only integer values then it is
   (a) Mixed integer linear programming (b) Pure integer linear programming
   (c) Dynamic programming (d) Goal programming
9. The best known search method for integer programming is
   (a) Cutting plane method (b) Branch and bound method
   (c) Graphical method (d) Simplex method
10. Conflicting goals could be .....
   (a) Minimize stock outs (b) Minimize storage cost
   (c) Minimize lead-time (d) All of the above.

15.12 SUMMARY

- "Operational Research is the application of the methods of science to complex problems arising in the direction and management of large systems of men, machines, materials and money in industry, business, government and defence.
- Linear Programming (LP) is a mathematical technique, which is used for allocating limited resources to a number of demands in an optimal manner.
- LP technique establishes a linear relationship between two or more variables involved in management decisions described above.
- In decision making all the decisions are taken through some variables which are known as decision variables. In engineering design, these variables are known as design vectors.
- When the value of the objective function is maximum/minimum at more than one corner points then '*multiple optima*' solutions are obtained.
- When there does not exist any common feasible region, then there does not exist any solution. Then the given LPP is called *infeasible i.e.,* having *no solution*.
- Simplex method is an algebraic procedure in which a series of repetitive operations are used and we progressively approach the optimal solution.
- The process of reaching the optimal solution through different stages is also called iterative, because the same computational steps are repeated a number of times before the optimum solution is reached.
- Consider the symmetric primal (max. type) and Dual (min. type). The value of the objective function of the (dual) minimum problem for any feasible solution is always greater than or equal to that of the maximum problem (primal) for any feasible solution.
- ***Proportionality:*** It requires that usage of resources is directly proportional to the value of the variables
- ***Additivity:*** It demands that the objective function (Z) must be the direct sum of the individual contribution of different variables.
- In Linear Programming it is assumed that the decision variables can take continuous values, *i.e.*, these could be fractions or in integer. Integer Programming deals with solutions in which some or all the variables can assume integers non-negative values only.
- ***Cutting Plane Method:*** In this method of Integer Linear Programming, certain 'secondary' conditions are added in such a manner that the ultimate result satisfies the conditions of only integer solutions.

- ***Search Methods:*** Here all the possible feasible integers only are considered as the solution. The best known search method is called the *Branch-and-Bound techniques*.
- The most common use of integer programming is found in the real world problems related to investment decisions, budgeting, Production Planning and Control (PPC) in manufacturing industry, travelling salesmen, etc.
- In certain type of problems of the variables of an Integer Programming Problem (IPP) have the constraint of a upper limit or a lower limit or both upper and lower bounds. The method used to solve such problem is called *Branch and Bound Method* and is applicable to pure as well mixed IPP.
- Goal Programming is a mathematical problem in which the constraints of linear programming problems are treated as goals in the objective function. Effort is made to come as close as possible to the achievement of the goals in order of priority set by the decision-makers.
- Goal Programming aims at satisfaction of the goals set by the management or decision makers. Exact achievement of the objective is not aimed.

15.13 GLOSSARY

1. ***Convex Set:*** A set X is said to be convex if $x_1, x_2 \in X$ for $0 \leq \lambda \leq 1$
   then $x_3 = \lambda x_1 + (1 - \lambda)\, x_2 \in X$
2. ***Extreme Point or Corner Point of Convex Set:*** It is a point in the convex set which cannot be expressed as $\lambda x_1 + (1 - \lambda) x_2$ where x_1 and x_2 are any two points on the convex set.
3. ***Feasible Solution:*** A solution which satisfies all the constraints in LPP is called feasible solution.
4. ***Basic Solution:*** Let m = number of constraints, n = number of variables
   If $m < n$, then the solution from the system $Ax = b$ is called basic solution.
5. ***Basic Feasible Solution (BFS):*** A solution which is basic as well as feasible is called basic feasible solution.
6. ***Degenerate BFS:*** If a basic variable takes the value zero in a BFS, then the solution is said to be degenerate.
7. ***Optimal BFS:*** The BFS which optimizes the objective function is called optimal BFS.
8. ***Unbounded Solution:*** Sometimes the optimum solution is obtained at infinity, then the solution is called unbounded solution.

9. ***Simplex Table:*** Calculations are done in a table in which for each constraint there will be a row and for each variable there will be a column. This table is called simplex table.
10. ***Bounded Variables:*** In certain LP problems, some or all of the variables may have lower and upper limits to their values. We may convert these constraints to equalities by introducing slack and surplus variables. Such constraints are called bounded variables.

15.14 ANSWERS TO CHECK YOUR PROGRESS

1. linear relationship
2. mathematical technique
3. design vectors
4. degenerate
5. feasible solution
6. False
7. True
8. False
9. True
10. True
11. (b)
12. (b)
13. (b)
14. (b)
15. (d)

15.15 TERMINAL AND MODEL QUESTIONS

Using graphical method solve the following LPP:

1. Maximize $z = 13x_1 + 117x_2$

Subject to, $x_1 + x_2 \leq 12$,

$x_1 - x_2 \geq 0$,

$4x_1 + 9x_2 \leq 36$,

$0 \leq x_1 \leq 2$ and $0 \leq x_2 \leq 10$.

2. Maximize $z = 3x_1 + 15x_2$

Subject to, $4x_1 + 5x_2 \leq 20$,

$x_2 - x_1 \leq 1$,

$0 \leq x_1 \leq 4$ and $0 \leq x_2 \leq 3$.

3. Minimize $z = 2x_1 + 3x_2$

Subject to, $x_2 - x_1 \geq 2$,

$5x_1 + 3x_2 \leq 15$,

$2x_1 \geq 1$,

$x_2 \leq 4$,

$x_1, x_2 \geq 0$.

4. Minimize $z = 10x_1 + 9x_2$

Subject to, $x_1 + 2x_2 \leq 10$,

$x_1 - x_2 \leq 0$,

$x_1 \leq 0, x_2 \geq 0$.

Duality of simplex method Obtain the dual of the following LP problems:

5. Maximize $z = 4x_1 + 2x_2 + x_3 + 6x_4$

Subject to, $6x_1 - 3x_2 + x_3 + 5x_4 \leq 15$,

$x_1 - x_2 + 6x_3 + 2x_4 \geq 8$,

$x_1, x_2, x_3, x_4 \geq 0$

6. Maximize $z = 2x_1 + x_2$

Subject to, $2x_1 + 3x_2 \geq 4$,

$3x_1 + 4x_2 \leq 10$,

$x_1 + 5x_2 = 9$,

$x_1 \geq 0, x_2 \geq 0$.

7. Minimize $z = 3x_1 + 4x_2 - x_3$

Subject to, $2x_1 + 3x_2 + 5x_3 \geq 10$,

$3x_1 + 10x_3 \leq 14$,

$x_1 \geq 0, x_2 \leq 0, x_3 \geq 0$.

8. Minimize $z = 10x_1 + 15x_2$

Subject to, $3x_1 + 2x_2 = 15$,

$5x_1 + 4x_2 = 20$,

x_1, x_2 unrestricted in sign.

9. Use principle of duality to solve the following LP problems:

(a) Minimize $z = 4x_1 + 3x_2$

Subject to: $2x_1 + x_2 \geq 40,\ x_1 + 2x_2 \geq 50,\ x_1 + x_2 \geq 35$

$x_1, x_2 \geq 0$

(b) Maximize $z = 2x_1 + x_2$

Subject to: $x_1 + 2x_2 \leq 10,\ x_1 + x_2 \leq 6,\ x_1 - x_2 \leq 2,\ x_1 - 2x_2 \leq 1$

$x_1, x_2 \geq 0$

(c) Minimize $z = 6x_1 + x_2$

Subject to: $2x_1 + x_2 \geq 3,\ x_1 - x_2 \geq 0,\ x_1, x_2 \geq 0$

(d) Minimize $z = 30x_1 + 30x_2 + 10x_3$

Subject to: $2x_1 + x_2 + x_3 \geq 6,\ x_1 + x_2 + 2x_3 \leq 8,\ x_1, x_2, x_3 \geq 0$

(e) Maximize $z = 5x_1 + 2x_2$

Subject to: $x_1 - x_2 \leq 1,\ x_1 + x_2 \geq 4,\ x_1 - 3x_2 \leq 3,\ x_1, x_2 \geq 0$

10. Using the complementary slackness condition solve the following LP problem :

Maximize $z = 2x_1 + 3x_2 + 6x_3$

Subject to: $x_1 + 3x_2 + 4x_3 \leq 4,\ 2x_1 + x_2 + 3x_3 \leq 2,\ x_1, x_2, x_3 \geq 0.$

11. With the help of the following example, verify that the dual of the dual is the primal.

Maximize $z = 3x_1 + 2x_2 + 5x_3$

Subject to: $4x_1 + 3x_2 - x_3 \leq 20,\ 3x_1 + 2x_2 + 5x_3 = 18,$

$0 \leq x_1 \leq 4,\ x_2 \geq 0,\ x_3 \leq 0.$

12. Verify the fundamental theorem of duality using the following LP problems:

(a) Maximize $z = 2x_1 + 10x_2$

Subject to: $2x_1 + 5x_2 \leq 16,\ 6x_1 \leq 30,\ x_1, x_2 \geq 0.$

(b) Minimize $z = 2x_1 - x_2$

Subject to: $x_1 + x_2 \leq 5,\ x_1 + 2x_2 \geq 8,\ x_1, x_2 \geq 0.$

Formulation of Linear Programming

13. A manufacturer of furniture makes only chair and tables. A chair requires two hours on m/c A and six hours on m/c B. A table requires five hours on m/c

A and two hours on m/c B. 16 hours are available on m/c A and 22 hours on m/c B per day. Profits for a chair and table be ₹ 1 and ₹ 5 respectively. Formulate the LPP of finding daily production of these items for maximum profit and solve graphically.

14. A tailor has 80 sq. m of cotton material and 120 sq. m of woolen material. A suit requires 1 sq. m of cotton and 3 sq. m of woolen material and a dress requires 2 sq. m of each. A suit sells for ₹ 500 and a dress for ₹ 400. Pose a LPP in terms of maximizing the income.
15. A company owns two mines: mine A produces 1 tonne of high grade ore, 3 tonnes of medium grade ore and 5 tonnes of low grade ore each day; and mine B produces 2 tonnes of each of the three grades of ore each day. The company needs 80 tonnes of high grade ore, 160 tonnes of medium grade ore and 200 tonnes of low grade ore. If it costs ₹ 200 per day to work each mine, find the number of days each mine has to be operated for producing the required output with minimum total cost.
16. A company manufactures two products A and B. The profit per unit sale of A and B is ₹ 10 and Rs. 15 respectively. The company can manufacture at most 40 units of A and 20 units of B in a month. The total sale must not be below ₹ 400 per month. If the market demand of the two items be 40 units in all, write the problem of finding the optimum number of items to be manufactured for maximum profit, as a problem of LP. Solve the problem graphically or otherwise.

Goal Programming

17. What is the concept of Goal Programming? Discuss how it can be used for solving problems.
18. How is the goal programming mathematical model prepared?
19. How is the goal programming used in decision-making, discuss by taking examples from industry?
20. What are the applications of goal programming? Give examples.
21. The manufacturing plant of an electronics firm produces two types of television sets with colour and black and white. According to the past experience, production of either a colour of a black and white set requires an average of one hour in the plant. The plant has a normal production capacity of 60 hours a week. The marketing department reports that, because of limited sales opportunity, the maximum numbers of colour and black and white sets that can be sold are 20 and 24 respectively for the week. The gross margin from the sale of a colour set is ₹ 1200 whereas it is ₹ 200 from a black and white set.

    Managing Director of the Company has set the following goals as arranged in the order of their importance to the organisation of:

(i) First he wants to avoid any under utilization of normal production capacity (no layoffs of production workers).

(ii) Second he wants to sell as many television sets as possible. Since the gross margin from the sale of colour television set is six times the amount from a black and white set, he has six times as much desire to achieve sales for colour television sets as for black and white sets.

(iii) Third, the MD wants to minimize the overtime operation of the plant as much as possible. Formulate this as a Goal Programming problem and solve it.

22. A textile company produces two types of materials, curtain cloth and dress material. The Curtain cloth is produced according to direct orders from the furniture manufactures. The dress material, on the other hand is distributed to retail stores. The average production rates for the curtain cloth and for the dress material are identical, 800 metres per hour. By running two shifts the operational capacity of the plant is 100 hours per week.

    The marketing department reports that the maximum estimated sales for the following week is 50000 metres of the curtain cloth and 40000 metres of dress material. According to accounts department, approximate profit from a metre of curtain cloth material is ₹ 5 and from the dress material is ₹ 3. The Chairman of the company of believes that a good employer-employee relationship is an important factor for business success. Hence, he decides that a stable employment level is a primary goal of the company. Therefore, whenever there is a demand exceeding normal production capacity, he simply expands production expands production capacity for providing overtime. However, he also feels overtime operation of the plant too more than 12 hours a week should avoided because of the increased costs. The Chairman has the following goals:

    (i) The first goal is to avoid any underutilization of production capacity (*i.e.*, to maintain stable employment at normal capacity).

    (ii) The second goal is to limit the overtime operation of the plant to 12 hours.

    (iii) The third goal is to achieve the sales target of 50000 metres of curtain cloth are 40000 metres of dress material.

    (iv) The last goal of the chairman is to minimize to minimize overtime operation of the plant as much as possible.

    Formulate and solve the problem as a Goal Programming Problem.

Integer Programming

23. What is integer programming? When can it be applied?

24. What do you understand be integer linear programming problem? Discuss giving examples.

25. How does the optimal solution of an integer linear programming problem compare with that of linear programming problem? Explain giving examples.

Minimize $Z = 8x_1 + 5x_2$

Subject to constraints

$$3x_1 + 2x_2 \leq 18$$
$$5x_1 + 6x_2 \leq 30$$

where x_1 and x_2 must be non-negative integer.

26. ABC Corporation is considering four possible investment opportunities Investment profits in ₹ thousands, for the investments are shown below.

Profit	PV of Expected Return	Capital Requirement by Projects		
		Year *1*	Year *2*	Year *3*
P–1	500	800	600	500
P–2	1000	1500	500	400
P–3	250	350	200	—
P–4	600	800	500	—
Capital available for investment		1800	1000	500

In addition, project 1 and 2 are mutually exclusive and 4 is possible only if 3 is accepted. Formulate an integer programming model to determine which product should be accepted to maximize the present value from accepted projects.

Parametric Programming

27. Min $Z = x - y$

subject to (i) $3x - y \geq 4$, (ii) $2x + y \leq 3$

where α is an arbitrary, and small scalar number but finite and β is an arbitrary and large scalar number but finite. Perform a complete parametric programming analysis.

28. (a) Max $Z = (\lambda - 1)\, x_1 + x_2$

subject to (i) $x_1 + 2x_2 \leq 10$ (ii) $2x_1 + x_2 \leq 11$

(iii) $x_1 - 2x_2 \leq 3$ and $x_1, x_2 \geq 0$.

(b) Min $Z = -\lambda x_1 - \lambda x_2 - x_3 + x_4$

subject to (i) $3x_1 - 3x_2 - x_3 + x_4 \geq 5$

(ii) $2x_1 - 2x_2 + x_3 - x_4 \leq 3$ and $x_1, x_2, x_3, x_4 \geq 0$.

Perform a complete parametric programming analysis.

15.16 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitive Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics,* Firewall Media, New Delhi

UNIT 16: TRANSPORTATION PROBLEM

NOTES

Structure

16.0 INTRODUCTION

In the previous unit, you have learnt about linear programming problems. Transportation problem is basically linear programmed. Simplex algorithm can be used to solve any LP model but since it is very laborious and time consuming, a model with simplified calculations is required.

Transportation problems deal with the determination of minimum cost for transporting one commodity from a number of resources for example, manufacturing units to a number of destinations *e.g.*, clearing and forwarding. (C & F) agents.

The name can be misleading because it is not only that transportation and destination problems can be solved. It can be extended to be used for facility and location (plant location, machine assignments) planning can be also for many Production Planning and Control (PPC) problems.

16.1 UNIT OBJECTIVES

After reading this unit, you will be able to:

- Define transportation problem
- Formulate transportation problem mathematically
- Find initial basic feasible solution of transportation problem
- Determine the minimum transportation cost
- Make an unbalanced transportation problem a balanced one
- Define degeneracy in transportation problem
- Understand maximization objective function related to transportation problem

16.2 MATHEMATICAL FORMULATION OF TRANSPORTATION PROBLEM

Transportation problem (T.P.) is generally concerned with the distribution of a certain commodity/product from several origins/sources to several destinations with minimum total cost through single mode of transportation. If different modes of transportation are considered then the problem is called 'solid T.P.'. In this unit, we shall deal with simple T.P.

Suppose there are m factories where a certain product is produced and n markets where it is needed. Let the supply from the factories be $a_1, a_2, ..., a_m$ units and demands at the markets be $b_1, b_2, ..., b_n$ units.

Also consider

c_{ij} = Unit of cost of shipping from factory i to market j.

x_{ij} = Quantity shipped from factory i to market j.

Then the LP formulation can be started as follows:

$$\text{Minimize } z = \text{Total cost of transportation}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\, x_{ij}$$

$$\text{Subject to, } \sum_{j=1}^{n} x_{ij} \le a_i,\ i = 1, 2, ..., m.$$

(Total amount shipped from any factory does not exceed its capacity)

NOTES

$$\sum_{i=1}^{m} x_{ij} \geq b_j, \; j = 1, 2, ..., n.$$

(Total amount shipped to a market meets the demand of the market)

$x_{ij} \geq 0$ for all i and j.

Here the market demand can be met if

$$\sum_{i=1}^{m} a_i \geq \sum_{j=1}^{n} b_j.$$

If $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$ *i.e.,* total supply = total demand, the problem is said to be "Balanced T.P." and all the constraints are replaced by equality sign.

$$\text{Minimize } z = \sum\sum c_{ij} \; x_{ij}$$

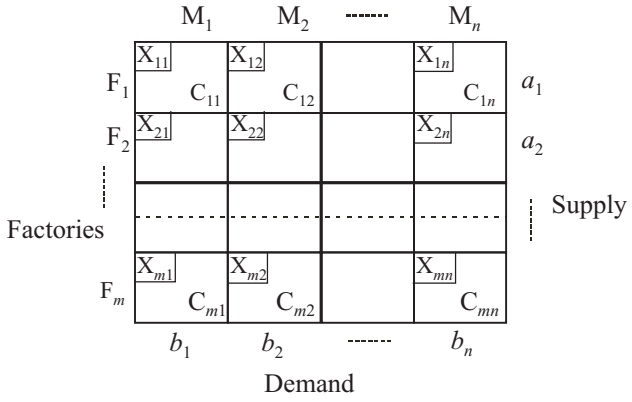
$$\text{Subject to, } \sum_{j=1}^{n} x_{ij} = a_i, \; i = 1, 2, ..., m.$$

$$\sum_{i=1}^{m} x_{ij} = b_j, \; j = 1, 2, ..., n.$$

$x_{ij} \geq 0$ for all i and j.

(Total $m + n$ constraints and mn variables)

The T.P. can be represented by *table form* as given below:



In the above, each cell consists of decision variable x_{ij} and per unit transportation cost c_{ij}.

Theorem 1: A necessary and sufficient condition for the existence of a feasible solution to a T.P. is that the T.P. is balanced.

Proof: (Necessary part)

Total supply from an origin $\sum_{j=1}^{n} x_{ij} = a_i, \quad i = 1, 2, ..., m.$

Overall supply, $\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \sum_{i=1}^{m} a_i$

Total demand met of a destination

$$\sum_{i=1}^{m} x_{ij} = b_j, j = 1, 2, ..., n.$$

Overall demand, $\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} = \sum_{j=1}^{n} b_j$.

Since overall supply exactly met the overall demand.

$$\sum_{i}\sum_{j} x_{ij} = \sum_{j}\sum_{i} x_{ij}$$

$$\Rightarrow \qquad \sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j.$$

(Sufficient part) Let $\sum_{i} a_i = \sum_{j} b_i = l$ and $x_{ij} = a_i b_j / l$ for all i and j.

Then $$\sum_{j=1}^{n} x_{ij} = \sum_{j=1}^{n} (a_i b_j)/l = a_i \left(\sum_{j=1}^{n} b_j \right) \Big/ l = a_i = 1, 2, ..., m.$$

$$\sum_{i=1}^{m} x_{ij} = \sum_{i=1}^{m} (a_i b_j)/l = b_j \left(\sum_{i=1}^{m} a_i \right) \Big/ l = b_j, j = 1, 2, ..., n.$$

$x_{ij} \geq 0$ since a_i and b_j are non-negative.

Therefore x_{ij} satisfies all the constraints and hence x_{ij} is a feasible solution.

Theorem 2: The number of basic variables in the basic feasible solution of an $m \times n$ T.P. is $m + n - 1$.

Proof: This is due to the fact that the one of the constraints is redundant in balanced T.P.

We have overall supply, $\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \sum_{i=1}^{m} a_i$

and overall demand $\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} = \sum_{j=1}^{n} b_j$

Since $\sum_i a_i = \sum_j b_i$, the above two equations are identical and we have only $m + n - 1$ independent constraints. Hence the theorem is proved.

NOTES

Note: 1. If any basic variable takes the value zero then the basic feasible solution (BFS) is said to be degenerate. Like LPP, all non-basic variables take the value zero.

2. If a basic variable takes either positive value or zero, then the corresponding cell is called 'Basic cell' or 'Occupied cell'. For non-basic variable the corresponding cell is called 'Non-basic cell' or 'Non-occupied cell' or 'Non-allocated cell'.

Loop: This means a closed circuit in a transportation table connecting the occupied (or allocated) cells satisfying the following:

(i) It consists of vertical and horizontal lines connecting the occupied (or allocated) cells.

(ii) Each line connects only two occupied (or allocated) cells.

(iii) Number of connected cells is even.

(iv) Lines can skip the middle cell of three adjacent cells to satisfy the condition (ii).

The following are the examples of loops.

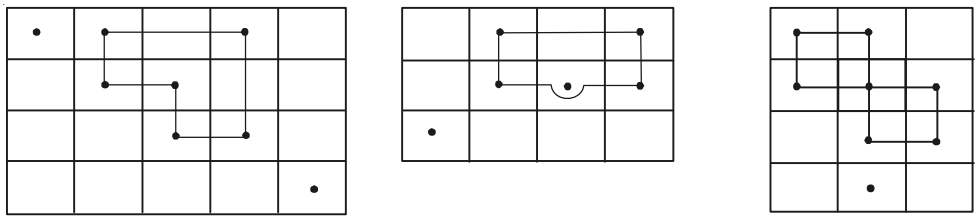


Fig. 16.1

Note: A solution of a T.P. is said to be basic if it does not consist of any loop.

Check Your Progress

Fill in the blanks:

1. If different modes of transportation are considered then the problem is called .....
2. The number of basic variables in the basic feasible solution of an $m \times n$ T.P. is .....
3. If the basic variable takes either positive value or zero, then the corresponding cell is called ..... or .....
4. ..... means a closed circuit in a transportation table connecting the occupied cells.
5. A solution of T.P. is said to be ..... if it does not consists of any loop.

16.3 FINDING INITIAL BASIC FEASIBLE SOLUTION

In this section three methods are to be discussed to find initial BFS of a T.P. In advance, it can be noted that the above three methods may give different initial BFS to the same T.P. Also allocation = minimum (supply, demand).

North-West Corner Rule (NWC)

(i) Select the north west corner cell of the transportation table.

(ii) Allocate the min (supply, demand) in that cell as the value of the variable.

If supply happens to be minimum, cross-off the row for further consideration and adjust the demand.

If demand happens to be minimum, cross-off the column for further consideration and adjust the supply.

(iii) The table is reduced and go to step (*i*) and continue the allocation until all the supplies are exhausted and the demands are met.

Example 1: *Find the initial BFS of the following T.P. using NWC rule.*

From \ To	M_1	M_2	M_3	M_4	Supply
F_1	3	2	4	1	20
F_2	2	4	5	3	15
F_3	3	5	2	6	25
F_4	4	3	1	4	40
Demand	30	20	25	25	

Solution: Here, total supply = 100 = total demand. So the problem is balanced T.P.

The north-west corner cell is (1, 1) cell. So allocate min. (20, 30) = 20 in that cell. Supply exhausted. So cross-off the first row and demand is reduced to 10. The reduced table is

	M_1	M_2	M_3	M_4	
F_2	2	4	5	3	15
F_3	3	5	2	6	25
F_4	4	3	1	4	40
	10	20	25	25	

Here the north-west corner cell is (2, 1) cell. So allocate min. (15, 10) = 10 in that cell. Demand met. So cross-off the first column and supply is reduced to 5. The reduced table is

	M_2	M_3	M_4	
F_2	4	5	3	5
F_3	5	2	6	25
F_4	3	1	4	40
	20	25	25	

Here the north-west corner cell is (2, 2) cell. So allocate min. (5, 20) = 5 in that cell. Supply exhausted. So cross-off the second row (due to F_2) and demand is reduced to 15. The reduced table is

	M_2	M_3	M_4	
F_2	5	2	6	25
F_3	3	1	4	40
	15	25	25	

Here the north-west corner cell is (3, 2) cell. So allocate min. (25, 15) = 15 in that cell. Demand met. So cross-off the second column (due to M_2) and supply is reduced to 10. The reduced table is

	M_3	M_4	
F_3	2	6	10
F_4	1	4	40
	25	25	

Here the north-west corner cell is (3, 3) cell. So allocate min. (10, 25) = 10 in that cell. Supply exhausted. So cross-off the third row (due to F_3) and demand is reduced to 15. The reduced table is

	M_3	M_4	
F_4	1	4	40
	25	25	

continuing we obtain the allocation 15 to (4, 3) cell and 25 to (4, 4) cell so that supply exhausted and demand met. The **complete allocation** is shown below:

	M_1	M_2	M_3	M_4
F_1	[20] 3	2	4	1
F_2	[10] 2	[5] 4	5	3
F_3	3	[15] 5	[10] 2	6
F_4	4	3	[15] 1	[25] 4

Thus, the initial BFS is

$$x_{11} = 20, x_{21} = 10, x_{22} = 5, x_{32} = 15, x_{33} = 10, x_{43} = 15, x_{44} = 25.$$

The transportation cost

$$= 20 \times 3 + 10 \times 2 + 5 \times 4 + 15 \times 5 + 10 \times 2 + 5 \times 1 + 25 \times 4$$

$$= ₹\ 310.$$

Least Cost Entry Method (LCM) (or Matrix Minimum Method)

(i) Find the least cost from transportation table. If the least value is unique, then go for allocation.

If the least value is not unique then select the cell for allocation for which the contributed cost is minimum.

(ii) If the supply is exhausted cross-off the row and adjust the demand.

If the demand is met cross-off the column and adjust the supply.

Thus the matrix is reduced.

(iii) Go to step (i) and continue until all the supplies are exhausted and all the demands are met.

Example 2: *Find the initial BFS of Example 1 using least cost entry method:*

	M_1	M_2	M_3	M_4	
F_1	3	2	4	1	20
F_2	2	4	5	3	15
	3	5	2	6	25
F_4	4	3	1	4	40
	30	20	25	25	

Solution: Here the least value is 1 and occurs in two cells (1, 4) and (4, 3). But the contributed cost due to cell (1, 4) is 1 × min (20, 25) *i.e.,* 20 and due to cell (4, 3) is 1 × min. (40, 25) *i.e.,* 25. So we selected the cell (1, 4) and allocate 20. Cross-off the first row since supply exhausted and adjust the demand to 5. The reduced table is given below:

2	4	5	3	15
3	5	2	6	25
4	3	1	4	40
30	20	25	5	

NOTES

The least value is 1 and unique. So allocate min. (40, 25) = 25 in that cell. Cross-off the third column (due to M_3) since the demand is met and adjust the supply to 15. The reduced table is given below:

2	4	3	15
3	5	6	25
4	3	4	15
30	20	5	

The least value is 2 and unique. So allocate min. (15, 30) = 15 in that cell. Cross-off the second row (due to F_2) since the supply exhausted and adjust the demand to 15. The reduced table is given below:

3	5	6	25
4	3	4	15
15	20	5	

The least value is 3 and occurs in two cells (3, 1) and (4, 2). The contributed cost due to cell (3, 1) is 3 × min. (25, 15) = 45 and due to cell (4, 2) is 3 × min. (15, 20) = 45. Let us select the (3, 1) cell for allocation and allocate 15. Cross-off the first column (due to M_1) since demand is met and adjust the supply to 10. The reduced table is given below:

5	6	10
3	4	15
20	5	

Continuing the above method and we obtain the allocations in the cell (4, 2) as 15, in the cell (3, 2) as 5 and in the cell (3, 4) as 5. The complete allocation is shown below:

	M_1	M_2	M_3	M_4
F_1	3	2	4	[20] 1
F_2	[15] 2	4	5	3
F_3	[15] 3	5 5	2	[5] 6
F_4	4	[15] 3	[25] 1	4

The initial BFS is

$$x_{14} = 20,\ x_{21} = 15,\ x_{31} = 15,\ x_{32} = 5,\ x_{34} = 5,\ x_{42} = 15,\ x_{43} = 25.$$

The transportation cost

$= 20 \times 1 + 15 \times 2 + 15 \times 3 + 5 \times 5 + 5 \times 6 + 15 \times 3 + 25 \times 1$

$= ₹ 220.$

Note: If the least cost is only selected columnwise then it is called 'column minima' method. If the least cost is only selected row wise then it is called 'row minima' method.

Vogel's Approximation Method (VAM)

(i) Calculate the row penalties and column penalties by taking the difference between the lowest and the next lowest costs of every row and of every column respectively.

(ii) Select the largest penalty by encircling it. For tie cases, it can be broken arbitrarily or by analyzing the contributed costs.

(iii) Allocate in the least cost cell of the row/column due to largest penalty.

(iv) If the demand is met, cross-off the corresponding column and adjust the supply.

If the supply is exhausted, cross-off the corresponding row and adjust the demand.

Thus the transportation table is reduced.

(v) Go to Step (i) and continue until all the supplies exhausted and all the demands are met.

Example 3: *Find the initial BFS of example 1 using Vogel's approximation method.*

Solution:

	M_1	M_2	M_3	M_4		Row Penalties
F_1	3	2	4	[20] 1	20	(1)
F_2	2	4	5	3	15	(1)
F_3	3	5	2	6	25	(1)
F_4	4	3	1	4	40	(2)
	30	20	25	25		
Column Penalties	(1)	(1)	(1)	②		

Since there is a tie in penalties, let us break the tie by considering the contributed costs. Due to M_4, the contributed cost is $1 \times$ min. $(20, 25) = 20$. While due to F_4, the contributed cost is $1 \times$ min. $(40, 25) = 25$. So select the column due to M_4 for allocation and we allocate min. (20, 25) *i.e.,* 20 in (1, 4) cell. Then cross-off the first row as supply is exhausted and adjust the corresponding demand as 5. The reduced table is

	M_1	M_2	M_3	M_4	Row Penalties
F_2	2	4	5	3	15 (1)
F_3	3	5	2	6	25 (1)
F_4	4	3	1	4	40 ②
Column Penalties	30 (1)	20 (1)	25 (1)	5 (1)	

Here the largest penalty is 2 which is due to F_4. Allocate in (4, 3) cell as min. (40, 25) = 25. Cross-off the third column due to M_3, since demand is met and adjust the corresponding supply to 15. The reduced table is

	M_1	M_2	M_4	Row Penalties
F_2	2	4	3	15 (1)
F_3	3	5	6	25 ②
F_4	4	3	4	15 (1)
Column Penalties	30 (1)	20 (1)	5 (1)	

Here the largest penalty is 2 which is due to F_3. Allocate in (3, 1) cell as min. (25, 30) = 25. Cross-off the third row due to F_3 since supply is exhausted and adjust the corresponding demand to 5. The reduced table is

	M_1	M_2	M_4	Row Penalties
F_2	2	4	3	15 (1)
F_4	4	3	4	15 (1)
Column Penalties	5 ②	20 (1)	5 (1)	

Here the largest penalty is 2 which is due to M_1. Allocate in (2, 1) cell as min. (15, 5) = 5. Corss-off the first column due to M_1 since demand is met and adjust the supply to 10. The reduced table is

	M_2	M_4	Row Penalties
F_2	4	3	10 (1)
F_4	3	4	15 (1)
Column Penalties	20 (1)	5 ①	

Here tie has occurred. The contributed cost is minimum due to (2, 4) cell which is $3 \times \min.(10, 5) = 15$. So allocate min. (10, 5) = 5 in (2, 4) cell. Cross-off the fourth column which is due to M_4 since demand is met and adjust the corresponding supply to 5. On continuation we obtain the allocation of 5 in (2, 2) cell and 15 in (4, 2) cell. The complete allocation in shown below:

	M_1	M_2	M_3	M_4
F_1	3	2	4	[20] 1
F_2	[5] 2	[5] 4	5	[5] 3
F_3	[25] 3	5	2	6
F_4	4	[15] 3	[25] 1	4

The initial BFS is

$$x_{14} = 20, x_{21} = 5, x_{22} = 5, x_{24} = 5, x_{31} = 25, x_{42} = 15, x_{43} = 25.$$

The transportation cost

$$= 1 \times 20 + 2 \times 5 + 4 \times 5 + 3 \times 5 + 3 \times 25 + 3 \times 15 + 1 \times 25$$

$$= ₹\ 210.$$

UV-Method/Modi Method

Taking the initial BFS by any method discussed above, this method find the optimal solution to the transportation problem. The steps are given below:

(i) For each row consider a variable u_i and for each column consider another variable v_j.

Find u_i and v_j such that

$u_i + v_j = c_{ij}$ for every basic cells.

(ii) For every non-basic cells, calculate the net evaluations as follows:

$\bar{c}_{ij} = u_i + v_j - c_{ij}$

If all $\bar{c}_{ij}$ are non-positive, the current solution is optimal.

If at least one $\bar{c}_{ij} > 0$, select the variable having the largest positive net evaluation to enter the basis.

(iii) Let the variable x_{rc} enter the basis. Allocate an unknown quantity θ to the cell (r, c).

Identify a loop that starts and ends in the cell (r, c).

Subtract and add θ to the corner points of the loop clockwise/anticlockwise.

NOTES

(iv) Assign a minimum value of θ in such a way that one basic variable becomes zero and other basic variables remain non-negative. The basic cell which reduces to zero leaves the basis and the cell with θ enters into the basis.

If more than one basic variables become zero due to the minimum value of θ, then only one basic cell leaves the basis and the solution is called degenerate.

(v) Go to step (i) until an optimal BFS has been obtained.

Note: In step (ii), if all $\bar{c}_{ij} < 0$, then the optimal solution is unique. If at least one $\bar{c}_{ij} < 0$, then we can obtain alternative solution. Assign θ in that cell and repeat one iteration (from step (*iii*)).

Example 4: *Consider the initial BFS by LCM of Example 2, find the optimal solution of the T.P.*

Solution: **Iteration 1**

	M_1	M_2	M_3	M_4	
F_1	3	2	4	[20] 1	$u_1 = -5$
F_2	[15] 2	4	5	3	$u_2 = -1$
F_3	[15] 3	[5] 5	2	[5] 6	$u_3 = 0$ (Let)
F_4	4	[15] 3	[25] 1	4	$u_4 = -2$
	$V_1 = 3$	$V_2 = 5$	$V_3 = 3$	$V_4 = 6$	

For non-basic cells: $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$$\bar{c}_{11} = -5,\ \bar{c}_{12} = -2,\ \bar{c}_{13} = -6,\ \bar{c}_{22} = 0,\ \bar{c}_{23} = -3,\ \bar{c}_{24} = 2,$$

$$\bar{c}_{33} = 1,\ \bar{c}_{41} = -3,\ \bar{c}_{44} = 0.$$

Since all $\bar{c}_{ij}$ are not non-positive, the current solution is not optimal.

Select the cell (2, 4) due to largest positive value and assign an unknown quantity θ in that cell. Identify a loop and subtract and add θ to the corner points of the loop which is as shown below:

3	2	4	[20] 1
[15] –θ			
2	4	5	3
[15] +θ	[5]		[5] –θ
3	5	2	6
	[15]	[25]	
4	3	1	4

Select θ = min. (5, 15) = 5. The cell (3, 4) leaves the basis and the cell (2, 4) enters into the basis. Thus the current solution is updated.

Iteration 2

3	2	4	[20] 1	$u_1 = -2$
[10] 2	4	5	[5] 3	$u_2 = 0$ (Let)
[20] 3	[5] 5	2	6	$u_3 = 1$
4	[15] 3	[25] 1	4	$u_4 = -1$
$V_1 = 2$	$V_2 = 4$	$V_3 = 2$	$V_4 = 3$	

For non-basic cells: $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$$\bar{c}_{11} = -3,\ \bar{c}_{12} = 0,\ \bar{c}_{13} = -4,\ \bar{c}_{22} = 0,\ \bar{c}_{23} = -3,$$

$$\bar{c}_{33} = 1,\ \bar{c}_{34} = -2,\ \bar{c}_{41} = -3,\ \bar{c}_{44} = -2.$$

Since all $\bar{c}_{ij}$ are not non-positive, the current solution is not optimal.

Select the cell (3, 3) due to largest positive value and assign an unknown quantity θ in that cell. Identify a loop and subtract and add θ to the corner points of the loop which is shown below:

3	2	4	[20] 1
[10] 2	4	5	[5] 3
[20] 3	[5] –θ	θ	
	[5]	2	6
4	[15] +θ 3	[25] –θ 1	4

Select θ = min. (5, 25) = 5. The cell (3, 2) leaves the basis and the cell (3, 3) enters into the basis. Thus the current solution is updated.

Iteration 3

3	2	4	[20] 1	$u_1 = -2$
[10] 2	4	5	[5] 3	$u_2 = 0$ (Let)
[20] 3	5	[5] 2	6	$u_3 = 1$
4	[20] 3	[20] 1	4	$u_4 = 0$
$V_1 = 2$	$V_2 = 3$	$V_3 = 1$	$V_4 = 3$	

NOTES

For non-basic cells: $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$$\bar{c}_{11} = -3,\ \bar{c}_{12} = -1,\ \bar{c}_{13} = -5,\ \bar{c}_{22} = -1,\ \bar{c}_{23} = -5,$$

$$\bar{c}_{32} = -1,\ \bar{c}_{34} = -2,\ \bar{c}_{41} = -2,\ \bar{c}_{44} = -1.$$

Since all $\bar{c}_{ij}$ are non-positive, the current solution is optimal. Thus, the optimal solution is

$$x_{14} = 20,\ x_{21} = 10,\ x_{24} = 5,\ x_{31} = 20,\ x_{33} = 5,\ x_{42} = 20,\ x_{43} = 20.$$

The optimal transportation cost

$$= 1 \times 20 + 2 \times 10 + 3 \times 5 + 3 \times 20 + 2 \times 5 + 3 \times 20 + 1 \times 20 = ₹\ 205.$$

Example 5: *Consider the initial BFS by VAM of Example 3, find the optimal solution of the T.P.*

Solution: **Iteration 1**

3	2	4	[20] 1	$u_1 = -2$
[5] 2	[5] 4	5	[5] 3	$u_2 = 0$ (Let)
[25] 3	5	2	6	$u_3 = 1$
4	[15] 3	[25] 1	4	$u_4 = -1$
$V_1 = 2$	$V_2 = 4$	$V_3 = 2$	$V_4 = 3$	

For non-basic cells: $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$$\bar{c}_{11} = -3,\ \bar{c}_{12} = 0,\ \bar{c}_{13} = -4,\ \bar{c}_{23} = -3,\ \bar{c}_{32} = 0,\ \bar{c}_{33} = 1,$$

$$\bar{c}_{34} = -2,\ \bar{c}_{41} = -3,\ \bar{c}_{44} = -2.$$

Since all $\bar{c}_{ij}$ are not non-positive, the current solution is not optimal.

Select the cell (3, 3) due to largest positive value and assign an unknown quantity θ in that cell. Identify a loop and subtract and add θ to the corner points of the loop which is shown below:

3	2	4	[20] 1
[5] +θ	[5] –θ		[5] 3
2	4	5	
[25] –θ		θ	
3	5	2	6
	[15] +θ	[25] –θ	
4	3	1	4

Select θ = min. (5, 25, 25) = 5. The cell (2, 2) leaves the basis and the cell (3, 3) enters into the basis. Thus the current solution is updated.

Iteration 2

3	2	4	[20] 1	$u_1 = -2$
[10] 2	4	5	[5] 3	$u_2 = 0$ (Let)
[20] 3	5	[5] 2	6	$u_3 = 1$
4	[20] 3	[20] 1	4	$u_4 = 0$
$V_1 = 2$	$V_2 = 3$	$V_3 = 1$	$V_4 = 3$	

For non-basic cells: $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$$\bar{c}_{11} = -3,\ \bar{c}_{12} = -1,\ \bar{c}_{13} = -5,\ \bar{c}_{22} = -1,\ \bar{c}_{23} = -5,$$

$$\bar{c}_{32} = -1,\ \bar{c}_{34} = -2,\ \bar{c}_{41} = -2,\ \bar{c}_{44} = -1.$$

Since all $\bar{c}_{ij}$ are non-positive, the current solution is optimal. Thus the optimal solution is

$$x_{14} = 20,\ x_{21} = 10,\ x_{24} = 5,\ x_{31} = 20,\ x_{33} = 5,\ x_{42} = 20,\ x_{43} = 20.$$

The optimal transportation cost = ₹ 205.

Note: To find optimal solution to a T.P., the number of iterations by uv-method is always more if we consider the initial BFS by NWC.

16.4 DEGENERACY IN TRANSPORTATION PROBLEM

A basic feasible solution of a Transportation Problem is said to be degenerate if one or more basic variables assume a zero value. This degeneracy may occur in initial BFS or in the subsequent iterations of *uv*-method. An initial BFS could become degenerate when the supply and demand in the intermediate stages of any one method (NWC/LCM/VAM) are equal corresponding to a selected cell for allocation. In *uv*-method it is identified only when more than one corner points in a loop vanishes due to minimum value of θ.

For the degeneracy in initial BFS, arbitrarily we can delete the row due to supply adjusting the demand to zero or delete the column due to demand adjusting the supply to zero whenever there is a tie in demand and supply.

For the degeneracy in *uv*-method, arbitrarily we can make one corner as non-basic cell and put zero in the other corner.

Example 6: *Find the optimal solution to the following T.P.* :

Source	Destination 1	2	3	Available
1	50	30	190	10
2	80	45	150	30
3	220	180	50	40
Requirement	40	20	20	80

Solution: Let us find the initial BFS using VAM:

	1	2	3	Row Penalties
1	50	30	190	10 (20)
2	80	45	150	30 (35)
3	220	180	50	40 (130)
Column Penalties	40 (30)	20 (15)	20 (100)	

Select (3, 3) cell for allocation and allocate min (40, 20) = 20 in that cell. Cross-off the third column as the requirement is met and adjust the availability to 20. The reduced table is given below:

	1	2	Row Penalties
1	50	30	10 (20)
2	80	45	30 (35)
3	220	180	20 (40)
Column Penalties	40 (30)	20 (15)	

Select (3, 2) cell for allocation. Now there is a tie in allocation. Let us allocate 20 in (3, 2) cell and cross-off the second column and adjust the availability to zero. The reduced table is given below:

	1	
1	50	10
2	80	30
3	220	0
	40	

On continuation we obtain the remaining allocations as 0 in (3, 1) cell, 30 in (2, 1) cell and 10 in (1, 1) cell. The complete initial BFS is given below and let us apply the first iteration of *uv*-method:

Iteration 1

10 50		30	190	$u_1 = -170$
30 80		45	150	$u_2 = -140$
0 220	20	180	20 50	$u_3 = 0$ (Let)
$V_1 = 220$	$V_2 = 180$		$V_3 = 50$	

For non-basic cells: $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$$\bar{c}_{12} = -20,\ \bar{c}_{13} = -310,\ \bar{c}_{22} = -5,\ \bar{c}_{23} = -240.$$

Since all $\bar{c}_{ij} < 0$, the current solution is optimal. Hence, the optimal solution is

$$x_{11} = 10, x_{21} = 30, x_{31} = 0, x_{32} = 20, x_{33} = 20.$$

The transportation cost

$$= 50 \times 10 + 80 \times 30 + 0 + 180 \times 20 + 50 \times 20$$

$$= ₹\ 7500.$$

16.5 MAX-TYPE TRANSPORTATION PROBLEM

Instead of unit cost in transportation table, unit profit is considered then the objective of the Transportation Problem changes to maximize the total profits subject to supply and demand restrictions. Then this problem is called 'max-type' Transportation Problem.

To obtain optimal solution, we consider

$$\text{Loss} = -\text{Profit}$$

and convert the max type transportation matrix to a loss matrix. Then all the methods described in the previous sections can be applied. Thus the optimal BFS obtained for the loss matrix will be the optimal BFS for the max-type Transportation Problem.

Example 7: *A company has three plants at locations A, B and C, which supply to four markets D, E, F and G. Monthly plant capacities are 500, 800 and 900 units respectively. Monthly demands of the markets are 600, 700, 400 and 500 units respectively. Unit profits (in rupees) due to transportation are given below:*

	D	E	F	G
A	8	5	3	6
B	7	4	5	2
C	6	8	4	2

Determine an optimal distribution for the company in order to maximize the total transportation profits.

Solution: The given problem is balanced max type T.P. All profits are converted to losses by multiplying – 1.

	D	E	F	G	
A	–8	–5	–3	–6	500
B	–7	–4	–5	–2	800
C	–6	–8	–4	–2	900
	600	700	400	500	2200

The initial BFS by LCM is given below:

500 –8	–5	–3	–6
100 –7	–4	400 –5	300 –2
–6	700 –8	–4	200 –2

To find optimal solution let us apply *uv*-method.

Iteration 1

500 –θ	–8	–5	–3 θ	–6	$u_1 = -1$
100 +θ	–7	–4	400 –5	300 –θ –2	$u_2 = 0$
–6	700 –8	–4	200 –2		$u_3 = 0$ (Let)
$V_1 = -7$	$V_2 = -8$	$V_3 = -5$	$V_4 = -2$		

For non-basic cells: $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$$\bar{c}_{12} = -4,\ \bar{c}_{13} = -3,\ \bar{c}_{14} = 3,\ \bar{c}_{22} = -4,\ \bar{c}_{31} = -1,\ \bar{c}_{33} = -1.$$

Since all $\bar{c}_{ij}$ are not non-positive, the current solution is not optimal. Select the cell (1, 4) due to largest positive value and assign an unknown quantity θ in that cell. Identify a loop and subtract and add θ to the corner points of the loop which is shown above.

Select θ = min. (500, 300) = 300. The cell (2, 4) leaves the basis and the cell (1, 4) enters into the basis. Thus the current solution is updated.

Iteration 2

200 –θ			300 +θ	$u_1 = -4$
–8	–5	–3	–6	
400		400		$u_2 = -3$
–7	–4	–5	–2	
θ	700		200 –θ	$u_3 = 0$ (Let)
–6	–8	–4	–2	
$V_1 = -4$	$V_2 = -8$	$V_3 = -2$	$V_4 = -2$	

For non-basic cells,

$$\bar{c}_{12} = -7,\ \bar{c}_{13} = -3,\ \bar{c}_{22} = -7,\ \bar{c}_{24} = -3,\ \bar{c}_{31} = 2,\ \bar{c}_{33} = 2.$$

Since all the $\bar{c}_{ij}$ are not non-positive, the current solution is not optimal. There is a tie in largest positive values. Let us select the cell (3, 1) and assign an unknown quantity θ in that cell. Identify a loop and subtract and add θ to the corner points of the loop which is shown above.

Select θ = min. (200, 200) = 200. Since only one cell will leave the basis, let the cell (3, 3) leaves the basis and assign a zero in the cell (1, 1). The cell (3, 1) enters into the basis. Thus the current solution is updated.

Iteration 3

0			500	$u_1 = -2$
–8	–5	–3	–6	
400		400		$u_2 = -1$
–7	–4	–5	–2	
200	700			$u_3 = 0$ (Let)
–6	–8	–4	–2	
$V_1 = -6$	$V_2 = -8$	$V_3 = -4$	$V_4 = -4$	

For non-basic cells,

$$\bar{c}_{12} = -5,\ \bar{c}_{13} = -3,\ \bar{c}_{22} = -5,\ \bar{c}_{24} = -3,\ \bar{c}_{33} = 0,\ \bar{c}_{34} = -4.$$

Since all the $\bar{c}_{ij}$ are non-positive, the current solution is optimal.

Thus the optimal solution, which is degenerate, is

$$x_{11} = 0,\ x_{14} = 500,\ x_{21} = 400,\ x_{23} = 400,\ x_{31} = 200,\ x_{32} = 700.$$

NOTES

The maximum transportation profit

$$=0 + 3000 + 2800 + 2000 + 1200 + 5600 = ₹\ 14600.$$

Since $\bar{c}_{33} = 0$, this indicates that there exists an alternative optimal solution. Assign an unknown quantity θ in the cell (3, 3). Identify a loop and subtract and add θ to the corner points of the loop which is as follow:

0			500
–8	–5	–3	–6
400 +θ		400 –θ	
–7	–4	–5	–2
200	700	θ	
–θ –6	–8	–4	–2

Select θ = min. (200, 400) = 200. The cell (3, 1) leaves the basis and the cell (3, 3) enters into the basis.

Iteration 4

0			500	$u_1 = -2$
–8	–5	–3	–6	
600		200		$u_2 = -1$
–7	–4	–5	–2	
	700	200		$u_3 = 0$ (Let)
–6	–8	–4	–2	
$V_1 = -6$	$V_2 = -8$	$V_3 = -4$	$V_4 = -4$	

For non-basic cells,

$$\bar{c}_{12} = -5,\ \bar{c}_{13} = -3,\ \bar{c}_{22} = -5,\ \bar{c}_{24} = -3,\ \bar{c}_{31} = 0,\ \bar{c}_{34} = -2.$$

Since all the $\bar{c}_{ij}$ are non-positive, the current solution is optimal. Thus the alternative optimal solution is

$$x_{11} = 0,\ x_{14} = 500,\ x_{21} = 600,\ x_{23} = 200,\ x_{32} = 700,\ x_{33} = 200.$$

and the maximum transportation profit is ₹ 14,600.

16.6 UNBALANCED TRANSPORTATION PROBLEM

If total supply ≠ total demand, the problem is called unbalanced Transportation Problem. To obtain feasible solution, the unbalanced problem should be converted to balanced problem by introducing dummy source or dummy destination, whichever is required. Suppose, (supply =) $\sum a_i > \sum b_j$ (= demand). Then add one dummy destination with demand = $\left(\sum a_i - \sum b_j\right)$ with either zero transportation costs or

some penalties, if they are given. Suppose (supply =) $\sum a_i < \sum b_j$ (= demand). Then add one dummy source with supply = $\left(\sum b_j - \sum a_i\right)$ with either zero transportation costs or some penalties, if they are given.

After making it balanced the mathematical formulation is similar to the balanced Transpiration Problem.

Example 8: *A company wants to supply materials from three plants to three new projects. Project I requires 50 truck loads, project II requires 40 truck loads and project III requires 60 truck loads. Supply capacities for the plants P_1, P_2 and P_3 are 30, 55 and 45 truck loads. The table of transportation costs are given below:*

	I	II	III
P_1	7	10	12
P_2	8	12	7
P_3	4	9	10

Determine the optimal distribution.

Solution: Here total supplies = 130 and total requirements = 150. The given problem is unbalanced T.P. To make it balanced consider a dummy plants with supply capacity of 20 truck loads and zero transportation costs to the three projects. Then the balanced T.P. is

		To I	To II	To III	
From	P_1	7	10	12	30
From	P_2	8	12	7	55
From	P_3	4	9	10	45
From	P_4 (Dummy)	0	0	0	20
		50	40	60	

Using VAM, we obtain the initial BFS as given below:

[5] 7	[20] 10	[5] 12
8	12	[55] 7
[45] 4	9	10
0	[20] 0	0

NOTES

To find optimal solution let us apply *uv*-method.

Iteration 1

5	20 +θ	5 –θ	$u_1 = 0$ (Let)
7	10	12	
		55	$u_2 = -5$
8	12	7	
45			$u_3 = -3$
4	9	10	
20		θ	$u_4 = -10$
0 –θ	0	0	
$V_1 = 7$	$V_2 = 10$	$V_3 = 12$	

For non-basic cells, $\bar{c}_{ij} = u_i + v_j - c_{ij}$

$\bar{c}_{21} = -6,\ \bar{c}_{22} = -7,\ \bar{c}_{32} = -2,\ \bar{c}_{33} = -1,\ \bar{c}_{41} = -3,\ \bar{c}_{43} = 2,$

Since $\bar{c}_{43}$ is only positive value assign an unknown quantity θ in (4, 3) cell. Identify a loop and subtract and add θ to the corner points of the loop which is shown in figure.

Select θ = min. (5, 20) = 5 so that the cell (1, 3) leaves the basis and the cell (4, 3) enters into the basis.

Iteration 2

5	25		$u_1 = 0$ (Let)
7	10	12	
		55	$u_2 = -3$
8	12	7	
45			$u_3 = -3$
4	9	10	
	15	5	$u_4 = -10$
0	0	0	
$V_1 = 7$	$V_2 = 10$	$V_3 = 10$	

For non-basic cells, we obtain

$$\bar{c}_{13} = -2,\ \bar{c}_{21} = -4,\ \bar{c}_{22} = -5,\ \bar{c}_{32} = -2,\ \bar{c}_{33} = -3,\ \bar{c}_{41} = -3$$

Since $\bar{c}_{ij} < 0,$ the current solution is optimal. Thus the optimal solution is

Supply 15 truck loads from P_1 to I, 25 truck loads from P_1 to II, 55 truck loads from P_2 to III, 45 truck loads from P_3 to I. Demands of 15 truck loads for II and 5 truck loads for III will remain unsatisfied.

16.7 PERFORMING OPTIMALITY TEST

We have found out a feasible solution. Now, we must find out whether this feasible solution is optimal or not. Such an optimality test can be performed only on such feasible solutions where

1. The number of allocation is $m + n - 1$

where m = number of rows and n = number of columns.

We can test the optimality of a feasible solution by carrying out an examination of each vacant cell to find out whether or not an allocation in that cell reduces the total transportation cost. This can be done by the use of the following two methods:

The Stepping-Stone Method

Let us consider the matrix of the above problem where we have already found out the feasible solution.

		Distribution Centres X	Distribution Centres Y
Plants	A	X_{11} ₹ 2000 [1000]	X_{12} ₹ 5380
	B	X_{21} ₹ 2500 1300 [–100]	X_{11} ₹ 2700 2000 [+100]
	C	X_{23} ₹ 2550 +100	X_{11} ₹ 1700 1200 [–100]

Let us make up any arbitrary empty cell, *i.e.,* CX and allocate +100 units to this cell. Now in order to maintain the restrictions of column X, we must allocate –100 to cell BX and to maintain the row B restriction we must allocate +100 to cell BY. This will result in unbalance of column Y conditions and so we must allot –100 to cell CY.

Now, let us work out the net change in the transportation cost by the changes we have made in allocations.

Evaluation of cell CX

$$= ₹ (2550 \times 100 - 2500 \times 100 + 2700 \times 100 - 1700 \times 100)$$

$$= 255000 - 250000 + 270000 - 170000$$

$$= ₹ 105000$$

As the evaluation of the empty cell CX results in a positive value the total transportation cost cannot be reduced. The feasible solution is an optimal solution already.

We must carryout evaluation of all the empty cells to be sure that optimal solution has been arrived. The total number of empty cells are $m \times n - (m + n - 1) = (m - 1)(n - 1)$. Hence $(m - 1)(n - 1)$ cells must be evaluated. In the present problem $m = 3$ and $n = 2$, so only two empty cells are there but in other problems, the number of empty cells could be much more and this procedure becomes very lengthy and cumbersome.

The Modified Distribution (MODI) Method or UV Method

The problem encountered in the stepping stone method of optimality test can be overcome by MODI method because we don't have to evaluate the empty cells one by one, all of them can be evaluated simultaneously. This is considerably time saving. The method has the following steps:

Step I. Set-up the cost matrix of the problem only with the costs in those cells in which allocations have been made.

	X	Y
A	₹ 2000	
B	₹ 2500	₹ 2700
C		₹ 1700

Step II. Let there be set of number V_j (V_1, V_2) across the top of the matrix and a set of number U; (U_1, U_2, U_3) across the left side so that their sums equal the costs entered in the matrix shown above.

		$V_1 = 0$	200 $V_2 = 200$
2000	U_1	₹ 2000	
2500	U_2	₹ 2500	₹ 2700
1500	U_3		₹ 1700

$U_1 + V_1 = 2000$ $\quad U_2 + V_2 = 2700$

$U_2 + V_1 = 2500$ $\quad U_3 + V_2 = 1700$

Let $V_1 = 0$ then $U_1 = 2000$, $\quad U_2 = 2500$

$V_2 = 2700 - 2500 = 200$

$U_3 = 1500$

Step III. Leave the already filled cells vacant and fill the vacant cells with sums of U_i and V_j. This is shown in the matrix below.

		0	V_1	200	V_2
2000	U_1	...		2200	$(V_1 + V_2)$
2500	U_2	...			...
1500	U_3	1500	$(U_3 + V_1)$		...

Step IV. Subtract the vacant values now filled in step III from the original cost matrix. This will result in cell evaluation matrix and is shown below from the example in hand.

...	5380 – 2200 = 3180
...	...
2550 – 1500 = 1050	...

Step V. If any of the cell evaluation turns out to be negative, then the feasible solution is no optimal. If the values are positive the solution is optimal. In the present example, since both the cell evaluation values are positive, the feasible solution is optimal.

Check Your Progress

Choose the correct option for the following statements:

6. In degenerate transportation problem one or more basic variables assume ..... .
   (a) positive value (b) negative value
   (c) zero value (d) multiple values
7. To obtain optimal solution in Max-type T.P. We consider
   (a) Loss = Profit (b) Loss = – Profit
   (c) Loss > Profit (d) Loss < Profit
8. Transportation problem is said to be unbalanced if
   (a) Total supply = Total demand (b) Total supply ≠ Total demand
   (c) Total supply > Total demand (d) Total supply < Total demand
9. Transportation model is basically ..... programmed.
   (a) computer (b) mathematically
   (c) linearly (d) inversely

16.8 SUMMARY

- Transportation problem is basically linear programmed.
- Transportation problems deal with the determination of minimum cost for transporting one commodity from a number of resources for example, manufacturing units to a number of destinations *e.g.*, clearing and forwarding.
- A necessary and sufficient condition for the existence of a feasible solution to a T.P. is that the T.P. is balanced.

- The number of basic variables in the basic feasible solution of an $m \times n$ T.P. is $m + n - 1$.
- A basic feasible solution of a T.P. is said to be degenerate if one or more basic variables assume a zero value. This degeneracy may occur in initial BFS or in the subsequent iterations of *uv*-method.
- For the degeneracy in *uv*-method, arbitrarily we can make one corner as non-basic cell and put zero in the other corner.
- Instead of unit cost in transportation table, unit profit is considered then the objective of the T.P. changes to maximize the total profits subject to supply and demand restrictions. Then this problem is called 'max-type' T.P.
- If total supply ≠ total demand, the problem is called unbalanced T.P. To obtain feasible solution, the unbalanced problem should be converted to balanced problem by introducing dummy source or dummy destination, whichever is required.
- We can test the optimality of a feasible solution by carrying out an examination of each vacant cell to find out whether or not an allocation in that cell reduces the total transportation cost.

16.9 GLOSSARY

- ***Feasible Solution:*** Non-negative values of x_{ij} where $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$ which satisfy the constraints of availability (supply) and requirement (demand) is called the feasible solution to the transportation problem.
- ***Basic Feasible Solution:*** It is the feasible solution that contains only $m + n - 1$ non-negative allocation

  where m = number of resources

  n = number of destinations
- ***Optimal Solution:*** A feasible solution is said to be optimal solution when the transportation cost is minimum.
- ***Balanced Transportation Problem:*** A transportation problem in which the total supply from all the sources equals the total demand in all the destinations.

  Mathematically, $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$

  where a_i = number of units of the commodity available at source ($i = 1, 2, 3, ..., m$).

  b_j = number of units required at destination ($j = 1, 2, 3, ..., n$)

- ***Unbalanced Transportation Problem:*** Such problem which are not balanced are called unbalanced transportation problems.

  Mathematically $\sum_{i=1}^{m} a_i \neq \sum_{j=1}^{n} b_j$

- ***Degeneracy:*** Basic feasible solution of a transportation problem is said to be degenerate if one or more basic variables assume a zero value.

NOTES

16.10 ANSWERS TO CHECK YOUR PROGRESS

1. solid T.P.
2. $m + n - 1$
3. basic cell, occupied cell
4. Loop
5. basic
6. (b)
7. (b)
8. (b)
9. (c)

16.11 TERMINAL AND MODEL QUESTIONS

1. What is transportation problem? How is it useful in business and industry?
2. Explain the use of transportation problem in business and industry giving suitable examples.
3. What do you understand by
   (a) Feasible solution;
   (b) North-West solution;
   (c) Vogel's Approximation Method (VAM)?
4. Discuss various steps involved in finding initial feasible solution of a transportation problem.
5. Discuss any two methods of solving a transportation problem. State the advantages and disadvantages of these methods.
6. How can an unbalanced transportation problem be balanced? How do you interpret the optimal solution of an unbalanced transportation problem?

7. There are three sources which store a given product. The sources supply these products to four dealers. The capacities of the sources and the demands of the dealers are given. Capacities $S_1 = 150$, $S_2 = 40$, $S_3 = 80$, Demands $D_1 = 90$, $D_2 = 70$, $D_3 = 50$, $D_4 = 60$. The cost matrix is given as follows :

		To D_1	D_2	D_3	D_4
From	S_1	27	23	31	69
	S_2	10	45	40	32
	S_3	30	54	35	57

Find the minimum cost of T.P.

8. There are three factories F_1, F_2, F_3 situated in different areas with supply capacities as 200, 400 and 350 units respectively. The items are shipped to five markets M_1, M_2, M_3, M_4 and M_5 with demands as 150, 120, 230, 200, 250 units respectively. The cost matrix is given as follows:

	M_1	M_2	M_3	M_4	M_5
F_1	2	5	6	4	7
F_2	4	3	5	8	8
F_3	4	6	2	1	5

Determine the optimal shipping cost and shipping patterns.

9. Find the initial basic feasible solution to the following T.P. using (a) NWC, (b) LCM, and (c) VAM:

(i)

		To D	E	F	G	H	
From	A	11	7	5	8	9	50
	B	10	11	8	4	5	90
	C	9	6	12	5	5	60
		20	40	20	40	80	

(ii)

		To A	To B	To C	To D	To E	
From	I	9	10	0	8	9	90
	II	11	12	5	8	3	20
	III	4	9	1	2	0	50
	IV	8	0	3	5	6	50
		80	60	20	40	10	

10. Solve the following transportation problem:

		To D_1	To D_2	To D_3	To D_4	To D_5	
From	S_1	3	5	2	1	3	45
	S_2	2	1	–	4	6	55
	S_3	5	4	3	1	2	65
	S_4	–	4	6	5	7	50
		27	42	51	62	33	

(Supply from S_2 to D_3 and S_4 to D_1 are restricted)

11. A transportation problem for which the costs, origin and availabilities, destinations and requirements are given below:

	D_1	D_2	D_3	
O_1	2	1	2	40
O_2	9	4	7	60
O_3	1	2	9	10
	40	50	20	

Check whether the following basic feasible solution $x_1 = 20$, $x_{13} = 20$, $x_{21} = 10$, $x_{22} = 50$, and $x_{31} = 10$ is optimal. If not, find an optimal solution.

12. Goods have to be transported from sources S_1, S_2 and S_3 to destinations D_1, D_2 and D_3. The T.P. cost per unit capacities of the sources and requirements of the destinations are given in the following table:

	D_1	D_2	D_3	Capacity
S_1	8	5	6	120
S_2	15	10	12	80
S_3	3	9	10	80
Requirement	150	80	50	

Determine a T.P. schedule so that the cost is minimized.

13. Four products are produced in four machines and their profit margins are given by the table as follows:

	P_1	P_2	P_3	P_4	Capacity
M_1	10	7	8	6	40
M_2	5	9	6	4	55
M_3	7	4	11	5	60
M_4	4	10	7	8	45
Requirement	35	42	68	55	

Find a suitable production plan of products in machines so that the profit is maximized while the capacities and requirements are met.

14. Identical products are produced in four factories and sent to four warehouses for delivery to the customers. The costs of transportation, capacities and demands are given as below:

		Warehouses				
		W_1	W_2	W_3	W_4	
Factories	F_1	9	6	11	5	200
	F_2	4	5	8	5	150
	F_3	7	8	4	6	350
	F_4	3	3	10	10	250
Demands		260	100	340	200	Capacities

Find the optimal schedule of delivery for minimization of cost of transportation. Is there any alternative solution? If yes, then find it.

15. Starting with LCM initial BFS, find the optimal solution to the following T.P. problem:

		To					
From		5	1	2	4	3	60
		1	4	2	3	6	55
		4	2	3	5	2	40
		3	5	6	3	7	50
Demands		42	33	41	52	27	Supply

16. A company manufacturing air coolers has two plants located at Mumbai and Kolkata with a weekly capacity of 200 units and 100 units respectively. The company supplies air coolers to its 4 show-rooms situated at Ranchi, Delhi, Lucknow and Kanpur which have a demand of 75, 100, 100 and 30 units respectively. The cost per unit (in ₹.) is shown in the following table:

	Ranchi	**Delhi**	**Lucknow**	**Kanpur**
Mumbai	90	90	100	100
Kolkata	50	70	130	85

Plan the production programmes so as to minimize the total cost of transportation.

16.12 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitative Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics, Laxmi Publications,* New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

NOTES

UNIT 17: ASSIGNMENT PROBLEM

Structure

17.0 INTRODUCTION

In the previous unit, you have learnt transportation problems and their mathematical formulation. In real life situations, problem arise where a number of resources have to be allocated to a number of activities. In a sense, special case of transportation problem is Assignment problem. This model is used when resources have to be assigned to the tasks *i.e.*, assign *n* persons to *n* different types of resources whether human *i.e.*, men or materials, machines etc. have different efficiency of performing different types of jobs and it involves different costs, the problem is how to assign such resources to jobs so that total cost is minimized or given objective is optimized.

In this unit, you will learn various types of assignment problems and their mathematical formulation. How assignment model can be applied in real life situation would also be explained with illustrations.

17.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Describe use of assignment models in industry and business
- Formulate assignment problems mathematically
- Solve assignment problems
- Solve unbalanced assignment problems
- Solve Max-type assignment problems

17.2 MATHEMATICAL FORMULATION OF ASSIGNMENT PROBLEM

Consider n machines $M_1, M_2, ..., M_n$ and n different jobs $J_1, J_2, ..., J_n$. These jobs to be processed by the machines one to one basis *i.e.,* each machine will process exactly one job and each job will be assigned to only one machine. For each job the processing cost depends on the machine to which it is assigned. Now we have to determine the assignment of the jobs to the machines one to one basis such that the total processing cost is minimum. This is called an *assignment problem.*

If the number of machines is equal to the number of jobs then the above problem is called *balanced* or *standard* assignment problem. Otherwise, the problem is called *unbalanced* or *non-standard* assignment problem. Let us consider a balanced assignment problem.

For linear programming problem formulation, let us define the decision variables as

$$x_{ij} = \begin{cases} 1, \text{ if job } j \text{ is assigned to machine } i \\ 0, \text{ otherwise} \end{cases}$$

and the cost of processing job j on machine i as c_{ij}. Then we can formulate the assignment problem as follows :

$$\text{Minimize } z = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} \qquad ...(1)$$

subject to, $$\sum_{j=1}^{n} x_{ij} = 1, \qquad i = 1, 2, ..., n$$

(Each machine is assigned exactly to one job)

$$\sum_{i=1}^{n} x_{ij} = 1, \qquad j = 1, 2, ..., n.$$

(Each job is assigned exactly to one machine)

$$x_{ij} = 0 \text{ or } 1 \text{ for all } i \text{ and } j$$

In **matrix form**,

$$\text{Minimize } z = Cx$$

$$\text{subject to, } Ax = 1,$$

$$x_{ij} = 0 \text{ or } 1,\ i, j = 1, 2, ..., n.$$

where A is a $2n \times n^2$ matrix and total unimodular *i.e.,* the determinant of ever sub-square matrix formed from it has value 0 or 1. This property permits us to replace the constraint $x_{ij} = 0$ or 1 by the constraint $x_{ij} \geq 0$. Thus we obtain

$$\text{Minimize } z = Cx$$

$$\text{subject to, } Ax = 1,\ x \geq 0$$

The **dual** of (1) with the non-negativity restrictions replacing the 0-1 constraintt can be written as follows :

$$\text{Maximize W} = \sum_{i=1}^{n} u_i + \sum_{j=1}^{n} v_j$$

subject to, $u_i + v_j \leq c_{ij}$, $\quad i, j = 1, 2, ..., n.$

u_i, v_j unrestricted in signs $\quad i, j = 1, 2, ..., n.$

Example 1: *A company is facing the problem of assigning four operators to four machines. The assignment cost in rupees in given below :*

		Machine			
		M_1	M_2	M_3	M_4
Operator	I	5	7	–	4
	II	7	5	3	2
	III	9	4	6	–
	IV	7	2	7	6

In the above, operators I and III can not be assigned to the machines M_3 and M_4 respectively. Formulate the above problem as a LP model.

Solution: Let $x_{ij} = \begin{cases} 1, \text{ if the } i\text{th operator is assigned to } j\text{th machine} \\ 0, \text{ otherwise} \end{cases}$

$$i, j = 1, 2, 3, 4.$$

be the decision variables.

By the problem, $x_{13} = 0$ and $x_{34} = 0$.

The LP model is given below :

$$\text{Minimize } z = 5x_{11} + 7x_{12} + 4x_{14} + 7x_{21} + 5x_{22} + 3x_{23} + 2x_{24} + 9x_{31} + 4x_{32} + 6x_{33} + 7x_{41} + 2x_{42} + 7x_{43} + 6x_{44}$$

subject to,

(Operator assignment constraints)

$$x_{11} + x_{12} + x_{14} = 1$$
$$x_{21} + x_{22} + x_{23} + x_{24} = 1$$
$$x_{31} + x_{32} + x_{33} = 1$$
$$x_{41} + x_{42} + x_{43} + x_{44} = 1$$

(Machine assignment constraints)

$$x_{11} + x_{21} + x_{31} + x_{41} = 1$$
$$x_{12} + x_{22} + x_{32} + x_{42} = 1$$
$$x_{23} + x_{33} + x_{43} = 1$$
$$x_{14} + x_{24} + x_{44} = 1$$
$$x_{ij} \geq 0 \text{ for all } i \text{ and } j.$$

17.3 HUNGARIAN ALGORITHM

This is an efficient algorithm for solving the assignment problem developed by the Hungarian mathematician König. Here the optimal assignment is not affected if a constant is added or subtracted from any row or column of the balanced assignment cost matrix. The **algorithm** can be started as follows :

(a) Bring at least one zero to each row and column of the cost matrix by subtracting the minimum of the row and column respectively.

(b) Cover all the zeros in cost matrix by *minimum* number of horizontal and vertical lines.

(c) If number of lines = order of the matrix, then select the zeros as many as the order of the matrix in such a way that they cover all the rows and columns.

(Here $A_{n \times n}$ means *n*th order matrix)

(d) If number of lines ≠ order of the matrix, then perform the following and create a new matrix :

(i) Select the minimum element from the uncovered elements of the cost matrix by the lines.

(ii) Subtract the uncovered elements from the minimum element.

(iii) Add the minimum element to the junction (*i.e.,* crossing of the lines) elements.

NOTES

(iv) Other elements on the lines remain unaltered.

(v) Go to Step (b).

Example 2: *A construction company has four engineers for designing. The general manager is facing the problem of assigning four designing projects to these engineers. It is also found that Engineer 2 is not competent to design project 4. Given the time estimate required by each engineer to design a given project, find an assignment which minimizes the total time.*

		Projects			
		P1	P2	P3	P4
Engineers	E1	6	5	13	2
	E2	8	10	4	–
	E3	10	3	7	3
	E4	9	8	6	2

Solution: Let us first bring zeros rowwise by subtracting the respective minima from all the row elements respectively.

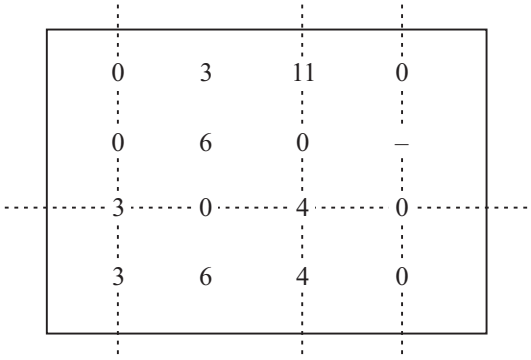
4	3	11	0
4	6	0	–
7	0	4	0
7	6	4	0

Let us bring zero columnwise by subtracting the respective minima from all the column elements respectively. Here the above operations is to be performed only on first column, since at least one zero has appeared in the remaining columns.

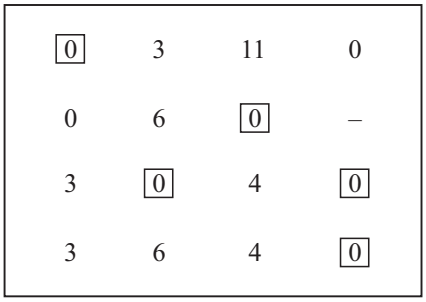
0	3	11	0
0	6	0	–
3	0	4	0
3	6	4	0

(This completes Step-a)

Now (Step-*b*) all the zeros are to be covered by minimum number of horizontal and vertical lines which is shown below. It is also to be noted that this covering is not unique.



It is seen that no. of lines = 4 = order of the matrix. Therefore by Step-*c*, we can go for assignment *i.e.,* we have to select 4 zeros such that they cover all the rows and columns which is shown below:



Therefore the optimal assignment is

E1→P1, E2→P3, E3→P2, E4→P4

and the minimum total time required = 6 + 4 + 3 + 2 = 15 units.

Example 3: *Solve the following job machine assignment problem. Cost data are given below:*

Machines

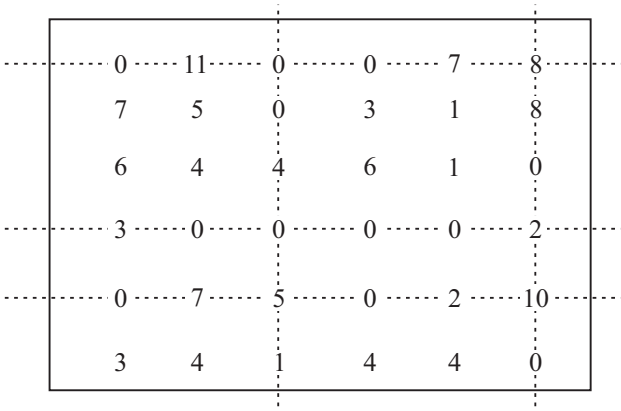
		1	2	3	4	5	6
	A	21	35	20	20	32	28
	B	30	31	22	25	28	30
	C	28	29	25	27	27	21
Jobs	D	30	30	26	26	31	28
	E	21	31	25	20	27	30
	F	25	29	22	25	30	21

NOTES

Solution: Let us first bring zeros first rowwise and then columnwise by subtracting the respective minima elements from each row and each column respectively and the cost matrix, thus obtained, is as follows :

0	11	0	0	7	8
7	5	0	3	1	8
6	4	4	6	1	0
3	0	0	0	0	2
0	7	5	0	2	10
3	4	1	4	4	0

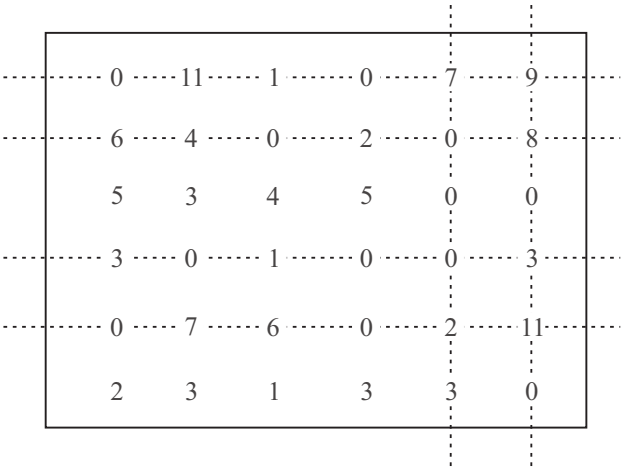
By Step-b, all the zeros are covered by minimum number of horizontal and vertical lines which is shown below:



Here no. of lines ≠ order of the matrix. Hence, we have to apply Step-*d*. The minimum uncovered element is 1. By applying Step-*d* we obtain the following matrix:

0	11	1	0	7	9
6	4	0	2	0	8
5	3	4	5	0	0
3	0	1	0	0	3
0	7	6	0	2	11
2	3	1	3	3	0

Now, by Step-b, we cover all the zeros by minimum number of horizontal and vertical straight lines.



Now, the no. of lines = order of the matrix. So we can go for assignment by Step-*c*. The assignment is shown below:

[0]	11	1	0	7	9
6	4	[0]	2	0	8
5	3	4	5	[0]	0
3	[0]	1	0	0	3
0	7	6	[0]	2	11
2	3	1	3	3	[0]

The optimal assignment is A→1, B→3, C→5, D→2, E→4, F→6. An alternative assignment is also obtained as A→4, B→3, C→5, D→2, E→1, F→6. For both the assignments, the minimum cost is 21 + 22 + 27 + 30 + 20 + 21 *i.e.,* ₹ 141.

Check Your Progress

Fill in the blanks:

1. Assignment problem is special case of .....
2. If the number of machines is equal to the ..... then the problem is called balanced or standard assignment problem.
3. For each job the processing cost depends on the ..... to which it in assigned.
4. ..... is an efficient algorithm for solving the assignment problem developed by König.
5. is not affected if a constant in added from any row or columan.

NOTES

17.4 UNBALANCED ASSIGNMENTS

For unbalanced or non-standard assignment problem no. of rows ≠ no. of columns in the assignment cost matrix *i.e.,* we deal with a rectangular cost matrix. To find an assignment for this type of problem, we have to first convert this unbalanced problem into a balanced problem by adding dummy rows or columns with zero costs so that the defective function will be unaltered. For machine-job problem, if no. of machines (say, m) > no. of jobs (say, n), then create m-n dummy jobs and the processing cost of dummy jobs as zero. When a dummy job gets assigned to a machine, that machine stays idle. Similarly the other case *i.e.,* $n > m$, is handled.

Example 4: *Find an optimal solution to an assignment problem with the following cost matrix:*

	M1	M2	M3	M4	M5
J1	13	5	20	5	6
J2	15	10	16	10	15
J3	6	12	14	10	13
J4	13	11	15	11	15
J5	15	6	16	10	6
J6	6	15	14	5	12

Solution: The above problem is unbalanced. We have to create a dummy machine M6 with zero processing time to make the problem as balanced assignment problem. Therefore we obtain the following:

	M1	M2	M3	M4	M5	M6 **(dummy)**
J1	13	5	20	5	6	0
J2	15	10	16	10	15	0
J3	6	12	14	10	13	0
J4	13	11	15	11	15	0
J5	15	6	16	10	6	0
J6	6	15	14	5	12	0

Let us bring zeros columnwise by subtracting the respective minima elements from each column respectively and the cost matrix, thus obtained, is as follows:

7	0	6	0	0	0
9	5	2	5	9	0
0	7	0	5	7	0
7	6	1	6	9	0
9	1	2	5	0	0
0	10	0	0	6	0

Let us cover all the zeros by minimum number of horizontal and vertical lines and is given below:

7	0	6	0	0	0
9	5	2	5		
0	7	0	5	7	0
7	6	1	6	9	0
9	1	2	5	0	0
0	10	0	0	6	0

Now, the number of lines ≠ order of the matrix. The minimum uncovered element by the lines is 1. Using Step-*d* of the Hungarian algorithm and covering all the zeros by minimum no. of lines we obtain as follows:

7	0	6	0	1	1
8	4	1	4	9	0
0	7	0	5	8	1
6	5	0	5	9	0
8	0	1	4	0	0
0	10	0	0	7	1

NOTES

Now, the number of lines = order of the matrix and we have to select 6 zeros such that they cover all the rows and columns. This is done in the following:

7	[0]	6	0	1	1
8	4	1	4	9	[0]
[0]	7	0	5	8	1
6	5	[0]	5	9	0
8	0	1	4	[0]	0
0	10	0	[0]	7	1

Therefore, the optimal assignment is

J1→M2, J2→M6, J3→M1, J4→M3, J5→M5, J6→M4

and the minimum cost = ₹ (5 + 0 + 6 + 15 + 6 + 5) = ₹ 37.

In the above, the job J2 will not get processed since the machine M6 is dummy.

17.5 MAX-TYPE ASSIGNMENT PROBLEMS

When the objective of the assignment is to maximize, the problem is called 'Max-type assignment problem'. This is solved by converting the profit matrix to an opportunity loss matrix by subtracting each element from the highest element of the profit matrix. Then the minimization of the loss matrix is the same as the maximization of the profit matrix.

Example 5: *A company is faced with the problem of assigning 4 jobs to 5 persons. The expected profit in rupees for each person on each job are as follows:*

	Job			
Persons	***J1***	***J2***	***J3***	***J4***
I	86	78	62	81
II	55	79	65	60
III	72	65	63	80
IV	86	70	65	71
V	72	70	71	60

Find the assignment of persons to jobs that will result in a maximum profit.

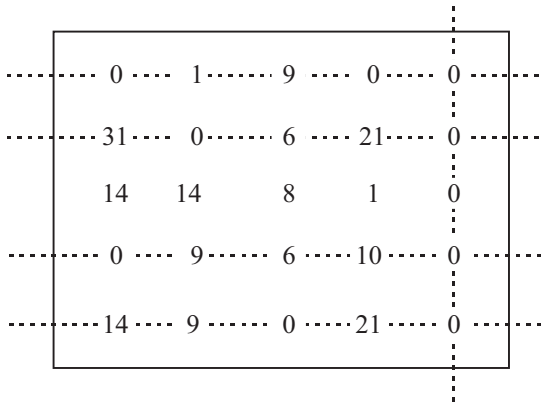
Solution: The above problem is unbalanced Max-type assignment problem. The maximum element is 86. By subtracting all the elements from it obtain the following opportunity loss matrix.

0	8	24	5
31	7	21	26
14	21	23	6
0	16	21	15
14	16	15	26

Now, a dummy job J5 is added with zero losses. Then bring zeros in each column by subtracting the respective minimum element from each column we obtain the following matrix.

0	1	9	0	0
31	0	6	21	0
14	14	8	1	0
0	9	6	10	0
14	9	0	21	0

Let us cover all the zeros by minimum number of lines and is as follows:



Since, the no. of lines = order of the matrix, we have to select 5 zeros such that they cover all the rows and columns. This is done in the following :

0	1	9	[0]	0
31	[0]	6	21	0
14	14	8	1	[0]
[0]	9	6	10	0
14	9	[0]	21	0

NOTES

The optimal assignment is

I→J4, II→J2, III→J5, IV→J1, V→J3 and maximum profit

= ₹ (81 + 79 + 86 + 71) = ₹ 317. Here person III is idle.

Note: The max-type assignment problem can also be converted to a minimization problem by multiplying all the elements of the profit matrix by – 1. Then the Hungarian method can be applied directly.

17.6 ROUTING PROBLEMS

There are various types of routing problems which occurs in a network. The most widely discussed problem is the 'Travelling Salesman Problem (TSP)'. Suppose there is a network of n cities and a salesman wants to make a tour *i.e.,* starting from a city 1 he will visit each of the other $(n - 1)$ cities once and will return to city 1. In this tour the objective is to minimize either the total distance travelled or the cost of travelling by the salesman.

(a) Mathematical Formulation

Let the cities be numbered as 1, 2, ..., n and the distance matrix as follows:

$$D = \begin{array}{c} \text{From} \backslash \text{To} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} 1 \quad 2 \cdots\cdots n \\ \begin{bmatrix} d_{11} & d_{12} \cdots d_{1n} \\ d_{21} & d_{22} \cdots d_{2n} \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ d_{n1} & d_{n2} \cdots d_{nn} \end{bmatrix} \end{array}$$

Generally an infinity symbol is placed in the principal diagonal elements where there is no travelling. So d_{ij} represents the distance from city i to city j $(i \neq j)$. If the cost of travelling is considered then D is referred as cost matrix. It is also to be noted that D may be symmetric in which case the problem is called 'Symmetric TSP' or asymmetric in which case the problem is called 'Asymmetric TSP'.

Let us define the decision variables as follows :

$$x_{ij} = \begin{cases} 1, \text{ if he travels from city } i \text{ to city } j \\ 0, \text{ otherwise} \end{cases}$$

where $i, j = 1, 2, ..., n$

Then the linear programming formulation can be stated as follows :

Minimize $$z = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}\, x_{ij}$$

Subject to, $\sum_{i=1}^{n} x_{ij} = 1, \quad j = 1, 2, ..., n$

$\sum_{i=1}^{n} x_{ij} = 1, \quad i = 1, 2, ..., n$

$x_{ij} = 0$ or 1 for all i and $j = 1, 2, ..., n$

and $x = (x_{ij})$ is a tour.

The above problem has been solved with various approaches *e.g.,* Graph Theoretic Approach, Dynamic Programming, Genetic algorithm etc.

The above problem looks like a special type of Assignment problem. Consider a 4 × 4 assignment problem and a solution as 1 – 4, 2 – 3, 3 – 1, 4 – 2 which can also be viewed as a tour *i.e.,* 1 – 4 – 2 – 3 – 1. If the solution is 1 – 4, 2 – 3, 3 – 2, 4 – 1 then this consists of two sub-tours 1 – 4 – 1, 2 – 3 – 2.

Here one algorithm known as 'Branch and Bound' algorithm is described below:

(b) Branch and Bound Algorithm for TSP

(i) Ignoring tour, solve [D] using Hungarian Algorithm. The transformed matrix is denoted as $[D_0]$. If there is a tour, stop, else go to next step while storing the solution in a node denoted by TSP.

(ii) Calculate the **evaluation** for the variables in $[D_0]$ whose values are zero *i.e.,* $x_{ij} = 0$ where evaluation means the sum of smallest elements of the i-th row and the j-th column excluding the (i, j)th entry.

(iii) Select the variable with highest evaluation, say x_{ij}. If there is a tie, break it arbitrarily. The variable x_{ij} is called the branching variable.

(iv) Create a left branch (TSP1) with $x_{ij} = 0$. To implement this put $d_{ij} = \infty$ in $[D_0]$ *i.e.,* travelling from city i to city j is restricted.

Set [D] = transformed $[D_0]$ and go to step (i).

(v) Create a right branch (TSP2) with $x_{ij} = 1$. This means the salesman must visit city j from city i. To implement this take $[D_0]$ of the parent node. Delete the i-th row and j-th column and put $d_{ji} = \infty$ (to prevent a subtour).

Set [D] = transformed $[D_0]$ and go to step (i).

Note:

(a) There may be a situation arises in step (i) where further solution is not possible then we shall stop that branch.

(b) There may be multiple tours. We shall select the tour with minimum distance or travelling cost.

(c) Calculate total distance (TD) from the given [D] which increases with the level of the tree.

NOTES

NOTES

Example 6: *Solve the following travelling salesman problem using branch and bound algorithm.*

$$D = \begin{array}{c|cccc} \text{From}\backslash\text{To} & 1 & 2 & 3 & 4 \\ 1 & \infty & 3 & 6 & 5 \\ 2 & 3 & \infty & 5 & 8 \\ 3 & 6 & 5 & \infty & 2 \\ 4 & 5 & 8 & 2 & \infty \end{array}$$

Solution: Let us apply the Hungarian Algorithm on [D] and obtain the following matrix :

$D_0 =$

	1	2	3	4
1	∞	(0)	3	2
2	(0)	∞	2	5
3	4	3	∞	(0)
4	3	6	(0)	∞

The solution is 1 – 2, 2 – 1, 3 – 4, 4 – 3. *i.e.,* there exists two subtours 1 – 2 – 1, 3 – 4 – 3. The total distance (TD) = 3 + 3 + 2 + 2 = 10 units.

Then we have to calculate the evaluations for the variables having the value zero in $[D_0]$.

Variable	Evaluation
x_{12}	2 + 3 = 5
x_{21}	2 + 3 = 5
x_{34}	3 + 2 = 5
x_{43}	3 + 2 = 5

Since there are ties in the values, let us select x_{12} as branching variable.

Subproblem TSP1

Let $x_{12} = 0 \Rightarrow$ Put $d_{12} = \infty$ in $[D_0]$ and obtain

$D_0 =$

	1	2	3	4
1	∞	∞	3	2
2	0	∞	2	5
3	4	3	∞	0
4	3	6	0	∞

	1	2	3	4
1	∞	∞	1	⓪
2	⓪	∞	2	5
3	4	⓪	∞	0
4	3	1	⓪	∞

(Apply Hungarian Algorithm)

The solution is 1 – 4, 2 – 1, 3 – 2, 4 – 3 *i.e.,* 1 – 4 – 3 – 2 – 1 which is a tour and TD = 5 + 3 + 5 + 2 = 15 units from [D].

Subproblem TSP2

Let $x_{12} = 1 \Rightarrow$ Delete row 1 and column 2 from $[D_0]$ and put $d_{21} = \infty$ to prevent subtour. The resultant transformed matrix is obtained as follows :

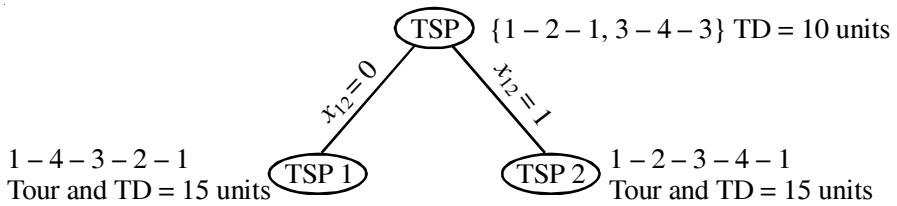
	1	3	4
2	∞	2	5
3	4	∞	0
4	3	0	∞

⇓

	1	3	4
2	∞	⓪	3
3	1	∞	⓪
4	⓪	0	∞

(Applying Hungarian Algorithm)

The solution is 1 – 2, 2 – 3, 3 – 4, 4 – 1 *i.e.,* 1 – 2 – 3 – 4 – 1 which is a tour and TD = 3 + 5 + 2 + 5 = 15 units from [D]. The above calculations is presented in the following tree diagram :



Since, there are two tours with same TD, the given problem has two solutions.

Check Your Progress

Choose the correct option for the following statements:

6. In Max-type assignment problem objective of assignment is to ..... the profit matrix.
   (a) maximize (b) minimize
   (c) equalize (d) add
7. To convert unbalanced problem into a balanced one we add ..... rows or columns with zero costs.
   (a) real (b) infinite
   (c) dummy (d) fixed number of
8. Various types of routing problems occur in a .....
   (a) medium (b) combination
   (c) network (d) loop
9. TSP is
   (a) Total sale problem
   (b) Transportation Sequencing Problem
   (c) Travelling Salesman Problem
   (d) The Sequencing Problem.
10. The assignment schedule will be optimal if there is exactly ..... in each row and each column.
    (a) zero assignment (b) one assignment
    (c) two assignments (d) multiple assignments

17.7 SUMMARY

- This model is used when resources have to be assigned to the tasks *i.e.*, assign *n* persons to *n* different types of resources whether human *i.e.*, men or materials, machines etc. have different efficiency of performing different types of jobs *and it involves different costs*.
- Now we have to determine the assignment of the jobs to the machines one to one basis such that the total processing cost is minimum. This is called an *assignment problem.*
- This is an efficient algorithm for solving the assignment problem developed by the Hungarian mathematician König. Here the optimal assignment is not affected if a constant is added or subtracted from any row or column of the balanced assignment cost matrix.

- When the objective of the assignment is to maximize, the problem is called 'Max-type assignment problem'. This is solved by converting the profit matrix to an opportunity loss matrix by subtracting each element from the highest element of the profit matrix.
- Generally an infinity symbol is placed in the principal diagonal elements where there is no travelling.

17.8 GLOSSARY

- ***Balanced Assignment Problem:*** If the number of machines = number of jobs then the above problem is called balanced or standard assignment problem.
- ***Unbalanced Assignment Problem:*** If number of machines ≠ number of jobs in an assignment problem then this is called unbalanced.
- ***Cost Matrix:*** If the cost of travelling is considered in routing problems then distance matrix D is referred to as cost matrix.
- ***Symmetric:*** **Travelling Salesman Problem (TSP)**. If distance matrix is symmetric then it is called symmetric TSP.
- ***Asymmetric TSP:*** If distance matrix D is asymmetric then TSP is termed as asymmetrix.

17.9 ANSWERS TO CHECK YOUR PROGRESS

1. transportation
2. the number of jobs
3. machine
4. Hungariam alogorithm
5. (a)
6. (c)
7. (c)
8. (c)
9. (c)

17.10 TERMINAL AND MODEL QUESTIONS

NOTES

1. (a) Show that assignment model is a special case of transportation problem.

   (b) A machine tool decides to make six sub-assemblies through six contractors A, B, C, D, E and F. Each contractor is to receive only one sub-assembly from A1, A2, A3, A4, A5 and A6. But the contractors C and E are not competent for the A4 and A2 assembly respectively. The cost of each subassembly by the bids submitted by each contractor is shown below (in hundred rupees):

	A1	A2	A3	A4	A5	A6
A	15	10	11	18	13	22
B	9	12	18	10	14	11
C	9	15	11	–	22	11
D	14	13	9	12	15	10
E	10	–	11	22	13	18
F	10	14	15	12	13	14

   Find the optimal assignments of the assemblies to contractors so as to minimize the total cost.

2. Solve the following assignment problems:

	A	B	C	D	E
I	12	20	20	18	17
II	20	12	5	11	8
III	20	5	12	5	9
IV	18	11	5	12	10
V	17	8	9	10	12

3. (a) If in an assignment problem we add a constant to every element of a row (or column) in the effectiveness matrix, prove that an assignment that minimizes the total effectiveness in one matrix also minimizes the total effectiveness in the other matrix.

   (b) A national car-rental serivce has a surplus of one car in each of the cities 1, 2, 3, 4, 5, 6 and a deficit of one car in each of the cities 7, 8, 9, 10, 11, 12. The distance in miles between cities with a surplus and cities with a deficit are displayed in matrix. How should the cars be despatched so as to minimize the total mileage travelled.

From \ To	7	8	9	10	11	12
1	41	72	39	52	25	51
2	22	29	49	65	81	50
3	27	39	60	51	32	32
4	45	50	48	52	37	43
5	29	40	39	26	30	33
6	82	40	40	60	51	30

4. (a) Distinguish between transportation model and assignment model.

   (b) Four different jobs are to be done on four different machines. The set-up and production times are prohibitively high for change over. Following table indicates the cost of producting job *i* on machine *j* in rupees.

Jobs \ Machines	*1*	*2*	*3*	*4*
1	5	7	11	6
2	8	5	9	6
3	4	7	10	7
4	10	4	8	3

   Assign jobs to different machines so that the total cost is minimized.

5. Five programmers, in a computer centre, write five programmes which run successfully but with different times. Assign the programmers to the programmes in a such a way that the total time taken by them is minimum taking the following time matrix:

Programmers \ Programmes	P1	P2	P3	P4	P5
A	80	66	65	65	73
B	76	75	70	70	75
C	74	73	72	70	66
D	75	75	71	71	73
E	76	66	66	70	75

6. Consider the problem of assigning seven jobs to seven persons. The assignment costs are given as follows:

		Jobs						
		I	II	III	IV	V	VI	VII
	A	9	6	12	11	13	15	11
	B	14	13	14	14	10	20	15
	C	18	6	17	11	15	13	11
Persons	D	10	11	12	15	15	14	13
	E	15	6	18	15	10	14	12
	F	9	18	15	20	14	13	11
	G	14	15	12	13	11	17	20

Determine the optimal assignment schedule.

7. A company wants to assign five salesperson to five different regions to promote a product. The expected sales (in thousand) are given below:

		Regions				
		I	II	III	IV	V
	S1	27	54	37	100	85
	S2	55	66	45	80	32
Salesperson	S3	72	58	74	80	85
	S4	39	88	74	59	72
	S5	72	66	45	69	85

Solve the above assignment problem to find the maximum total expected sale.

8. A company makes profit (₹) while processing different jobs on different machines (one machine to one job only). Now, the company is facing problem of assigning 4 machines to 5 jobs. The profits are estimated as given below:

		Job				
		J1	J2	J3	J4	J5
	A	21	16	35	42	16
	B	15	20	30	35	15
Machine	C	20	16	30	27	18
	D	15	18	32	27	15

Determine the optimal assignment for maximum total profits.

9. There are five operators and six machines in a machine shop. The assignment costs are given in the table below:

		Machine					
		M1	M2	M3	M4	M5	M6
Operator	A	5	–	22	6	8	6
	B	14	9	15	9	14	15
	C	8	12	12	10	8	5
	D	11	13	11	6	9	14
	E	8	9	11	13	–	12

Operator A cannot operate machine M2 and operator E cannot operate machine M5. Find the optimal assignment schedule.

10. A batch of 4 jobs can be assigned to 5 different machines. The setup time for each job on various machines is given below:

		Machine				
		1	2	3	4	5
Jobs	J1	3	9	6	5	6
	J2	4	5	5	7	4
	J3	5	5	3	4	4
	J4	6	8	4	5	5

Find an optimal assignment of jobs to machines which will minimize the total setup time.

17.11 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitative Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 18: QUEUEING THEORY AND DECISION THEORY

Structure

18.0 INTRODUCTION

Queueing systems are prevalent throughout society. The formation of waiting lines is a common phenomenon which occurs whenever the current demand for a service exceeds the current capacity to provide that service. Commuters waiting to board a bus, cars waiting at signals, machines waiting to be serviced by a repairman, letters waiting to be typed by a typist, depositors waiting to deposit the money to a counter in a bank provide some examples of queues. There are applications of queueing theory in several disciplines. A schematic diagram of a queueing system is given below:

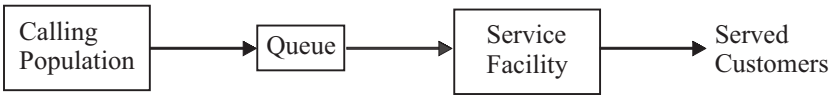


Fig. 18.1: The Basic Queueing System

A queueing system involves a number of servers (or serving facilities) which we will also call *service channels* (in deference to the source field of the theory telephone communication system). The serving channels can be communications links, workstations, check out counters, retailers, elevators, buses, to mention but a few. According to the number of servers, queueing systems can be single and multi-channel type.

Customers arriving at a queuing system at random intervals of time are serviced generally for random times too. When a service is completed, the customer leaves the servicing facility rendering it empty and ready and gets next arrival. The radom nature of arrival and service times may be a cause of congestion at the input to the system at some periods when the incoming customers either queue up for service or leave the system unserviced; in other periods the system might not be completely busy because of the lack of customers, or even be idle altogether.

The subject matter of queueing theory is to build mathematical models, which relate the specified operating conditions for the system (number of channels, their servicing mechanism, distribution of arrivals) to the concerned characteristics of value-measures of effectiveness describing the ability of the system to handle the incoming demands. Depending on the circumstances and the objective of the study, such measures may be: the expected (mean) number of arrivals served per unit time, the expected number of busy channels, the expected number of customers in the queue and the mean waiting time for service, the probability that the number in queue is above some limit, and so on. We do not single out purposely among intended for the given operating conditions, those intended for decision variables, since they may be either or these characteristics, for example, the number of channels, their capacity, service mechanism, etc. The most important part of a study is model establishment (primal problems) while its optimisation (inverse problem) is indeed depending on which parameters are selected to work with or to change. We are not going to consider optimization of queueing models in this text with the exception madel only for the simplest queueing situation.

18.1 UNIT OBJECTIVES

After this unit going through you will be able to:

- Define basic elements of queueing model
- List various notations of queueing theory

- Explain various assumptions and models associated with queueing models based on birth-and-death processes
- Define conenient rotation to denote queueing system
- Explain various non-poisson queueing models
- List various benefits and limitations of queueing theory
- Define decision theory
- Explain decision theory approach

18.2 BASIC ELEMENTS OF QUEUEING MODEL

The basic elements of a queueing model depend on the following factors:

(a) ***Arrival's Distribution:*** Customers arrive and join in the queue according to a probability distribution. The arrival may be single or bulk.

(b) ***Service-time Distribution:*** The service offered by the server also follows a probability distribution. The server(s) may offer single or bulk services *e.g.*, one man barber shop, a computer with parallel processing.

(c) ***Design of Service Facility:*** The services can be offered by the servers in a series, parallel or network stations. A facility comprise a number of series stations through which the customer many pass for service is called 'Tandem queues'. Waiting lines may or may not be allowed between the stations. Similarly parallel queue and network queue are defined.

(d) ***Service Discipline/Queue Discipline:*** There are three types of discipline *e.g.*,

FIFO – First In First Out

LIFO – Last In First Out

SIRO – Service In Random Order

'Stack' is an example of LIFO and selling tickets in a bus is an example of SIRO. Sometimes FIFO is referred as GD (*i.e.,* General Discipline).

Also there is priority service which is two types.

Preemptive. The customers of high priority are given service over the low priority customer.

Non-preemptive. A customer of low priority is served before a high priority customer.

(e) ***Queue Size:*** Generally it is referred as length of the queue or line length. Queue size may be finite or infinite (*i.e.,* a very large queue). Queue size along with the server(s) form the capacity of the system.

(f) ***Calling Population:*** It is also called calling source. Customers join in the queue from a source is known as calling population which may be finite or infinite (*i.e.,* a very large number). To reserve a ticket in a railway reservation counter, customers may come from anywhere of a city. Then the population of the city forms the calling population which can be considered as infinite.

(g) ***Human Behaviour:*** In a queueing system three types of human behaviours are observed.

Jockeying – If one queue is shorter then one join from a larger queue to it.

Balking – If the length of the queue is large, one decides not to enter into it.

Reneging – When a person becomes tired *i.e.,* loses patience on standing on a queue, the person leaves the queue.

18.3 NOTATIONS

P_n = Probability of n customers in a system (steady state)

$P_n(t)$ = Probability of n customers at time t in a system (transient state)

L_s = Expected number of customers in a system

L_q = Expected number of customers in queue

W_s = Expected waiting time in a system (in queue + in service)

W_q = Expected waiting time in queue

λ_n = Mean arrival rate when n customers are in the system.

(If λ_n = Constant for all n, this constant is denoted by λ)

μ_n = Mean service rate for overall system when n customers are in the system.

(If μ_n = Constant for all n, this constant is denoted by μ)

$\rho = \frac{\lambda}{\mu}$ = Traffic intensity/utilization factor

s = Number of servers

$N(t)$ = Number of customers in queueing system at time t.

It has been known that $L_s = \lambda W_s$, $L_q = \lambda W_q$. These relations are called 'Little's formula'. Also we have,

$$W_s = W_q + \frac{1}{\mu}, \text{ for } n \geq 1$$

and $$L_s = L_q + \frac{\lambda}{\mu}$$

NOTES

18.4 KENDALL'S NOTATIONS

A convenient notation to denote queueing system is as follows:

a/*b*/*c*: *d*/*e*/*f*

where

a = Arrivals' distribution

b = Service time distribution

c = Number of servers or channels

d = Service discipline

e = System capacity

f = Calling population.

In the above, *a* and *b* usually take one of the following distribution with its symbol:

M = Exponential distribution for inter-arrival or service time and Poisson arrival.

E_k = Erlangian or gamma distribution with parameter *k*.

D = Constant or deterministic inter-arrival or service time.

G = General distribution (of service time)

Generally *f* is taken as ∞ (infinity) and it is omitted while representing a queue.

Check Your Progress

Fill in the blanks:

1. A queueing system involves a number of servers (or serving facilities) which are also called .....
2. Stack is an example of .....
3. Calling population is also called .....
4. Customers arrive and join in the queue according to .....
5. If the length of queue is large, one decides not to enter into it. It is called.....

18.5 QUEUEING MODELS BASED BIRTH-AND-DEATH PROCESSES

The assumptions of the birth-and-death processes are the following:

(a) Given $N(t) = n$, ($n = 0, 1, 2, ...$), the current probability distribution of the remaining time until the next arrival is exponential with parameter λ_n.

(b) Given N(t) = n, (n = 1, 2, ...), the current probability distribution of the remaining time until the next service completion is exponential with parameter μ_n.

(c) Only one birth or death can occur at a time.

In the queueing models, arrival of a customer implies a birth and departed customer implies a death. In queueing system both arrivals and departures take place simultaneously. This makes difference from birth-and-death process. In the following queueing models mean arrival rate and mean service rate are constant.

M/M/1: FIFO/∞ Model

In this queueing model, arrivals and departures are Poisson with rates λ and μ respectively. There is one server and the capacity of the system is infinity *i.e.,* very large. We shall derive the steady-state probabilities and other characteristics.

Let $P_n(t)$ = Probability of n arrivals during time interval t

If $h > 0$ and small then

$P_n(t + h)$ = P (n arrivals during t and none during h)

or

P ($n - 1$ arrivals during t and one during h)

or P ($n + 1$ arrivals during t and one departure during h)

$$\left\{\text{For Poisson process, } P[X = n] = \frac{(\alpha t)^n . e^{-\alpha t}}{n!}\right\}$$

Now,

P (zero arrival in h) = $e^{-\lambda h} \approx 1 - \lambda$

P (one arrival in h) = $1 - e^{-\lambda h} \approx \lambda h$

Similarly,

P (zero departure in h) = $he^{-\mu h} \approx 1 - \mu h$

P (one departure in h) = $1 - e^{-\mu h} \approx \mu h$.

Then for $n > 0$, we can write

$$P_n(t + h) = P_n(t).(1 - \lambda h)(1 - \mu h) + P_{n-1}(t)(\lambda h)(1 - \mu h) + P_{n+1}(t)(1 - \lambda h).(\mu h)$$

$$\Rightarrow \quad \frac{P_n(t + h) - P_n(t)}{h} = \lambda P_{n-1}(t) + \mu P_{n+1}(t) - (\lambda + \mu) P_n(t).$$

Taking $h \to 0$, we obtain

$$P'_n(t) = \lambda P_{n-1}(t) + \mu P_{n+1}(t) - (\lambda + \mu) P_n(t) \qquad ...(1)$$

For $n = 0$, we can write

$$P_0(t + h) = P_0(t).(1 - \lambda h).1 + P_1(t) (1 - \lambda h).(\mu h).$$

$$\Rightarrow \quad \frac{P_0(t+h) - P_0(t)}{h} = -\lambda P_0(t) + \mu P_1(t)$$

Taking $h \to 0$, we obtain

NOTES

$$P'_0(t) = -\lambda P_0(t) + \mu P_1(t) \qquad ...(2)$$

These are called difference-differential equations.

The solution of (1) and (2) will give the transient-state probabilities, $P_n(t)$. But the solution procedure is complex. So with certain assumptions we shall obtain the steady state solution.

For steady state, let us consider

$$t \to \infty, \lambda < \mu,$$

$$P'_n(t) \to 0, P_n(t) \to P_n \text{ for } n = 0, 1, 2, ...$$

(Here $\lambda = \mu \Rightarrow$ No queue and $\lambda > \mu \Rightarrow$ explosive state).

From (1) and (2) we obtain,

$$\lambda P_{n-1} + \mu P_{n+1} - (\lambda + \mu) P_n = 0, \qquad n > 0 \qquad ...(3)$$

$$-\lambda P_0 + \mu P_1 = 0, \qquad n = 0 \qquad ...(4)$$

From (4), $$P_1 = \frac{\lambda}{\mu} . P_0.$$

From (3), for $n = 1$,

$$P_2 = \left(\frac{\lambda + \mu}{\mu}\right) P_1 - \frac{\lambda}{\mu} P_0$$

$$= \left(\frac{\lambda + \mu}{\mu}\right)\left(\frac{\lambda}{\mu}\right) P_0 - \left(\frac{\lambda}{\mu}\right) P_0$$

$$= \left(\frac{\lambda}{\mu}\right)^2 P_0$$

Similarly, for $n = 2, 3, ...$

$$P_3 = \left(\frac{\lambda}{\mu}\right)^3 P_0$$

.....

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

.....

Also, we have $$\sum_{n=0}^{\infty} P_n = 1$$

NOTES

$$\Rightarrow \quad P_0 \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = 1$$

$$\Rightarrow \quad P_0 \left[1 - \frac{\lambda}{\mu}\right]^{-1} = 1, \ \frac{\lambda}{\mu} < 1$$

$$\Rightarrow \quad P_0 = \frac{1}{1 - \frac{\lambda}{\mu}} = \frac{1}{1-\rho}, \ \rho < 1.$$

Therefore, $$P_n = P[X = n] = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

$$= P^n (1 - \rho), \ \rho < 1, \ n \geq 0.$$

Now L_s = Expected number of customers in the system.

$$= \sum_{n=0}^{\infty} n.P_n \qquad = \sum_{n=0}^{\infty} n.\rho^n (1-\rho)$$

$$= (1-\rho)\rho \sum_{n=0}^{\infty} n.\rho^{n-1} = (1-\rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right)$$

$$= (1-\rho).\rho.\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) \qquad (\because \rho < 1)$$

$$= (1-\rho).\rho.\frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

W_s = Expected waiting time in the system

$$= \frac{L_s}{\lambda} \text{ (By Little's formula)} = \frac{1}{\mu - \lambda}$$

L_q = Average queue length

$$= L_s - \frac{\lambda}{\mu} = \frac{\rho^2}{1-\rho}$$

W_q = Expected waiting time in queue

$$= \frac{L_q}{\lambda} \text{ (By Little's formula)} = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu - \lambda)}.$$

P (at least *n* customers in the system)

$$= P \text{ (queue size} \geq n)$$

$$= \sum_{j=n}^{\infty} P_j = \sum_{j=n}^{\infty} \rho^j (1-\rho)$$

NOTES

$$= (1-\rho)\rho^n \sum_{j=n}^{\infty} \rho^{j-n} = (1-\rho)\rho^n \sum_{k=0}^{\infty} \rho^k \text{ (let } k = j - n)$$

$$= (1-\rho)\rho^n . \frac{1}{1-\rho} = \rho^n.$$

Let m = No. of customers in the queue

$$P[m>0] = P[n>1] = 1 - P[n \le 1]$$

$$= 1 - \{P[n=0] + P[n=1]\}$$

$$= 1 - \{(1-\rho) + \rho(1-\rho)\} = \rho^2.$$

Therefore, **Average length of non-empty queue**

$$= E[m|m>0]$$

$$= \frac{E(m)}{P[m>0]} = \frac{L_q}{P[m>0]} = \frac{\rho^2}{1-\rho}, \frac{1}{\rho^2} = \frac{1}{1-\rho}.$$

Variance of system length/Fluctuation of queue

$$= \sum_{n=0}^{\infty} [n - L_s]^2 P_n = \sum_{n=0}^{\infty} n^2 . P_n - [L_s]^2$$

$$= \frac{\rho}{(1-\rho)^2}, \text{ after simplification.}$$

(a) Waiting Time Distributions

Let the time spent by a customer in the system be given as follows

$$T_s = t'_1 + t_2 + ... + t_n + t_{n+1}$$

where t'_1 is the additional time taken by the customer in service, t_2, ..., t_n are the service times of other customers ahead of him and t_{n+1} is the service time of arriving customer. Here T_s is the sum of $(n + 1)$ independently identically exponentially distributed random variables and follows a gamma distribution with parameters μ and $n + 1$. The conditional *pdf* $w(t \mid n + 1)$ of T_s is given by

$$w(t \mid n+1) = \frac{\mu}{n!}(\mu t)^n . e^{-\mu t}, \; t > 0.$$

Then the *pdf* of T_s is obtained by first multiplying the expression $w(t \mid n + 1)$ with the probability that there are n customers in the system and then summing over-all values of n from 0 to ∞ and is given below:

$$pdf \text{ of } T_s = (u - \lambda)\, e^{-(\mu - \lambda)t}, \; t > 0$$

which is an exponential distribution with parameter $(\mu - \lambda)$. We can also compute the *pdf* T_q of waiting time of an incoming customer before he receives the service following a similar line of argument. Thus

$$pdf \text{ of } T_q = \begin{cases} \rho(\mu - \lambda) e^{-(\mu-\lambda)t}, & t > 0 \\ 1 - \rho & t = 0 \end{cases}$$

The second component means the customer starts receiving service immediately after the arrival if there is no customer in the system.

Also we can obtain $W_s = \int_0^\infty t.T_s \, dt = \frac{1}{\mu - \lambda}$, $W_q = \int_0^\infty t.T_q \, dt = \frac{\lambda}{\mu(\mu - \lambda)}$.

Example 1: *At a public telephone booth arrivals are considered to be Poisson with an average inter-arrival time of 10 minutes. The length of a phone call may be treated as service, assumed to be distributed exponentially with mean = 2.5 minutes. Calculate the following:*

(*a*) *Average number of customers in the booth.*

(*b*) *Probability that a fresh arrival will have to wait for a phone call.*

(*c*) *Probability that a customer completes the phone call in less than 10 minutes and leave.*

(*d*) *Probability that queue size exceeds at least 5.*

Solution: Here $\lambda = \frac{1}{10}$ customers per minute.

$$\mu = \frac{1}{2.5} \text{ customers per minute.}$$

and $$\rho = \frac{\lambda}{\mu} = \frac{2.5}{10} = 0.25 < 1$$

(a) Average number of customers in the booth

$$= L_s = \frac{\rho}{1 - \rho} = \frac{0.25}{1 - 0.25} = 0.33$$

(b) P (a fresh arrival will have to wait)

$$= 1 - P \text{ (a fresh arrival will not have to wait)}$$

$$= 1 - P \text{ (no customers in the booth)}$$

$$= 1 - P[X = 0]$$

$$= 1 - (1 - \rho) = \rho = 0.25$$

NOTES

(c) P (phone call completes in less than 10 min.)

$$= P\left[T_s < 10\right]$$

$$= \int_0^{10} (\mu - \lambda)\, e^{-(\mu-\lambda)t}\, dt = 1 - e^{-(\mu-\lambda).10} = 0.95$$

Example 2: *At a one-man barber shop, customers arrive according to the Poisson distribution with a mean arrival rate of 4 per hour and his hair cutting time was exponentially distributed with an average hair cut taking 12 minutes. There is no restriction in queue length. Calculate the following:*

(*a*) *Expected time in minutes that a customer has to spend in the queue.*

(*b*) *Fluctuations of the queue length.*

(*c*) *Probability that there is at least 5 customers in the system.*

(*d*) *Percentage of time the barber is idle in 8-hr. day.*

Solution: $\lambda =$ 4 per hour $= \frac{1}{15}$ per minute.

$$\mu = \frac{1}{12} \text{ per minute.}$$

$$\rho = \frac{\lambda}{\mu} = \frac{12}{15} = 0.8 < 1$$

(a) $$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{1/15}{\frac{1}{12}\left(\frac{1}{12} - \frac{1}{15}\right)} = 48 \text{ minutes}$$

(b) Fluctuations of queue length $= \frac{\rho}{(1-\rho)^2} = \frac{0.8}{(0.2)^2} = 20$

(c) P (at least 5 customers in the system) $= \rho^5 = (0.8)^5 = 0.33$

(d) P (barber is idle) $= P$ (no customers in the shop)

$$= 1 - \rho = 1 - 0.8 = 0.2$$

Percentage of time barber is idle $= 8 \times .2 = 1.6$.

M/M/1: FIFO/N Model

In this model the system capacity is restricted to N. Therefore, (N + 1)th customer will not join and the difference-differential equations of the previous model are valid if $n < N$. Then for $n = N$, we have

$$P_N(t + h) = P_N(t)\,(1 - \mu h) + P_{N-1}(t)(\lambda h)(1 - \mu h)$$

On simplification, the additional difference-differential equation is obtained as

$$P'_N(t) = -\mu P_N(t) + \lambda P_{N-1}(t).$$

Under steady-state this equation reduces to

$$0 = -\mu P_N + \lambda P_{N-1}.$$

Hence, we have three difference equations

$$\lambda P_{n+1} = (\lambda + \mu)P_n - \lambda P_{n-1},\ 1 \le n \le N-1$$

$$\mu P_1 = \lambda P_0,\ n = 0$$

and

$$\mu P_N = \lambda P_{N-1},\ n = N.$$

As before, the first two equations give

$$P_n = \left(\frac{\lambda}{\mu}\right)^n . P_0,\ n \le N-1$$

$$\Rightarrow \quad P_n = \rho^n . P_0.$$

This equation satisfies the third difference equation for $n = N$.

To determine P_0, we use, $\sum_{n=0}^{N} P_n = 1$.

$$\Rightarrow \quad 1 = P_0 \cdot \sum_{n=0}^{N} \rho^n = \begin{cases} P_0 \cdot \left(\dfrac{1-\rho^{n+1}}{1-\rho}\right), & \rho \neq 1 \\ P_0\,(N+1), & \rho = 1 \end{cases}$$

$$\Rightarrow \quad P_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{N+1}}, & \rho \neq 1 \\ \dfrac{1}{N+1}, & \rho = 1 \end{cases}$$

Hence

$$P_n = \begin{cases} \dfrac{(1-\rho)\,\rho^n}{1-\rho^{N+1}}, & \rho \neq 1 \\ \dfrac{1}{N+1}, & \rho = 1 \end{cases}.$$

So in this model, ρ can be > 1 or < 1.

$$L_s = \sum_{n=0}^{N} n . P_n$$

For $\rho \neq 1$,

$$L_s = P_0 \cdot \sum_{n=0}^{N} n . \rho^n$$

NOTES

$$= P_0.\rho \sum_{n=0}^{N} n.\rho^{n-1} = P_0.\rho \frac{d}{d\rho}\left(\sum_{n=0}^{N}\rho^n\right)$$

$$= P_0.\rho \frac{d}{d\rho}\left(\frac{1-\rho^{N+1}}{1-\rho}\right)$$

$$= P_0.\rho .\frac{(1-\rho)\left(-(N+1)\rho^N\right)-(1-\rho^{N+1})(-1)}{(1-\rho)^2}$$

$$= P_0.\rho .\frac{1-(N+1)\rho^N+N\rho^{N+1}}{(1-\rho)^2}$$

$$= \left(\frac{1-\rho}{1-\rho^{N+1}}\right).\rho .\frac{1-(N+1)\rho^N+N\rho^{N+1}}{(1-\rho)^2}$$

$$= \frac{\rho\left[1-(N+1)\rho^N+N\rho^{N+1}\right]}{(1-\rho)(1-\rho^{N+1})}$$

For $\rho = 1$,

$$L_s = \sum_{n=0}^{N} nP_n = \sum_{n=0}^{N} n.\frac{1}{(N+1)} = \frac{1}{N+1}.\sum_{n=0}^{N} n$$

$$= \frac{1}{N+1}.\frac{N(N+1)}{2} = \frac{N}{2}.$$

Thus

$$L_s = \begin{cases} \dfrac{\rho(1-(N+1)\rho^N+N\rho^{N+1})}{(1-\rho)\ (1-\rho^{N+1})}, & \rho \neq 1 \\ \dfrac{N}{2}, & \rho = 1 \end{cases}$$

Let $\bar{\lambda}$ = Effective arrival rate.

Here $\lambda_n = 0$ for $n \geq N$

and $\sum_{n=0}^{N} P_n = 1$

$\Rightarrow$ $P_N + \sum_{n=0}^{N-1} P_n = 1$

$$\Rightarrow \qquad \sum_{n=0}^{N-1} P_n = 1 - P_N.$$

NOTES

Therefore, $$\bar{\lambda} = \sum_{n=0}^{N-1} \lambda_n . P_n = \lambda . \sum_{n=0}^{N-1} P_n \qquad (\because \lambda_n = \lambda)$$

$$= \lambda(1 - P_N).$$

The **other measures** are obtained as follows:

$$L_s = L_q + \frac{\bar{\lambda}}{\mu} \Rightarrow L_q = L_s - \frac{\bar{\lambda}}{\mu}$$

$$W_q = \frac{L_q}{\bar{\lambda}}$$

$$W_s = W_q + \frac{1}{\mu}.$$

All will give two values *i.e.,* one for $\rho \neq 1$ and the other for $\rho = 1$, *e.g.,*

$$L_q = \begin{cases} \dfrac{\rho^2\left[1 - N\rho^{N-1} + (N-1)\rho^N\right]}{(1-\rho)(1-\rho^{N+1})}, & \rho \neq 1 \\ \dfrac{N(N-1)}{2(N+1)}, & \rho = 1. \end{cases}$$

Example 3: *Assume that the trucks with goods are coming in a market yard at the rate of 30 trucks per day and suppose that the inter-arrival times follows an exponential distribution. The time to unload the trucks is assumed to be exponential with an average of 42 minutes. If the market yard can admit 10 trucks at a time, calculate P (the yard is empty) and find the average queue length.*

If the unload time increases to 48 minutes, then again calculate the above two questions.

Solution: Here $\lambda = \dfrac{30}{60 \times 24} = \dfrac{1}{48}$ trucks per minute.

and $\mu = \dfrac{1}{42}$ trucks per minute.

$$N = 10, \rho = \frac{\lambda}{\mu} = \frac{42}{48} = 0.875$$

P (yard is empty) = P (no trucks in the yard)

NOTES

$$= P_0 = \frac{1-\rho}{1-\rho^{N+1}} = \frac{1-0.875}{1-(0.875)^{11}}$$

$$= \frac{0.125}{0.7698} = 0.16$$

$$\text{Average queue length} = \frac{\rho^2\left[1 - N\rho^{N-1} + (N-1)\rho^N\right]}{(1-\rho)(1-\rho^{N+1})}$$

$$= \frac{(0.875)^2\left[1 - 10.(0.875)^9 + 9.(0.875)^{10}\right]}{(1-0.875)\left(1-(0.875)^{11}\right)}$$

$$= \frac{0.2765}{0.0962} = 2.87.$$

Next part: $\mu = \frac{1}{48}$ trucks per minute

and $\rho = \frac{\lambda}{\mu} = 1$

$$\text{P (yard is empty)} = P_0 = \frac{1}{N+1} = \frac{1}{11} = 0.09.$$

$$\text{Average queue length} = \frac{N(N-1)}{2(N+1)} = \frac{10 \times 9}{2 \times 11} = 4.09.$$

Example 4: *Cars arrive in a pollution testing centre according to poisson distribution at an average rate of 15 cars per hour. The testing centre can accommodate at maximum 15 cars. The service time (i.e., testing time) per car is an exponential distribution with mean rate 10 per hour.*

(*a*) *Find the effective arrival rate at the pollution testing centre.*

(*b*) *What is the probability that an arriving car has not to wait for testing.*

(*c*) *What is the probability that an arriving car will find a vaccant place in the testing centre.*

(*d*) *What is the expected waiting time until a car is left from the testing centre.*

Solution: λ = 15 cars/hour

μ = 10 cars/hour

$$\rho = \frac{\lambda}{\mu} = 1.5,\ N = 15$$

NOTES

(a) $\bar{\lambda} = \lambda\,(1 - P_N) = 15(1 - P_{15}) = 15(1 - 0.333) = 10$ cars/hour.

(b) P (arriving car has not to wait for testing)

$$= P_0 = \frac{1 - 1.5}{1 - (1.5)^{16}} = 0.00076.$$

(c) $P_0 + P_1 + ... + P_{14} = 1 - P_{15} = 1 - 0.333 = 0.667.$

(d) $$L_s = \frac{\rho\left[1 - (N+1)\,\rho^N + N\rho^{N+1}\right]}{(1-\rho)\,(1-\rho^{N+1})} = 13.016$$

$$\therefore \quad W_s = \frac{L_s}{\lambda} = 1.301 \text{ hours}.$$

M/M/S: FIFO/● Model

In this model, the arrival rate of the customers is λ, but maximum of *s* customers can be served simultaneously.

If μ be the average number of services per unit time per server, then we have

$$\lambda_n = \lambda,\ n = 0, 1, 2, ...$$

$$\mu_n = \begin{cases} n\mu, & 0 \le n \le s \\ s\mu, & n \ge s \end{cases}$$

When $n < s$, there is no queue $\rho = \frac{\lambda}{\mu}$.

In this model the condition of existence of steady state solution is $\frac{\rho}{s} < 1$.

The **steady-state probabilities** are obtained as

$$P_n = \begin{cases} \frac{\rho^n}{n!}.P_0, & 0 \le n \le s \\ \frac{\rho^n}{s^{n-s}\,s!}.P_0, & n \ge s \end{cases}$$

where $$P_0 = \left[\sum_{n=0}^{s-1}\frac{\rho^n}{n!} + \frac{\rho^s}{s!\,(1-\rho/s)}\right]^{-1}$$

NOTES

$$L_q = \sum_{n=s}^{\infty}(n-s)P_n$$

$$= \sum_{n=s}^{\infty}(n-s)\frac{\rho^n}{s^{n-s}.s!}.P_0 = \frac{\rho^s}{s!}.P_0.\sum_{i=0}^{\infty} i.\left(\frac{\rho}{s}\right)^i \qquad (\text{let } i = n-s)$$

$$= \frac{\rho^{s+1}}{s.s!}.P_0.\frac{d}{du}\left(\sum_{i=0}^{\infty}u^i\right) \qquad (\text{let } u = \rho/s)$$

$$= \frac{\rho^{s+1}}{s.s!}.P_0.\frac{d}{du}\left(\frac{1}{1-u}\right) = \frac{\rho^{s+1}}{s.s!}P_0.\frac{1}{(1-u)^2}$$

$$= \frac{\rho^{s+1}}{s.s!}.\frac{1}{\left(1-\frac{\rho}{s}\right)^2}P_0 = \frac{\rho^{s+1}}{(s-1)!.(s-\rho)^2}.P_0.$$

The **other measures** are obtained as follows:

$$L_s = L_q + \rho,\ W_q = L_q/\lambda,\ W_s = W_q + \frac{1}{\mu}.$$

Expected number of customers in the service = ρ

Expected time for which a server is busy = $\frac{\rho}{s}$

Expected time for which a server is idle = $1 - \frac{\rho}{s}$

P (all servers are busy) = $P_s + P_{s+1} + ...$

$$= \sum_{n=s}^{\infty}\frac{\rho^n}{s!.s^{n-s}}.P_0 = \frac{\rho^s}{s!}\left(1-\frac{\rho}{s}\right)^{-1}.P_0.$$

P (an arrival has to wait) = P (all servers are busy).

Example 5: *A post office has two counters, which handles the business of money orders, registration letters etc. It has been found that the service time distributions for both the counters are exponential with mean service time of 4 minutes per customer. The customers are found to come in each counter in a Poisson fashion with mean arrival rate of 11 per hour. Calculate*

(*a*) *Probability of having to wait for service of a customer.*

(*b*) *Average waiting time in the queue.*

(*c*) *Expected number of idle counters.*

Solution: Here $\lambda = 11$ customers/hour.

$$\mu = \frac{60}{4} = 15 \text{ customers/hour.}$$

$$s = 2.$$

$$\rho = \frac{\lambda}{\mu} = \frac{11}{15}, \frac{\rho}{s} = \frac{11}{30} < 1.$$

$$P_0 = \left[\sum_{n=0}^{1} \frac{\rho^n}{n!} + \frac{\rho^2}{2!.(1-\rho/2)}\right]^{-1}, \; P_1 = \rho.P_0 = 0.34$$

$$= \left[1 + \frac{11}{15} + \frac{(11/15)^2}{2.(1-11/30)}\right]^{-1} = 0.463.$$

(a) P (an arrival has to wait) = P (all servers are busy)

$$= \left(\frac{11}{15}\right)^2 . \frac{1}{2!}\left(1 - \frac{11}{30}\right)^{-1} .(0.463) = 0.197.$$

(b) $$L_q = \frac{\rho^{s+1}}{(s-1)!.(s-\rho)^2}.P_0$$

$$= \frac{(11/15)^3}{(2-11/15)^2}.(0.463) = 0.114.$$

$\therefore$ Average waiting time in queue $(W_q) = \frac{L_q}{\lambda} = \frac{0.114}{11} = 0.01$ hour $= 0.62$ min.

(c) Expected number of idle counters

$$= 2.P_0 + 1.P_1 = 1.266.$$

M/M/S: FIFO/N, S ≤ N Model

This is called *s*-server model with finite system capacity.

$$\text{Arrival rate } \lambda_n = \begin{cases} \lambda, & 0 \le n < N \\ 0, & n \ge N \end{cases}$$

$$\text{Service rate } \mu_n = \begin{cases} n\mu, & 0 \le n \le s \\ s\mu, & s \le n \le N \end{cases}$$

$$\rho = \frac{\lambda}{\mu}.$$

The **steady-state probability** are given as

$$P_n = \begin{cases} \frac{\rho^n}{n!}.P_0 & 0 \le n \le s \\ \frac{\rho^n}{s!.s^{n-s}}.P_0 & s \le n \le N \end{cases}$$

where $$P_0 = \left[\sum_{i=0}^{s-1}\frac{\rho^i}{i!} + \sum_{i=s}^{N}\frac{\rho^i}{s!.s^{i-s}}\right]^{-1}$$

The **other characteristics** of this model are given below:

$$L_q = \sum_{n=s}^{N}(n-s).P_n = P_0.\frac{\rho^{s+1}}{(s-1)!.(s-\rho)^2}\left\{1 - x^{N-S} - (N-s)x^{N-S}(1-x)\right\}$$

where $$x = \frac{\rho}{s}$$

Expected number of idle servers $\bar{s} = \sum_{k=0}^{s}(s-k).P_k$, k = arrival

$$\bar{\lambda} = \lambda(1 - P_N) = \mu(s - \bar{s})$$

$$W_q = \frac{L_q}{\bar{\lambda}}$$

$$W_s = W_q + \frac{1}{\mu}$$

$$L_s = L_q + (s - \bar{s}).$$

Expected number of busy servers = Expected number of customers in service

$$= s - \bar{s} = \frac{\bar{\lambda}}{\mu}$$

Proportion of busy time for a server $= \frac{s - \bar{s}}{s}$.

Special Cases: (I) **M/M/S: FIFO/S**

In this model s = N, the steady state probabilities are given by

$$P_n = \frac{\rho^n}{n!}.P_0, \quad 0 \le n \le s$$

$$= 0, \quad n > s$$

where $$P_0 = \left[\sum_{n=0}^{s}\left(\frac{\rho^n}{n!}\right)\right]^{-1}$$

$$L_s = \sum_{n=1}^{s} n.P_n = P_0 \sum_{n=1}^{s} \frac{\rho^n}{(n-1)!}, \quad L_q = W_q = 0.$$

(II) M/M/∞: FIFO/∞ (Self-service queueing model)

In this model a customer joining the system becomes a server. So this is called self-service system. The steady state probabilities are given by

$$P_n = \frac{e^{-\rho}.\rho^n}{n!}, n = 0, 1, 2, \ldots. \text{ (Poisson distribution)}$$

$$\bar{\lambda} = \lambda, \; L_s = \rho, \; W_s = \frac{L_s}{\lambda} = \frac{1}{\mu}$$

$$L_q = 0, \; W_q = 0.$$

Example 6: *A barber shop has two barbers and four chairs for customers. Assume that customers arrive in a Poisson fashion at a rate of 4 per hour and that each barber services customers according to an exponential distribution with mean of 18 minutes. Further, if a customer arrives and there are no empty chairs in the shop he will leave. Calculate the following:*

(*a*) *Probability that the shop is empty.*

(*b*) *Effective arrival rate.*

(*c*) *Expected number of busy servers.*

(*d*) *Expected number of customers in queue.*

Solution: Here, $s = 2$, $N = 4$, $\lambda = \frac{4}{60} = \frac{1}{15}$ customers per minute

$$\mu = \frac{1}{18} \text{ customers per minute.}$$

$$\rho = \frac{\lambda}{\mu} = \frac{18}{15} = 1.2, \; \frac{\rho}{s} = 0.6$$

$$P_0 = \left[1 + \rho + \sum_{i=2}^{4} \frac{\rho^i}{2!.(2)^{i-2}}\right]^{-1} = [3.6112]^{-1} = 0.28$$

$$P_1 = \rho.P_0 = 0.34, \; P_4 = \frac{\rho^4.P_0}{2!.2^2} = 0.07, \; P_3 = 0.12$$

NOTES

(a) P (shop is empty) = $P_0 = 0.28$

(b) $$\bar{\lambda} = \lambda(1 - P_4) = \frac{1}{15}(1 - 0.07) = 0.062$$

(c) Expected no. of busy servers = $\frac{\bar{\lambda}}{\mu} = 18 \times 0.062 = 1.116$

(d) $$L_q = \sum_{n=2}^{4}(n-2).P_n = 1.P_3 + 2.P_4$$

$$= 0.12 + 2.(0.07) = 0.26.$$

18.6 NON-POISSON QUEUEING MODELS

In these queueing models, either arriving time distribution or service time distribution or both does not follow Poisson distribution. The results of two such models are summarize below:

(a) **$M/E_k/1$: FIFO/∞ Model**

This queueing model has single server and Poisson input process with mean arrival rate λ. However the service time distribution is Erlang distribution with *k*-phases. The density function of Erlang distribution is given as

$$f(t) = \frac{(\mu k)^k}{(k-1)!} t^{k-1}.e^{-k\mu t}, t \geq 0$$

where μ and *k* are positive parameters and *k* is integer.

In this *k* phases of service, a new customer enters the service channel if the previous customer finishes the *k*-phases of the service. The **characteristics** of this model are listed below:

If λ = mean arrival rate, $\frac{1}{k\mu}$ = Expected service time of each phase, then

(i) $$P_0 = 1 - k\rho$$

(ii) $$W_q = \frac{k+1}{2k}.\frac{\lambda}{\mu(\mu - \lambda)}$$

(iii) $$W_s = W_q + \frac{1}{\mu}$$

(iv) $L_s = \lambda.W_s$

(v) $L_q = \lambda.W_q$

(b) **M/G/1: FIFO/∞ Model**

This queueing model has single server and Poisson input process with mean arrival rate λ. However there is a general distribution for the service time whose mean and variance are $1/\mu$ and σ^2 respectively. For this system the steady state condition is $\rho = \frac{\lambda}{\mu} < 1$. The characteristics of this model are given below:

(i) $P_0 = 1 - \rho$

(ii) $L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$

(iii) $L_s = L_q + \rho$

(iv) $W_q = \frac{L_q}{\lambda}$

(v) $W_s = W_q + \frac{1}{\mu}$.

Example 7: *A barber shop with a one-man takes exactly 20 minutes to complete one haircut. If customers arrive in a Poisson fashion at an average rate of 2 customers per hour calculate the average waiting time in the queue and expected number of customers in the shop.*

Solution: $\lambda = 2$ customers/hour $= \frac{1}{30}$ customers/min.

$\mu = \frac{1}{20}$ customers/min.

$\rho = \frac{\lambda}{\mu} = \frac{2}{3} < 1.$

Since service time is constant, we can take $k = \infty$.

$\therefore \quad W_q = \underset{k\to\infty}{\text{Lt}} \frac{k+1}{2k}.\frac{\lambda}{\mu(\mu-\lambda)} = \frac{1}{2}.\frac{\lambda}{\mu(\mu-\lambda)} = 20$ min.

and $L_s = \lambda.W_s = \lambda\left(W_q + \frac{1}{\mu}\right) = \frac{1}{30}(20+20) = 1.33.$

Check Your Progress

State whether the following statements are True or False:

6. In M/M/1: FIFO/∞ model, there are multiple servers and the capacity of the system is infinity.
7. In M/M/S: FIFO/∞ model, the condition of existence of steady state solution is $\frac{\rho}{s} < 1$.
8. M/M/S: FIFO/N, S ≤ N model is called *s*-server with infinite system capacity.
9. M/M/∞: FIFO/∞ model is also called self-service queueing model.
10. Non-poisson queueing models does not follow poisson distribution.

18.7 BENEFITS AND LIMITATIONS OF QUEUEING THEORY

Queueing theory has been used for many real life applications to a great advantage. It is so because many problems of business and industry can be assumed/simulated to be arrival-departure or queueing problems. In any practical life situations, it is not possible to accurately determine the arrival and departure of customers when the number and types of facilities as also the requirements of the customers are not known. Queueing theory techniques, in particular, can help us to determine suitable number and type of service facilities to be provided to different types of customers. Queueing theory techniques can be applied to problems such as:

(a) Planning, scheduling and sequencing of parts and components to assembly lines in a mass production system.

(b) Scheduling of workstations and machines performing different operations in mass production.

(c) Scheduling and dispatch of war material of special nature based on operational needs.

(d) Scheduling of service facilities in a repair and maintenance workshop.

(e) Scheduling of overhaul of used engines and other assemblies of aircrafts, missile systems, transport fleet, etc.

(f) Scheduling of limited transport fleet to a large number of users.

(g) Scheduling of landing and take-off from airports with heavy duty of air traffic and limited facilities.

(h) Decision of replacement of plant, machinery, special maintenance tools and other equipment based on different criteria.

Special **benefit** which this technique enjoys is solving problems such as above are:

(i) Queueing theory attempts to solve problems based on a scientific understanding of the problems and solving them in optimal manner so that facilities are fully utilised and waiting time is reduced to minimum possible.

(ii) Waiting time (or queueing) theory models can recommend arrival of customers to be serviced, setting up of workstations, requirement of manpower, etc., based on probability theory.

Limitation of Queueing Theory

Though queueing theory provides us a scientific method of understanding the queues and solving such problems, the theory has certain limitations which must be understood while using the technique, some of these are:

(a) Mathematical distribution, which we assume while solving queuing theory problems, are only a close approximation of the behaviour of customers, time between their arrival and service time required by each customer.

(b) Most of the real life queueing problems are complex situation and are very difficult to use the queueing theory technique, even then uncertainty will remain.

(c) Many situations in industry and service are multi-channel queueing problems. When a customer has been attended to and the service provided, it may still have to get some other service from another service point and may have to fall in queue once again. Here the departure of one channel queue becomes the arrival of the other channel queue. In such situations, the problem becomes still more difficult to analyse.

(d) Queuing model may not be the ideal method to solve certain very difficult and complex problems and one may have to resort to other techniques like Monte-Carlo simulation method.

18.8 INTRODUCTION: DECISION THEORY

Management has to make decisions. We have dealt with certain situations in previous units where we had perfect information; such decisions are made under certainty. Most of the mangers make major financial investments and other decisions related with production, marketing, etc., with less than complete information. Decision theory, provides a rational approach in dealing with such situations, where the information is incomplete and uncertain about future conditions. The management must make decisions under such circumstances. With the help of decision theory best possible decision under a particular situation can be taken.

In decision theory, a number of statistical techniques can help the management in making rational decisions. One such statistical decision theory is known as Bayesin decision theory.

NOTES

18.9 DECISION THEORY APPROACH

While discussing the approach, it will be helpful for us to take a real life situation and relate it with the steps involved in taking decisions. Let us take the case of a manufacturing company, which is interested in increasing its production to meet the increasing market demand.

Step I. Determine all possible alternatives

The first obvious step involved before making a rational decision is to list all the viable alternatives available in a particular situation. In the example considered above, the following options are available to the manufacturer:

(a) Expand the existing manufacturing facilities (Expansion);

(b) Setup a new plant (New facilities);

(c) Engage other manufacturers to produce for him as much as is the extra demand (Sub contracting).

Step II. Identify the future scenario

It is very difficult to identify the exact events that may occur in future. However, it is possible to list all that can happen. The future events are not under the control of the decision-maker. In decision theory, identifying the future events is called the *state of nature*. In the case which we have taken of a particular manufacturing company, we can identify the following future events:

(a) Demand continues to increase (High demand)

(b) Moderate demand

(c) Demand starts coming down (Low demand)

(d) The product does not remain in demand (No demand).

Step III. Preparing a payoff table

The decision-maker has to now find out possible payoffs, in terms of profits, if any, of the possible events taking place in future. Putting all the alternatives together (Step I) in relation to the state of nature (Step II) gives us the payoff table. Let us prepare the payoff table for our manufacturing company.

Alternatives	State of nature			
	High demand	**Moderate demand**	**Low demand**	**No demand**
Expansion	1	2	3	4
Add New-Facilities	5	6	7	8
Sub-Contact	9	10	11	12

If expansion is carried out and the demand continues to be high (one of the 12 alternatives), the payoff is going to be maximum in terms of profit of say ₹ X. However, if expansion is carried out and there is no demand (situation 4), the company will suffer a loss.

Step IV. Select the best alternative

The decision-maker will, of course, select the best course of action in terms of payoff. However, it must be understood that the decision may not be based on purely quantitative payoff in terms of profit alone, the decision-maker may consider other qualitative aspect like the goodwill generated which can be encashed in future, increasing market share with an eye on specially designed pricing policy which ultimately gives profits to the company, etc.

18.10 ENVIRONMENT IN WHICH DECISIONS ARE MADE

Decision-maker faces the following situations while making decisions:

(a) **Decision Under Conditions of Certainty**

This is a hypothetical situation in which complete information about the future business environment is available to the decision-maker. It is very easy for him to take a very good decision, as there is no uncertainty involved. But in real life, such situations are never available.

(b) **Decision Under Conditions of Uncertainty**

The future state of events is not known, *i.e.*, there are more than one state of nature. As these uncertainties increase, the situation becomes more complex. The decision-maker does not have sufficient information and cannot assign probabilities to different occurrences.

(c) **Decision Under Risk**

Here, there are a number of states of nature like the above case. The only difference is the decision-maker has sufficient information and can allot probabilities to the different states of nature *i.e.*, the risk can be quantified.

Decision Under Certainty

This is a rare situation and no decision-maker is so fortunate to have complete information before making a decision. Hence, it is not a real life situation and is of no-consequence in managerial decisions.

In this situation only one state of nature exists and its probability is one. With one state of nature, possible alternatives could still be numerous and the decision-maker may use techniques like Linear programming, Transportation and Assignment technique, Economic Order Quantity (EOQ) model, input-output analysis, etc.

In our example of the manufacturing company, if the company had perfect information that the demand would be high: it would have three alternatives of expansion, construction of additional facilities and sub-contracting. Any one Alternative, which gives the best payoff, say constructing additional facilities may be picked up to get the maximum benefit. So, the job of decision-maker is simple just to pick-up the best payoff in the column of state of nature (high demand, low demand, no demand) and use the associated alternative (expand, add facilities, sub-contract).

Decision Under Uncertainty

Under conditions of uncertainty, one may know the state of nature in future but what is the probability of occurrence is not known. Since the data or information is incomplete the decision model becomes complex and the decision is not optimal or the best decision. Such situations and decision problems are called the *Game Theory*, which will be taken up subsequently.

Let us take the case of our manufacturing company. If the company wishes to launch a new product like a DVD player, it knows that the demand of DVDs in future is likely to rise, but the probability that it will increase is not known. Also, the company may face the uncertainty of manufacturing these profitably, because the imported DVDs may become very cheap because of the Government policy.

There are number of criterion available for making decision under uncertainty. The assumption, of course, is that no probability distributions are available under these conditions. The following are discussed in this unit:

(a) The maximax criterion

(b) The minimax (Maximin) criterion

(c) The savage criterion (The Minimax Regret Criterion)

(d) The Laplace criterion (Criterion of Rationality)

(e) The Hurwicz criterion (Criterion of Realism).

In the above criteria, the assumption is also made that the decision-maker does not have art ‘intelligent’ opponent whose interest will oppose the interest of decision-maker. For example, when two armies fight each other, they are a case of intelligent opponents ad such cases are dealt with and handled by Games Theory.

Decision-making Under Conditions of Risk

In real life situations managers have to make-decision under condition of risk. In decision-making under conditions of uncertainty, the decision-maker does not have sufficient information to assign probability to different states of nature. Whereas in decision-making under conditions of risk, the decision-maker has sufficient information to assign probabilities to each of the states of nature.

Decisions under risk are usually based on one of the following criterion:

(a) Expected value criterion (Expected monetary value–EMV criterion)

(b) Combined expected value and variance

(c) Known aspiration level

(d) Most likely occurrence of a future state

Check Your Progress

Choose the correct option for the following statements:

11. In decision theory, identifying ..... is called the state of nature.
    (a) present events (b) past events
    (c) future events (d) None of the above
12. If complete information about the future business is available to the decision-maker it is said to be
    (a) Decision under conditions of certainty
    (b) Decision under conditions of uncertainty
    (c) Decision under risk
    (d) None of these
13. If the decision is not optimal, decision problems are called .....
    (a) assignment theory (b) transportation theory
    (c) game theory (d) linear programming theory
14. Criterion of Rationality in decision making under uncertainty is also called .....
    (a) The Savage Criterion (b) The Laplace Criterion
    (c) The Hurwicz Criterion (d) The Maximum Criterion
15. In decision theory ..... techniques are used to make rational decision.
    (a) qualitative (b) quantitative
    (c) statistical (d) programming

18.11 Summary

- A queueing system involves a number of servers (or serving facilities) which we will also call *service channels* (in deference to the source field of the theory telephone communication system). The serving channels can be communications links, workstations, check out counters, retailers. elevators, buses, to mention but a few.
- The subject matter of queueing theory is to build mathematical models, which relate the specified operating conditions for the system (number of channels, their servicing mechanism, distribution of arrivals) to the concerned characteristics of value-measures of effectiveness describing the ability of the system to handle the incoming demands.
- In the queueing models, arrival of a customer implies a birth and departed customer implies a death. In queueing system both arrivals and departures take place simultaneously. This makes difference from birth-and-death process.
- In this queueing model, arrivals and departures are Poisson with rates λ and μ respectively. There is one server and the capacity of the system is infinity *i.e.,* very large.

18.12 GLOSSARY

- ***FIFO:*** A service discipline which means first in first out.
- ***LIFO:*** A service discipline which means last in first out.
- ***SIRO:*** A service discipline which means service in random order.
- ***Tandem Queues:*** A facility comprising a number of series stations through which the customer may pass for service is called 'Tandem Queues'.
- ***Arrival Pattern:*** It is the pattern of the arrival of a customer to be serviced. The pattern may be regular or at random.
- ***Exponential Distribution:*** This is based on the probability of completion of a service and is most commonly used distribution in queueing theory.
- ***Service Pattern:*** We have seen that arrival pattern is random and poissons distribution can be used for use in queue model. Service pattern are assumed to be exponential for the purpose of avoiding complex mathematical problem.
- ***Traffic Intensity:*** This is the rate at which the service facility is utilised by the components.

- ***State of Nature:*** In decision theory, identifying the future events is called the state of nature.
- ***Deterministic Queueing Models:*** Where the arrivals and service rates are known are called deterministic queueing models. This is rarely used as it is not a practical method.
- ***Probabilistic Model:*** Here both the parameters *i.e.*, the arrivals rate and service rate are unknown and are assumed random in nature.
- ***Mixed Model:*** Where one of the parameters out of the two is known and the other is unknown.

18.13 ANSWERS TO CHECK YOUR PROGRESS

1. service channels
2. LIFO
3. calling source
4. probability distribution
5. balking
6. False
7. True
8. False
9. True
10. True
11. (c)
12. (a)
13. (c)
14. (b)
15. (c)

18.14 TERMINAL AND MODEL QUESTIONS

1. Customers at a box office window, being manned by a single individual arrive according to a Poisson process with a rate of 30 per hr. The time taken to serve a customer has an exponential distribution with a mean of 80 sec. Find the average waiting time of a customer.

2. At a certain Petrol Pump, customers arrive according to a Poisson process with an average time of 6 min. between arrivals. The service time is exponentially distributed with mean time as 3 min. Calculate the following:
   (a) What would be the average number of customers in the petrol pump?
   (b) What is the average waiting time of a car before receiving petrol?
   (c) The per cent of time that the petrol pump is idle.
3. A xerox machine in an office is operated by a person who does other jobs also. The average service time for a job is 6 minutes per customer. On an average, every 12 minutes one customer arrives for xeroxing. Find:
   (a) the xerox machine utilisation.
   (b) percentage of times that an arrival has not to wait.
   (c) average time spent by a customer.
   (d) average queue length.
   (e) the arrival rate if the management is willing to deploy the person exclusively for xeroxing when the average time spent by a customer exceeds 15 minutes.
4. A repairman is to be hired to repair machines, which breakdown at an average rate of 3 per hour. Breakdowns are distributed in time in a manner that may be regarded as Poisson. Non-productive time on any one machine is considered to cost the company ₹ 5 per hour. The company has narrowed the choice to 2 repairmen–one slow but cheap, the other fast but expensive. The slow-cheap repairman asks ₹ 2 per hour, in return he will service breakdown machine exponentially at an average rate of 4 per hour. The second fast expensive repairman demands ₹ 8 per hour and will repair machines exponentially at an average rate of 5 per hour. Which repairman should be hired? (Assume a day is 8 hours).
5. A computer manufacturing firm has a troubleshooting station that can replace a computer component in an average time of 3 minutes. Service is provided on a FIFO basis and the service rate is Poisson distributed. Arrival rates are also Poisson distributed with a mean of 18 per hour. Assuming that this is a single channel, single phase system,
   (a) What is the average waiting time before a component is replaced?
   (b) What is the probability that a component is replaced in less than 10 minutes?
6. A bank has two tellers working on savings accounts. The first teller handles withdrawals only while the second teller on deposits only. For both tellers the service time is exponential with mean service time 3 minutes per customer. Arrivals follow Poisson distribution with mean arrival rate of 16 per hour for depositors and 14 per hour for withdrawers respectively. What would be the

effect on the average waiting time for depositors and withdrawers if each teller handles both withdrawals and deposits?

7. Customers arrive at a bank counter manned by a single person according to a Poisson input process with a mean rate of 10 per hour. The time required to serve a customer has an exponential distribution with a mean of 4 minutes. Find:
   (a) the average number in the system.
   (b) the probability that there would be 3 customer in the queue.
   (c) the probability that the time taken by a customer in the queue is more than 3 minutes.
8. There are two booking clerks at a railway ticket counter. Passengers arrive according to Poisson process with an average rate of 176 per 8 hour day. The mean service time is 5 minutes. Find the idle time of a booking clerk in a day and the average number of customers in the queue.
9. In a cycle repair shop, the inter-arrival times of the customers are exponential with an average time of 10 minutes. The length of service time is assumed to be exponentially distributed with mean 5 minutes. Services are offered by a mechanic. Find the following:
   (a) the probability that an arrival will have to wait for more than 10 minutes before getting a service.
   (b) the probability that an arrival will have to spent at most 15 minutes in the shop.
   (c) the probability that there will be two or more customers in the shop.
10. In a petrol pump vehicles are arriving to buy petrol or diesel according to Poisson distribution with an inter-arrival time of 6 minutes for petrol and 4 minutes for diesel and two different queues are maintained. The service time for both the queues are assumed to be distributed exponentially with an average of 3 minutes. Compare the (a) average waiting time in the two queues, (b) average length of queues from time to time.
11. In a self service queueing system the customers are coming in a Poisson process with an average inter-arrival time of 8 minutes. The service time is exponential with a mean of 4 minutes. Calculate
    (a) probability that there are more than one customer in the system.
    (b) average number of customers in the system.
12. Five components of a certain machine needs to be testing by a mechanic for efficiency in production. For machines the testing is done in a Poisson fashion at an average rate of 2 per hour. Mechanic tests the components in a prescribed

order. The testing times of all the five components are identical exponential distributions with mean 5 minutes.

(a) Find the average time spent before taking service.

(b) Find the average time the machine remains with the mechanic.

13. Explain clearly the various ingredients of a decision problem. What are the basic steps of a decision-making process?
14. What are different environment in which decision are made?
15. Explain clearly the following terms:

    (i) Action space, (ii) State-of-nature, (iii) Payoff table, and (iv) Opportunity loss.
16. Describe some methods, which are useful for decision-making under uncertainty. Illustrate each by an example.
17. Write a short note on decision-making under uncertainty.
18. Indicate the difference between decision under risk and decision under uncertainty in statistical decision theory.

18.15 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitative Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 19: REPLACEMENT THEORY AND SEQUENCING PROBLEMS

Structure

19.0 INTRODUCTION

In the previous unit you have learnt about queueing theory and its various types of queueing problems. In this unit you will learn about theory of replacement and sequencing problems.

The problem of replacement is felt when the job performing units such as men, machines, equipments, parts, etc., become less effective or useless due to either sudden, or gradual deterioration in their efficiency, failure or breakdown. By replacing them with new ones at frequent intervals, maintenance and other overhead costs can be reduced. However, such replacements would increase the need of capital cost for new ones. For example,

(i) A vehicle tends to wear out with time due to constant use. More money needs to be spent on it on account of increased repair and operating cost. A stage comes when it becomes uneconomical to maintain the vehicle and it is better to replace it with a new one. Here the replacement decision may be taken to balance the increasing maintenance cost with the decreasing money value of the vehicle, with the passing of time.

(ii) In case of highway tubelights where time of failure is not predictable for individual tubes, they are replaced only after their individual failure. However, it may be economical to replace such tubes on a schedule basis before their failure. Here the replacement decision may be taken to balance between the wasted life of a tube before failure and cost incurred when a tube completely fails during service.

Thus, the basic problem in such situations is to formulate a replacement policy in order to determine an age (or period) at which the replacement of the given machinery/ equipment is most economical, keeping in view all possible alternatives.

In this unit, we shall discuss the replacement policies in the context of the following three types of replacement situations:

(i) Items such as machines, vehicles, tyres, etc., whose efficiency deteriorates with age due to constant use and which need increased operating and maintenance costs. In such cases the deterioration level is predictable and is represented by (a) increased maintenance/operational cost, (b) its waste or scrap value and damage to items and safety risk.

(ii) Items such as light bulbs and tubes, electric motors, radio, television parts, etc., which do not give any indication of deterioration with time but fail all of a sudden and are rendered useless. Such cases require an anticipation of failures to specify the probability of failure for any future time period. With this probability distribution and the cost information, it is desired to formulate optimal replacement policy in order to balance the wasted life of an item, replaced before failure against the costs incurred when the item fails in service.

(iii) The existing working staff in an organization gradually reduces due to retirement, death, retrenchment and other reasons.

19.1 UNIT OBJECTIVES

After going through this unit you, will be able to:

- Define replacement theory and its need in real life
- Describe various types of failure of an item
- Apply replacement policy for items whose efficiency deteriorates with time
- Apply replacement policy for items that completely fails
- Appreciate the use of replacement analysis in handling problems like ‘staffing problem’ and ‘equipment renewal problem’, etc.
- Explain the meaning and concept of sequencing
- Explain assumptions made in sequencing problem

- Explain terminology associated with sequencing
- Formulate processing of jobs through different number of machines
- Formulate processing of two jobs through *m* machines and *n* jobs through *m* machines

19.2 TYPES OF FAILURE

The term 'failure' here will be discussed in the context of replacement decisions. There are two types of failures: (i) Gradual failure, and (ii) Sudden failure.

Gradual Failure

Gradual failure is progressive in nature. That is, as the life of an item increases, its operational efficiency also deteriorates. This results in:

- increased running (maintenance and operating) costs
- decrease in its productivity
- decrease in the resale or salvage value

Machanical items like pistons, rings, bearings, etc., and automobile tyres fall under this category.

Sudden Failure

This type of failure occurs in items after some period of desired service rather than deterioration while in service. The period of desired service is not constant but follows some frequency distribution which may be *progressive, retrogressive* or *random* in nature.

(a) ***Progressive Failure:*** If the probability of failure of an item increases with the increase in its life, then such a failure is called a progressive failure. For example, light bulbs and tubes fail progressively.

(b) ***Retrogressive Failure:*** If the probability of failure in the beginning of the life of an item is more but as time passes the chances of its failure become less, then such failure is said to be retrogressive.

(c) ***Random Failure:*** In this type of failure, the constant probability of failure is associated with items that air-born equipment have been found to fail at a rate independent of the age of the tube.

NOTES

19.3 REPLACEMENT OF ITEMS WHOSE EFFICIENCY DETERIORATES WITH TIME

When operational efficiency of an item deteriorates with time (gradual failure), it is economical to replace the same with a new one. For example, the maintenance cost of a machine increases with time and a stage is reached when it may not be economical to allow the machine to continue in the system. Besides, there could be a number of alternative choices and one may like to compare the available alternatives on the basis of the running costs (average maintenance and operating costs) involved. In this section, we shall discuss various techniques for making such comparisons under different conditions. While making such comparisons it is assumed that suitable expressions for running costs are available.

Model 1: Replacement Policy for items Whose Running Cost Increases with Time and Value of Money Remain Constant During a Period

Theorem 19.1 The cost of maintenance of a machine is given as a function increasing with time, whose scrap value is constant.

(a) If time is measured continuously then the average annual cost will be minimized by replacing the machine when the average cost to data becomes equal to the current maintenance cost.

(b) If time is measured in discrete units, then the average annual cost will be minimized by replacing the machine when the next period's maintenance cost becomes greater than the current average cost.

Proof: The aim here is to determine the optimal replacement age of a piece of equipment whose running cost increases with time and the value of money remains constant (*i.e*., value is not counted) during that period.

Let C = capital or purchase cost of new equipment.

S = scrap (or salvage) value of the equipment at the end of t years.

R(t) = running cost of equipment for the year t

n = replacement age of the equipment

(a) ***When time 't' is a continuous variable:*** If the equipment is used for t years, then the total cost incurred over this period is given by:

TC = Capital (or purchase) cost – Scrap value at the end of t years

+ Running cost for t years

$$= C - S + \int_0^n R(t)\, dt$$

Therefore, the average cost per unit time incurred over the period on n years is:

$$\text{ATC}_n = \frac{1}{n}\left\{\text{C} - \text{S} + \int_0^n \text{R}(t)\, dt\right\} \quad ...(1)$$

To obtain the optimal value n for which ATC_n is minimum, differentiate ATC_n with respect to n, and set the first derivative equal to zero. That is, for minimum of ATC_n.

$$\frac{d}{dn}\{\text{ATC}_n\} = -\frac{1}{n^2}\{\text{C} - \text{S}\} + \frac{\text{R}(n)}{n} - \frac{1}{n^2}\int_0^n \text{R}(t)\, dt = 0$$

or
$$\text{R}(n) = \frac{1}{n}\left\{\text{C} - \text{S} + \int_0^n \text{R}(t)\, dt\right\}, \quad n \neq 0 \quad ...(2)$$

$$\text{R}(n) = \text{ATC}_n$$

Hence, the following replacement policy can be derived with the help of Eq. (2).

Policy. *Replace the equipment when the average annual cost for n years becomes equal to the current/annual running cost. That is:*

$$\text{R}(n) = \frac{1}{n}\left\{\text{C} - \text{S} + \int_0^n \text{R}(t)\, dt\right\}$$

(b) ***When time 't' is a discrete variable:*** The average cost incurred over the period n is given by:

$$\text{ATC}_n = \frac{1}{n}\left\{\text{C} - \text{S} + \sum_{t=0}^{n} \text{R}(t)\right\} \quad ...(3)$$

If C – S and $\sum_{t=0}^{n} \text{R}(t)$ are assumed to be monotonically decreasing and increasing, respectively, then there will exist a value of n for which ATC_n is minimum. Thus, we shall have inequalities:

$$\text{ATC}_{n-1} > \text{ATC}_n < \text{ATC}_{n+1}$$

or
$$\text{ATC}_{n-1} - \text{ATC}_n > 0$$

and
$$\text{ATC}_{n+1} - \text{ATC}_n > 0$$

Eq. (3) for period $n + 1$, we get:

$$\text{ATC}_{n+1} = \frac{1}{n+1}\left\{\text{C} - \text{S} + \sum_{t=1}^{n+1} \text{R}(t)\right\} = \frac{1}{n+1}\left\{\text{C} - \text{S} + \sum_{t=1}^{n} \text{R}(t) + \text{R}(n+1)\right\}$$

$$= \frac{n}{n+1}\frac{\left\{\text{C} - \text{S} + \sum_{t=1}^{n} \text{R}(t)\right\}}{n} + \frac{\text{R}(n+1)}{n+1} = \frac{n}{n+1} \cdot \text{ATC}_n + \frac{\text{R}(n+1)}{n+1}$$

NOTES

Therefore, $\text{ATC}_{n+1} - \text{ATC}_n = \frac{n}{n+1}\ \text{ATC}_n + \frac{\text{A}(n+1)}{n+1} - \text{ATC}_n$

$$= \frac{\text{R}(n+1)}{n+1} + \text{ATC}_n\left(\frac{n}{n+1} - 1\right) = \frac{\text{R}(n+1)}{n+1} - \frac{\text{ATC}_n}{n+1}$$

Since $\text{ATC}_{n+1} - \text{ATC}_n > 0$, we get

$$\frac{\text{R}(n+1)}{n+1} - \frac{\text{ATC}_n}{n+1} > 0$$

$$\text{R}(n+1) - \text{ATC}_n > 0 \quad \text{or} \quad \text{R}(n+1) > \text{ATC}_n$$

Similarly, $\text{ATC}_{n-1} - \text{ATC}_n > 0$, implies that $\text{R}(n+1) < \text{ATC}_{n-1}$. This provides the following replacement policy.

Policy 1. *If the running cost of next year, R(n + 1) is more than the average cost of nth year, ATC_n, then it is economical to replace at the end of n years. That is:*

$$\text{R}(n+1) > \frac{1}{n}\left\{\text{C} - \text{S} + \sum_{t=0}^{n} \text{R}(t)\right\}$$

Policy 2. *If the present year's running cost is less than the previous year's average cost, ATC_{n-1}, then do not replace. That is:*

$$\text{R}(n) < \frac{1}{n-1}\left\{\text{C} - \text{S} + \sum_{t=0}^{n-1} \text{R}(t)\right\}$$

Example 1: *A firm is considering the replacement of a machine, whose cost price is ₹ 12,200, and its scrap value is ₹ 200. From experience the running (maintenance and operating) costs are found to be as follows:*

Year	:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Running cost (₹)	:	*200*	*500*	*800*	*1,200*	*1,800*	*2,500*	*3,200*	*4,000*

When should the machine be replaced?

Solution: We are given the running cost, R(*n*), the scrap value S = ₹ 200 and the cost of the machine, C = ₹ 12,200. In order to determine the optimal time *n* when the machine should be replaced, we first calculate the average cost per year during the life of the machine, as shown in Table 19.1.

Table 19.1: Calculations of Average Cost

Year of service n	Running cost (₹) $R(n)$	Cumulative running cost (₹) $\Sigma R\ (n)$	Depreciation cost (₹) $C - S$	Total cost (₹) TC	Average cost (₹) ATC_n
(1)	(2)	(3)	(4)	(5) = (3) + (4)	(6) = (5)/(1)
1	200	200	12,000	12,200	12,000
2	500	700	12,000	12,700	6,350
3	800	1,500	12,000	13,500	4,500
4	1,200	2,700	12,000	14,700	3,675
5	1,800	4,500	12,000	16,500	3,300
6	2,500	7,000	12,000	19,000	**3,167**
7	3,200	10,200	12,000	22,200	3,171
8	4,000	14,200	12,000	26,200	3,275

The average cost per year, ATC_n = ₹ 3,167 is minimum in the sixth year as shown in Table 19.1. Also the average cost, ₹ 3,171 in seventh year is more than the cost in the sixth year. Hence, the machine should be replaced after every six years.

Example 2: *The data collected in running a machine, the cost of which is* ₹ *60,000 are given below:*

Year	:	*1*	*2*	*3*	*4*	*5*
Resale value (₹)	:	*42,000*	*30,000*	*20,400*	*14,400*	*9,650*
Cost of spares (₹)	:	*4,000*	*4,270*	*4,880*	*5,700*	*6,800*
Cost of labour (₹)	:	*14,000*	*16,000*	*18,000*	*21,000*	*25,000*

Determine the optimum period for replacement of the machine.

Solution: The costs of spares and labour, together determine the running (operational or maintenance) cost. Thus, the running costs and the resale price of the machine in successive years are as follows:

Year	**:**	***1***	***2***	***3***	***4***	***5***
Resale value (₹)	:	42,000	30,000	20,400	14,400	9,650
Running cost (₹)	:	18,000	20,270	22,880	26,700	31,800

The calculations of average running cost per year during the life of the machine are shown in Table 19.2.

Table 19.2: Calculations of Average Running Cost

Year of service n	Running cost (₹) $R(n)$	Cumulative running cost (₹) $\Sigma R\ (n)$	Resale value (₹) S	Depreciation cost (₹) $C - S$	Total cost (₹) TC	Average cost (₹) ATC_n
(1)	(2)	(3)	(4)	(5) = 60,000 – (4)	(6) = (3) + (5)	(7) = (6)/(1)
1	18,000	18,000	42,000	18,000	36,000	36,000.00
2	20,270	38,270	30,000	30,000	68,270	34,135.00
3	22,880	61,150	20,400	39,600	1,00,750	33,583.30
4	26,700	87,850	14,400	45,600	1,33,450	**33,362.50**
5	31,800	1,19,650	9,650	50,350	1,70,000	34,000.00

The average cost, ATC_4 = ₹ 33,362.50 is the lowest during the fourth year as shown in Table 19.2. Hence, the machine should be replaced after every four years, otherwise the average cost per year for running the machine would start increasing.

Model II: Replacement Policy for Items Whose Running Cost Increases with Time but Value of Money Changes with Constant Rate During a Period

Value of money criterion: If the effect of the time-value of money is to be considered, then replacement decision must be based upon an equivalent annual cost. For example, if the interest rate on ₹ 100 is 10 percent year, then the value of ₹ 100 to be spent after one year will be ₹ 110. This is also called *value of money*. Also, the value of money that decreases with constant rate is known as its *depreciation ratio* or *discounted factor.* The discounted value is the amount of money required to build up funds at compound interest that is sufficient to pay the required cost when due. For example, if the interest rate on ₹ 100 is *r* per cent per year, then the *present value* (*or worth*) of ₹ 100 to be spent after *n* years will be:

$$d = \left(\frac{100}{100 + r}\right)^n$$

where *d* is the *discount rate* or *depreciation value.* After calculating depreciation value, we need to determine the critical age at which an item should be replaced so that the sum of all discounted costs is minimum.

Example 3: *Let the value of the money be assumed to be 10 % per year and suppose that machine A is replaced after every three years, whereas machine B is replaced every six years. The yearly costs (in ₹) of both the machines are given below:*

Year	:	*1*	*2*	*3*	*4*	*5*	*6*
Machine A	:	*1,000*	*200*	*400*	*1000*	*200*	*400*
Machine B	:	*1,700*	*100*	*200*	*300*	*400*	*500*

Determine which machine should be purchased.

Solution: The discounted cost (present worth) at 10 per cent rate per year for machine A for three years and B for six years is given in Tables 19.3 and 19.4, respectively.

Table 19.3: Discounted Cost of Machine A

Year	Discounted cost at *10%* Rate (₹)	
	Cost	**Present Worth**
1	1,000	1000×1 = 1,000
2	200	$200\left(\frac{100}{100+10}\right) = 200 \times 0.9091$ = 181.82
3	400	$400\left(\frac{100}{100+10}\right)^2 = 400 \times 0.8264$ = 330.58
		Total ₹ 1,512.40

Thus, the average yearly cost of machine A is 1,512.40/3 = ₹ 504.13

Table 19.4: Discounted Cost of Machine B

Year	Discounted cost at *10%* Rate (₹)	
	Cost	**Present worth**
1	1,700	$1,700 \times 1$ = 1,700.00
2	100	$100 \times (10/11) = 100 \times 0.9091$ = 90.91
3	200	$200 \times (10/11)^2 = 200 \times 0.8264$ = 165.28
4	300	$300 \times (10/11)^3 = 200 \times 0.7513$ = 225.39
5	400	$400 \times (10/11)^4 = 400 \times 0.6830$ = 273.20
6	500	$500 \times (10/11)^5 = 500 \times 0.6209$ = 310.45
		Total ₹ 2,765.23

Thus, the average yearly cost of machine B is 2,765.23/6 = ₹ 460.87.

With the data on average yearly cost of both the machines, the apparent advantage is in purchasing machine B. But, the periods for which the costs are considered are different. Therefore, let us first calculate the total present worth of machine A for six years.

Total present worth = $1,000 + 200 \times 0.9091 + 400 \times 0.8264 + 1,000 \times 0.7513$

$+ 200 \times 0.6830 + 400 \times 0.6209$ = ₹ 2,648.64

This is less than the total present worth of machine B. Thus machine A should be purchased.

Example 4: *A pipeline is due for repairs. The repair would cost* ₹ *10,000 and would last for three years. Alternatively, a new pipeline can be laid at a cost of* ₹ *30,000, which would for 10 years. Assuming the cost of capital to be 10 per cent and ignoring salvage value, which alternative should be choosen?*

Solution: Consider the two types of pipelines for infinite replacement cycles of ten years for the new pipeline and three years for the existing pipeline.

Since the discount rate of money per year is 10 per cent, the present worth of the money to be spent over a period of one year is: $d = 100/(100 + 10) = 0.9091$

Let D_n be the discounted value of all future costs associated with a policy of replacing the equipment after n years. Then

$$D_n = c + c \times d^n + c \times d^{2n} + ...$$

$$= c(1 + d^n + d^{2n} + ...) = \frac{c}{1 - d^n} \text{ (sum of infinite GP)}$$

where c is the initial cost.

Substituting the values of c, d's and n for two types of pipelines, we get:

$$D_3 = \frac{10,000}{1 - (0.9091)^3} = ₹ 4,021 \text{ (approx), for existing pipeline}$$

and

$$D_{10} = \frac{30,000}{1 - (0.9091)^{10}} = ₹ 48,820, \text{ for new pipeline}$$

Since the value of $D_3 < D_{10}$, the existing pipeline should be continued. Alternatively, the comparison may be made over $3 \times 10 = 30$ years.

Present worth factor criterion: In this case the optimal value of replacement age of an equipment can be determined under the following two situations:

(i) The running cost of an equipment that deteriorates over a period of time increases and the value of the money decreases with a constant rate. If r is the interest rate, then:

$$Pwf = (1 + r)^{-n}$$

is called the *present worth factor* (*Pwf*) or present value of one rupee spent in n years from time now onwards. But if $n = 1$ the *Pwf* is given by:

$$d = (1 + r)^{-1}$$

where d is called the *discount rate or depreciation value*.

(ii) The money to be spent is taken on loan for a certain period at a given rate under the condition of repayment in installments.

The replacement of items on the basis of present worth factor (*Pwf*) includes the present worth of all future expenditure and revenues for each replacement alternatives. An item for which the present worth factor is less, is preferred. Let:

C = purchase cost of an item,

R = annual running cost

n = life of the item in years,

r = annual interest rate

S = scrap (or salvage) value of the item at the end of its life

Then the present worth of the total cost during n years is given by:

Total cost = $C + R$ (*Pwf for r% interest rate for n years*)

$-$ S (*Pwf for r% interest rate for n years*)

If the running cost of the item is different for its different operational life, then the present worth of the total cost during n years is given by:

Total cost = $C + R$ (*Pwf for r% interest rate for i years*)

$-$ S (*Pwf for r% interest rate for i years*) where $i = 1, 2, \ldots n$.

Example 5: *A company is considering the purchase of a new machine at ₹ 15,000. The economic life of the machine is expected to be 8 years. The salvage value of the machine at the end of the life will be ₹ 3,000. The annual running cost is estimated to be ₹ 7,000.*

(a) *Assuming an interest rate of 5 per cent, determine the present worth of future costs of the proposed machine.*

(b) *Compare the new machine with the presently-owned machine that has an annual operating cost of ₹ 5,000 and cost of repair ₹ 1,500 in the second year, with an annual increase of ₹ 500 in the subsequent years of its life.*

Solution: (a) **New Machine**

(i) Purchase cost C = ₹ 15,000

(ii) Present worth of annual operating cost

= 7,000 × *Pwf* at 5% interest for 8 years

= 7,000 × 6.4632 = ₹ 45,242.40

(iii) Present worth of the salvage value

= 3,000 × *Pwf* at 5% interest for 8 years

= 3,000 × 0.6768 = ₹ 2,030.40

Thus, the present worth of total future costs for new machine for eight years will be

15,000 + 45,242.4 + 2,030 = ₹ 58,212.

NOTES

(b) **Old Machine**

The calculations of the present worth of the old machine are shown in Table 19.5.

Table 19.5: Present' Worth of Old Machine

Year of service	Operating cost (₹)	Repair cost (₹)	Total operating and repair cost (₹)	Pwf for single payment	Present worth (₹)
1	5,000	—	5,000	0.9524	5,000 × 0.9524 = 4762.00
2	5,000	1,500	6,500	0,9072	5,895.50
3	5,000	2,000	7,000	0,8638	6,046.60
4	5,000	2,500	7,500	0,8227	6.252.52
5	5,000	3,000	8,000	0,7835	6,268.00
6	5,000	3,500	8,500	0,7462	6,342.70
7	5,000	4,000	9,000	0,7107	6,396.30
8	5,000	4,500	9,500	0,6778	6,429.60
Total					44,103.20

Since the present worth of old machine, as shown in Table 19.5, is less than that of the new machine, the new machine should not be purchased.

Example 6: *A person is considering purchasing a machine for his own factory. Relevant data about alternative machines are as follows:*

	Machine A	***Machine B***	***Machine C***
Present investment (₹)	*10,000*	*12.000*	*15,000*
Total annual cost (₹)	*2,000*	*1,500*	*1,200*
Life (years)	*10*	*10*	*10*
Salvage value (₹)	*500*	*1,000*	*1,200*

As an adviser to the buyer, you have been asked to select the best machine, considering 12 per cent normal rate of return.

You are given that:

(a) Single payment present worth factor (*Pwf*) at 12 per cent interest for 10 years (= 0.322).

(b) Annual series present worth factor (*Pwf*) at 12 per cent interest for 10 years (= 5.650).

Solution: The present value of the total cost of each of the three machines, for a period of ten years, given in Table 19.6.

Table 19.6: Present Value of Total Cost for Three Machines

Machine	Present investment	Present value of total annual cost	Present value of salvage value	Net cost (₹)
(1)	*(2)*	*(3)*	*(4)*	*(5) = (2) + (3) – (4)*
A	10,000	2,000 × 5.65 = 11,300	500 × 0.322 = 161.00	21,139.00
B	12,000	1,500 × 5.65 = 8,475	1,000 × 0.322 = 322.00	20,153.00
C	15,000	1,200 × 5.65 = 6,780	1,200 × 0.322 = 386.40	21,393.60

Table 19.6 shows that the present value of total cost for machine B is the least and hence, machine B should be purchased.

General Cost Function

Theorem 19.2: If the maintenance cost increases with time and the money value decreases with constant rate, *i.e.*, its depreciation value is given, its replacement policy would then be based on the following:

(a) Replace if the running cost of next period is greater than the weighted average of previous cost.

(b) Do not replace if the running cost of the next period is less than the weighted average of the previous costs.

Proof: Suppose that the item (machine or equipment) is available for use over a series of time periods of equal length, say one year. Let us use the following notations:

C = purchase price of a new item

R_n = running cost of the item at the beginning of nth year ($R_{n+1} > R_n$)

r = annual interest rate

d = depreciation value per unit of money during a year $\{1/(1 + r)\}$.

Let us assume that the item is replaced after every n years of service and has no resale value (or price). To arrive at a replacement policy, calculating the total amount of money required for purchasing and running the item for n years. The year(s) for which the total money is minimum will represent best period for replacement.

The present worth (discounted value) of all future costs of purchasing and running the item with a policy of replacing it after every n years is given by:

$$\begin{aligned} D_n = {} & [(C + R_1) + d\,R_2 + d^2 R_3 + \ldots + d^{n-1} R_n] \qquad \text{(for 1 to } n \text{ years)} \\ & + [d^n (C + R_1) + d^{n+1} R_2 + d^{n+2} R_3 + \ldots + d^{2n-1} R_n] \\ & \qquad \text{(for } n + 1 \text{ to } 2n \text{ years)} \\ & + [d^{2n} (C + R_1) + d^{2n+1} R_2 + d^{2n+2} R_3 + \ldots + d^{3n-1} R_n] + \ldots \\ & \qquad \text{(for } 2n + 1 \text{ to } 3n \text{ years)} \\ = {} & [(C+R_1)(1 + d^n + d^{2n} + \ldots) + d\,R_2(1 + d^n + d^{2n} + \ldots) \\ & \qquad + \ldots + d^{n-1} R_n(1 + d^n + d^{2n} + \ldots)] \end{aligned}$$

NOTES

$$= [(C + R_1) + d\,R_2 + ... + d^{n-1}\,R_n]\,[1 + d^n + d^{2n} + ...]$$

$$= \left[C + \sum_{i=1}^{n} d^{i-1}\,R_i\right]\left[\frac{1}{1-d^n}\right] \text{ (sum of infinite G.P., D < 1)} \quad ...(5)$$

For Eqn. (5), if n is an optimal replacement interval, then D_n represents the minimum money required to pay all future costs of purchasing and running an item because of the following inequality:

$$D_{n+1} > D_n < D_{n-1}$$

From this, the two inequalities, $D_{n+1} - D_n > 0$ and $D_n - D_{n-1} < 0$ can be established.

From Eq. (5), since $D_n = \frac{P(n)}{1-d^n}$,

therefore, $$D_{n+1} = \frac{P(n+1)}{1-d^{n+1}} = \frac{(1-d^n)\,D_n + d^n\,R_{n+1}}{1-d^{n+1}}$$

$$= \frac{1-d^n}{1-d^{n+1}} \cdot D_n + \frac{d^n\,R_{n+1}}{1-d^{n+1}}$$

where, $$P(n) = C + R_1 + d\,R_2 + d^2\,R_3 + ... + d^{n-1}\,R_n$$

Now, considering the inequality

$$D_{n+1} - D_n = \frac{P(n+1)}{1-d^{n+1}} - \frac{P(n)}{1-d^n} = \frac{P(n+1)\,(1-d^n) - P(n)\,(1-d^{n+1})}{(1-d^n)\,(1-d^{n+1})}$$

$$= \frac{\{P(n+1) - P(n)\} + d^{n+1}\,P(n) - d^n P(n+1)}{(1-d^n)\,(1-d^{n+1})}$$

$$= \frac{d^n R_{n+1} + d^{n+1}\,P(n) - d^n\{P(n) - d^n R_{n+1}\}}{(1-d^n)\,(1-d^{n+1})}$$

[Because $P(n+1) = P(n) + d^n\,R_{n+1}$]

$$= \frac{d^n(1-d^n)\,R_{n+1} - d^n(1-d)\,P(n)}{(1-d^n)\,(1-d^{n+1})}$$

$$= \frac{d^n(1-d^n)}{(1-d^n)\,(1-d^{n+1})}\left[\frac{1-d^n}{1-d} \cdot R_{n+1} - P(n)\right] \quad ...(6)$$

Since $d < 1$, and so $1 - d_n > 0$, therefore, $D_{n+1} - D_n$ is always positive and has the same sign as the quantity in the bracket in Eq. (6).

Similarly, putting $n - 1$ for n in Eq. (6), we have:

$$D_n - D_{n-1} = \frac{d^{n-1}(1-d)}{(1-d^{n-1})\,(1-d^n)}\left[\frac{1-d^{n-1}}{1-d} \cdot R_n - P(n-1)\right]$$

$$= \frac{d^{n-1}(1-d)}{(1-d^n)(1-d^{n-1})}\left[\frac{1-d^{n-1}}{1-d}.\mathrm{R}_n - \{\mathrm{P}(n) - \mathrm{R}_n\, d^{n-1}\}\right]$$

$$= \frac{d^{n-1}(1-d)}{(1-d^n)(1-d^{n-1})}\left[\frac{1-d^n}{1-d}.\mathrm{R}_n - \mathrm{P}(n)\right] \quad ...(7)$$

Hence, the condition for minimum value of D_n from Eqs. (6) and (7) can be expressed as:

$$(\mathrm{D}_n - \mathrm{D}_{n-1}) < 0 < (\mathrm{D}_{n+1} - \mathrm{D}_n)$$

$$\left[\frac{1-d^n}{1-d}.\mathrm{R}_n - \mathrm{P}(n)\right] < 0 < \left[\frac{1-d^n}{1-d}.\mathrm{R}_{n+1} - \mathrm{P}(n)\right]$$

or $$\frac{1-d^n}{1-d}.\mathrm{R}_n < \mathrm{P}(n) < \frac{1-d^n}{1-d}.\mathrm{R}_{n+1}$$

or $$\mathrm{R}_n < \frac{\mathrm{C} + (\mathrm{R}_1 + d\,\mathrm{R}_2 + d^2\,\mathrm{R}_3 + ... + d^{n-1}\,\mathrm{R}_n)}{1 + d + d^2 + ... + d^{n-1}} < \mathrm{R}_{n+1} \quad ...(8)$$

The expression between R_n and R_{n+1} in Eq. (8) represents the weighted average $\mathrm{W}(n)$ of all costs up to the period $(n-1)$ with weights 1, d, d^2, ..., d^{n-1}, respectively. The given weights are actually the discounted factors of the costs in the previous years. Hence *the value of n, satisfying the relationship* (8), *will be the best age for replacing the given item.*

Check Your Progress

Fill in the blanks:

1. Gradual failure is ..... in nature.
2. If the probability of failure of an item increases with the increase in its life such a failure is called .....
3. Failure of vacuum tubes in air-bornel equipment is a type of .....
4. When ..... of an item deteriorates with time (gradual failure), it is economical to replace the same with the new one.
5. Mechanical items like piston, rings, bearings, etc. and automobile tyre fall under the category .....

19.4 REPLACEMENT OF ITEMS THAT COMPLETELY FAIL

It is somehow difficult to predict that a particular equipment will fail at a particular time. This uncertainty can be avoided by deriving the probability distribution of

NOTES

failures. Here it is assumed that the failures occur only at the end of the period, say t. Thus, the objective is to find the value of t that minimizes the total post involved for the replacement.

Mortality Tables: These tables are used to derive the probability distribution of life span of an equipment in question. Let:

$M(t)$ = number of survivors at any time t

$M(t-1)$ = number of survivors at any time $t-1$

N = initial number of equipments

Then the probability of failure during time period t is given by:

$$P(t) = \frac{M(t-1) - M(t)}{N}$$

The probability that an equipment has survived to an age $(t-1)$, and will fail during the interval $(t-1)$ to t can be defined as the *conditional probability* of failure. It is given by:

$$P_c(t) = \frac{M(t-1) - M(t)}{M(t-1)}$$

The probability of survival to an age t is given by:

$$P_s(t) = \frac{M(t)}{N}$$

Mortality Theorem 19.3: A large population is subject to a given mortality law for a very long period of time. All deaths are immediately replaced by births and there are no other entries or exits. Show that the age distribution ultimately becomes stable and that the number of deaths per unit time becomes constant and is equal to the size of the total population divided by the mean age at death.

Proof: Without any loss of generality, it is assumed that death (or failure) occurs just before the age of $(k+1)$ years, where k is an integer. That is, the life span of an item lies between $t = 0$ and $t = k$. Let us define,

$f(t)$ = number of births (replacements) at time t, and

$p(x)$ = probability of death (failure) just before the age $x+1$, *i.e.* failure at time x.

and $$\sum_{x=0}^{k} p(x) = 1$$

If $f(t-x)$ represents the number of births at time $t-x$, $t = k, k+1, k+2, \ldots$ then the age of newly born attain the age x at time t is illustrated in the figure below:

NOTES

Hence, the expected number of deaths of such newly borns before reaching the time $t + 1$ (*i.e.* at time t) would be:

$$\text{Expected number of death} = \sum_{x=0}^{k} p(x)\, f(t-x), \quad t = k, k+1 \ldots$$

Since all deaths (failures) at time t are replaced immediately by births (replacements) at time $t + 1$, the expected number of births are:

$$f(t+1) = \sum_{x=0}^{k} \mathrm{P}(x)\, f(t-x), \quad t = k, k+1 \ldots \qquad \ldots(9)$$

The solution to the difference Eq. (9) in t can be obtained by putting the value $f(t) = \mathrm{A}\alpha^t$, where A is some constant. Eq. (9) then becomes:

$$\mathrm{A}\,\alpha^{t+1} = \mathrm{A} \sum_{x=0}^{k} \mathrm{P}(x)\, \alpha^{t-x} \qquad \ldots(10)$$

Dividing both sides of Eq. (10) by $\mathrm{A}\alpha^{t-k}$, we get:

$$\alpha^{k+1} = \sum_{x=0}^{k} \mathrm{P}(x)\, \alpha^{k-x} = \alpha^k \sum_{x=0}^{k} \mathrm{P}(x)\, \alpha^{-x}$$

$$= \alpha^k \{p(0) + p(1)\, \alpha^{-1} + p(2)\, \alpha^{-2} + \ldots\}$$

or
$$\alpha^{k+1} - \{p(0)\alpha^k + p(1)\alpha^{k-1} + \ldots + p(k)\} = 0 \qquad \ldots(11)$$

Equation (11) is of degree $(k + 1)$ and will, therefore, have exactly $(k + 1)$ roots. Let us denote the roots of Eq. (11) by $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k$.

For $\alpha = 1$, the LHS of Eq. (11) becomes:

$$\text{LHS} = 1 - \{p(0) + p(1) + \ldots + p(w)\} = 1 - 1 \sum_{x=0}^{k} p(x) = 0 = \text{RHS}$$

Hence, one root of Eq. (11) is $\alpha = 1$. Let us denote this root by α_0. The general solution of Eq. (11) will then be of the form

$$f(t) = \mathrm{A}_0\, \alpha_0^t + \mathrm{A}_1\, \alpha_1^t + \ldots + \mathrm{A}_k\, \alpha_k^t$$

$$= \mathrm{A}_0 + \mathrm{A}_1\, \alpha_1^t + \mathrm{A}_2\, \alpha_2^t + \ldots + \mathrm{A}_k\, \alpha_k^t$$

where $\mathrm{A}_0, \mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_k$ are constant whose values are to be calculated. ...(12)

Since one of the roots of Eq. (11), $\alpha_0 = 1$ is positive, according to the *Descrate's sign rule* all other roots $\alpha_1, \alpha_2, \ldots, \alpha_k$ will be negative and their absolute value would be less than unity, *i.e.* $|\alpha_i| < 1$, $i = 1, 2, \ldots, k$. It follows that the value of these roots tends to zero as $t \to \infty$. With the result that Eq. (12) becomes $f(t) = \mathrm{A}_0$. This indicates that the number of deaths (as well as births) becomes constant at any time.

NOTES

Now the problem is to determine the value of the constant A_0. For this we can proceed as follows. Let us define:

$g(x)$ = Probability of survival for more than x years

or $g(x) = 1 - \text{prob (survivor will die before attaining the age } x)$

$= 1 - \{p(0) + p(1) + ... + p(x-1)\}$

Obviously, it can be assumed that $g(0) = 1$.

Since the number of births as well as deaths has become constant and equal to A_0, the expected number of survivors of age x is given by $A_0 \,.\, g(x)$.

As deaths are immediately replaced by births, size N of the population remains constant. That is,

$$N = A_0 \sum_{x=0}^{k} g(x) \quad \text{or} \quad A_0 = \frac{N}{\sum_{x=0}^{k} g(x)} \qquad ...(13)$$

The denominator in Eq. (13) represents the average age at death. This can also be proved as follows:

From finite differences, we know that:

$$\Delta(x) = (x+1) - x = 1$$

$$\sum_{x=a}^{b} f(x)\,\Delta\, h(x) = f(b+1)\, h(b+1) - f(a)\, h(a) - \sum_{x=a}^{b} h(x+1)\,\Delta f(x)$$

Therefore, we can write,

$$\sum_{x=0}^{k} g(x) = \sum_{x=0}^{k} g(x)\,\Delta\,(x) = [g(x)\,.\,x]_0^{k+1} - \sum_{x=0}^{k} (x+1)\,\Delta\, g(x)$$

$$= g\,(k+1)\,(k+1) - g(0)\,.\,0 - \sum_{x=0}^{k} (x+1)\,\Delta\, g(x)$$

$$= g\,(k+1)\,(k+1) - \sum_{x=0}^{k} (x+1)\,\Delta\, g(x) \qquad ...(14)$$

But $g(k+1) = 1 - \{p(0) + p(1) + p(2) + ... + p(k)\} = 0$

Since no one can survive for more than $(k+1)$ years of age and

$$\Delta g(x) = g(x+1) - g(x)$$

$$= \{1 - p(0) - p(1) - ... - p(x)\} - \{1 - p(0) - p(1) - ... - p(x-1)\} = -p(x)$$

Substituting the value of $g(k+1)$ and $\Delta g(x)$ in Eq. (14), we get

$$\sum_{x=0}^{k} g(x) = \sum_{x=0}^{k} (x+1)\, p(x) = \text{Mean age at death}$$

Hence from Eq. (13), we get $A_0 = \dfrac{N}{\text{Average age at death}}$

Individual Replacement Policy

Under this policy, an item (machine or equipment) is replaced individually as when it failed. This ensures smooth running of the system.

Group Replacement Policy

Sometime the immediate replacement on failure of the item(s) is costly. In such cases a *group replacement policy* is preferred. Under this policy items are replaced at the end of some suitable time period, without waiting for their failure, but if any item fails before the time specified, it may also be replaced individually. In group replacement policy, we need to notice the following:

(i) the rate of individual replacement during the specified time period

(ii) the total cost incurred for individual as well as group replacement during the specified time

Obviously, a time period shall be considered optimal time for replacement when the total cost of replacement is minimum. In order to calculate optimal time period for replacement, the data on (i) probability of failure, (ii) loss incurred due to these failures, (iii) cost of individual replacement, and (iv) cost of group replacement, are required.

Remark: The group replacement policy is suitable for a large number of identical low cost items that are likely to fail with age and for which it is difficult as well as not justified to keep the record of their individual ages.

Theorem 19.3: (*Group Replacement Policies*) (a) Group replacement should be made at the end of the period, t, if the cost of individual replacements for the period t is greater than the average cost per period through the end of period t.

(b) Group replacement is not advisable at the end of period t if the cost of individual replacements at the end of period $t-1$ is less than the average cost per period through the end of period t.

Proof: Let us consider the following notations:

n = total number of items in the system

$F(t)$ = number of items failing during time t

$C(t)$ = total cost of group replacement until the end of period t

C_1 = unit cost of replacement in a group

NOTES

C_2 = unit cost of individual replacement after time t, *i.e.* failure

L = maximum life of any item

$p(t)$ = probability of failure of any item at age t

Rate of Replacement at Time t: The number of failures at any time t is

$$F(t) = \begin{cases} np(t) + \sum_{x=1}^{t-1} p(x)\, F(t-x), & t \le L \\ \sum_{x=1}^{L} p(x)\, F(t-x), & t > L \end{cases} \quad ...(15)$$

Cost of Replacement at Time t: The cost of group replacement after time period t is given by:

$$C(t) = nC_1 + C_2 \sum_{x=1}^{t-1} F(x) \quad ...(16)$$

In Eq. (16) nC_1 is the cost of replacing the items as a group, and $C_2 \sum_{x=1}^{t-1} F(x)$ is the cost of replacing the individual failures at the end of each of $t-1$ periods before the group is replaced again.

The average cost per unit period is then given by:

$$\frac{C(t)}{t} = \frac{nC_1}{t} + \frac{C_2}{t} \sum_{x=1}^{t-1} F(x) \quad ...(17)$$

For optimal replacement period t, the value of average cost per unit period, given by Eq. (17), should be the minimum. The condition for minimum of $C(t)/t$ is:

$$\Delta \left\{ \frac{C(t)}{t-1} \right\} < 0 < \Delta \left\{ \frac{C(t)}{t} \right\}$$

Now for $\Delta \left\{ \frac{C(t)}{t} \right\} > 0$ we have $\Delta \left\{ \frac{C(t)}{t} \right\} = \frac{C(t+1)}{t+1} - \frac{C(t)}{t} > 0$

From Eq. (17), we get:

$$\frac{C(t+1)}{t+1} - \frac{C(t)}{t} = nC_1 \left\{ \frac{1}{t+1} - \frac{1}{t} \right\} + C_2 \sum_{x=1}^{t-1} F(x) \left\{ \frac{1}{t+1} - \frac{1}{t} \right\} + \frac{C_2 F(t)}{t+1}$$

$$= \left\{ -nC_1 - C_2 \sum_{x=1}^{t-1} F(x) + tC_2 F(t) \right\} \Big/ t(t+1)$$

For $\frac{C(t+1)}{t+1} - \frac{C(t)}{t} > 0$, it is necessary that:

$$t.C_2\,F(t) > nC_1 + C_2 \sum_{x=1}^{t-1} F(x)$$

$$C_2\,F(t) > \left\{nC_1 + C_2 \sum_{x=1}^{t-1} F(x)\right\} \Big/ t \qquad ...(18)$$

Similarly for $\Delta\left\{\frac{C(t)}{t}\right\} = \frac{C(t+1)}{t+1} - \frac{C(t)}{t} > 0$, the following condition can be derived

$$C_2F(t-1) < \left\{nC_1 + C_2 \sum_{x=1}^{t-1} F(x)\right\} \Big/ t-1 \qquad ...(19)$$

Inequalities (18) and (19) describe the necessary condition for optimal group replacement. In Eq. (18), the expression:

$$\left\{nC_1 + C_2 \sum_{x=1}^{t-1} F(x)\right\} \Big/ t$$

represents the average cost per period if all items are replaced at the end of period *t*. Whereas, expression $C_2 + F(t)$ represents the cost for the *t*th period if group replacement is not made at the end of period *t*.

Example 7: (a) *At time zero, all items in a system are new. Each item has a probability p of failing immediately before the end of the first month of life, and a probability q = 1 – p of failing immediately before the end of the second month (i.e. all items fail by the end of the second month). If all items are replaced as they fail, then show that the expected number of failures f(x) at the end of month x is given by:*

$$f(x) = \frac{N}{1+q}\,[-(-q)^{x+1}]$$

Where ***N*** *is the number of items in the system.*

(b) *If the cost per item of individual replacement is C_1, and the cost per item of group replacement is C_2, find the conditions under which (i) a group replacement policy at the end of each month is most profitable; (ii) no group replacement policy is better than that of pure individual replacement.*

Solution: (a) Let N_i, be the expected number of items to fail at the end of the *i*th month. Then:

N_0 = number of items in the system in the beginning (= N)

N_1 = expected number of items to fail at the end of the first month

= $N_0 p = N(1 - q)$, since $p = 1 - q$

N_2 = expected number of items to fail at the end of the second month

$$= N_0 q + N_1 p = N q + N_1(1 - q)$$

$$= Nq + N(1 - q)^2 = N(1 - q + q^2)$$

N_3 = expected number of items to fail at the end of the third month

$$= N_0 q + N_1 q + N_2 p = N q + N q(1 - q) + N(1 - q + q^2)(1 - q)$$

$$= N(1 - q + q^2 - q^3)$$

and so on. In general

$$N_k = N\{1 - q + q^2 - q^3 + ... + (- q)^k\},$$

$$N_{k+1} = N_{k-1} q + N_k p$$

$$= N\{1 - q + q^2 + ... + (- q)^{k-1}\} q + N\{1 - q + q^2 + ... + (- q)^k\}(1 - q)$$

$$= N\{1 - q + q^2 + ... + (- q)^{k+1}\}$$

By mathematical induction, the expected number of items to fail, $f(x)$ at the end of month x is given by

$$f(x) = N\{1 - q + q^2 + ... + (- q)^x\} = \frac{N\{1 - (- q)^{x+1}\}}{1 + q}$$

(sum of G.P. of $x + 1$ terms with common ratio $- q$)

(b) The value of $f(x)$ at the end of month x will vary for different values of $(- q)^{x+1}$ and it will reach in steady state as $x \to \infty$. Hence, in the steady-state the expected number of failures becomes:

$$\lim_{x \to \infty} f(x) = \lim_{x \to \infty} \frac{N}{1 + q} \{1 - (- q)^{x+1}\}$$

$$= \frac{N}{1 + q}, \quad q < 1 \quad \text{and} \quad (- q)^{x+1} \to 0 \text{ as } x \to \infty$$

where $(1 + q)$ represents the mean age at failure and is given by $p + 2q = (1 - q) + 2q = 1 + q$.

Since C_2 is the cost of replacement per item, individually, the average cost per month for an individual replacement policy will be $NC_2/(1 + q)$.

(i) The average cost for group replacement policy at the end of every month is given by:

$$NC_1 + Np\, C_2 = NC_1 + N(1 - q)C_2$$

A group replacement policy at the end of each month is most profitable, when:

$$NC_1 + Np\, C_2 < \frac{NC_2}{1 + q}, \quad i.e. \quad C_1 < \frac{q^2}{1 + q} C_2 \quad \text{or} \quad C_2 > \frac{1 + q}{q^2} C_1$$

(ii) The individual replacement policy is always better than any group replacement policy

Check Your Progress

State whether the following statements are True or False:

6. Mortality tables are used to derive the probability distribution of life span of an equipment in question.
7. According to Descrate's sign rule all other roots $\alpha_1, \alpha_2, ..., \alpha_k$ will be positive and their absolute value will be more than unity.
8. Individual replacement policy ensures smooth running of the system.
9. The value of money that increases with constant rate is known as its depreciation ratio.
10. Present worth factor is present value of one rupee spent in *n* years from time now onwards.

19.5 OTHER REPLACEMENT PROBLEMS

A few replacement problems that are different from those discussed earlier in this unit are as follows:

Staffing Problem

The principles of replacement may also be applied to formulate some useful recruitment and promotion policies for the staff working in an organization. To apply the principles of replacement is such a case, it is assumed that the life distribution for the service of staff in the organization is already known.

Example 8: *An airline requires 200 assistant hostesses, 300 hostesses, and 50 supervisors. Women are recruited at the age of 21, and if still in service retire at 60. Given the following life table, determine:*

(a) *How many women should be recruited each year?*

(b) *At what age should the women be promoted?*

Airline Hostesses' Life Record

Age	***21***	***22***	***23***	***24***	***25***	***26***	***27***	***28***
No. in Service	***1,000***	***600***	***480***	***384***	***307***	***261***	***228***	***206***
Age	29	30	31	32	33	34	35	36
No. in Service	109	181	173	167	161	155	150	146
Age	37	38	39	40	41	42	43	44
No. in Service	141	136	131	125	119	113	106	99
Age	45	46	47	48	49	50	51	52
No. in Service	93	87	80	73	66	59	53	46
Age	53	54	55	56	57	58	59	
No. in Service	39	33	27	22	18	14	11	

Solution: If 1,000 women had been recruited each year for the past 39 years, then the total number of them recruited at the age of 21 and those serving up to the age of 59 is 6,480. Total number of women recruited in the airline are: 200 + 300 + 50 = 550.

(a) Approx 550 × (1,000/6,480) = 85 new hostesses are to be recruited every year in order to maintain a strength of 550 hostesses.

(b) If the assistant hostesses are promoted at the age of x, then up to age $(x – 1)$, 200 assistant hostesses will be required. Since among a total of 550 hostesses, 200 are assistant hostesses, therefore, out of a strength of 1,000 hostesses there will be: 200 × (1,000/550) = 364 assistant hostesses. But from the life table, this number is available up to the age of 24 years. Thus, the promotion of assistant hostesses is due in the 25th year.

Since out of the 550 recruitments only 300 hostesses are needed, if 1,000 girls are recruited, then only 1,000 × (300/550) = 545 (approx.) will be hostesses. Hence, the total number of hostesses and assistant hostesses in a recruitment of 100 will be: 545 + 364 = 909. This means, only 1,000 – 909 = 91 supervisors are required. But from the life table this number is available up to the age of 46 years. Thus, the promotion hostesses to supervisors will be due in the 47th year.

Example 9: *It is planned to raise a research team to a strength of 50 chemists, which is to be maintained. The wastage of recruits depends on their length of service which is as follows:*

Year	:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Total percentage who have left by end of year	:	*5*	*36*	*55*	*63*	*68*	*73*	*79*	*87*	*97*	*100*

What is the required number of recruits recruitment per year necessary to maintain the required strength? There are 8 senior posts for which the length of service is the main criterion. What is the average length of service after which the new entrant expects promotion to one of these posts?

Solution: The probability of a chemist being in service at the end of the year can be calculated with the help of the given data as shown in Table 19.7.

Table 19.7 shows that if 100 chemists are recruited each year, then the total number of chemists present the end of the year will be 436. Thus, to maintain a strength of 50 chemists in the organization, [(100/436) × 50 = 12] chemists have to be recruited each year.

Table 19.7: Probability of Chemists in Service

Year	Number of chemists who left at the end of the year	Number of chemists in service at the end of the year	Probability of leaving at the end of year	Probability of in service at the end of the year
1	0	100	0	1.00
2	5	95	0.05	0.95
3	36	64	0.36	0.64
4	56	44	0.56	0.44
5	63	37	0.63	0.37
6	68	32	0.68	0.32
7	73	27	0.73	0.27
8	79	21	0.79	0.21
9	87	13	0.87	0.13
10	97	3	0.97	0.03
	100	0	1.00	0
		436		

If P_n is the probability of a person to be in service at the end of the year, then out of 12 new recruits as calculated above, the number of survivals (chemists who will remain in service) at the end of the year n will be 12. p_n. Thus, a table, as shown below, can be constructed to show the number of chemists in service he end of each year.

Year *(n)*	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Probability (p_n)	1.00	0.95	0.64	0.44	0.37	0.32	0.27	0.21	0.13	0.03	0
Number of chemists 12 (p_n)	12	11	8	5	4	4	3	2	2	0	0

This table shows that there are 3, 2, and 2 persons in service during 6th, 7th and 8th year, respectively. The total of such chemists is less than the number of senior posts, *i.e.* 8. Hence, promotions of the new chemist must start by the end of the 5th year.

Equipment Renewal Problem

The term *renewal* refers to either replacing an item (machine or equipment) by new or repairing it so that the probability density function of its future life time is equivalent to that of a new item. The future lifetime of the item is considered to be a random variable.

Definition: The probability that an item will need a renewal in the interval t to $t + dt$ is called the *renewal rate*, at time t, provided it is in running order at age t. This is given by $r(t)\ dt$ (also called *renewal density function*).

Example 10: *A certain piece of equipment is extremely difficult to adjust. During a period when no adjustment is made, the running cost increases linearly with time, at a rate of b rupees per hour. The running cost immediately after an adjustment is not precisely known until the adjustment has been made. Before the adjustment, the resulting running cost x is a random variable x with density function f(x). If each adjustment costs k rupees, when should the replacement be made?*

Solution: The running cost ₹ x is a random variable with density function $f(x)$. Suppose that the maximun of x be X.

If the adjustment is made when the running cost equals Z, then there can be two possibilities:

$$(i)\ Z > X \quad \text{and} \quad (ii)\ Z < X$$

Case 1 (When Z > X): Let ₹ x be the running cost at time $t = 0$. If the adjustment is made after t, then the running cost at time t will be ₹ $(x + bt)$, because the running cost increases at the rate of ₹ b per hour. Obviously,

$$Z = x + bt \quad \text{or} \quad t = \frac{(Z - x)}{b}$$

If C(Z) is the total cost incurred between the period of one adjustment and another, then:

C(Z) = Cost of one adjustment + Total running cost from $t = 0$ to $t = (Z - x)/b$

$$= k + \int_0^{(Z-x)/b} (x + bt)\, dt = k + \left[\frac{(x+bt)^2}{2b}\right]_0^{(z-x)/b} = k + \frac{1}{2b}(Z^2 - x^2)$$

Therefore, the average cost per hour is given by:

$$\text{Average cost per hour} = \frac{C(Z)}{t} = \frac{kb}{Z - x} + \frac{Z + x}{2}$$

Since the running cost x is a random variable with density function, therefore, the expected cost per hour is given by:

$$E\{C(Z)\} = \int_0^x \left(\frac{kb}{Z - x} + \frac{Z + x}{2}\right) f(x)\, dx$$

The value of E{C(Z)} will be minimum for some value of Z, for which:

$$\frac{d}{dZ}\,[E\{C(Z)\}] = 0 \quad \text{and} \quad \frac{d^2}{dZ^2}\,[E\{C(Z)\}] > 0$$

$$\text{Now,} \quad \frac{d}{dZ}\,[E\{C(Z)\}] = \frac{d}{dZ}\int_0^x \left(\frac{kb}{Z - x} + \frac{Z + x}{2}\right) f(x)\, dx$$

$$= \int_0^x \frac{d}{dZ}\left(\frac{kb}{Z-x}+\frac{Z+x}{2}\right) f(x)\, d(x)$$

$$= \int_0^x \left\{\frac{kb}{(Z-x)^2}+\frac{1}{2}\right\} f(x)\, dx$$

$$= \frac{1}{2} - kb\int_0^x \frac{f(x)}{(Z-x)^2}\, dx\ ;\ \int_0^x f(x)\, dx = 1$$

For E{C(Z)} to be minimum, we must have:

$$\frac{d}{dZ}\ [\mathrm{E}\{\mathrm{C}(Z)\}] = 0$$

which gives $\frac{1}{2} - kb\int_0^x \frac{f(x)}{(Z-x)^2}\, dx = 0$ or $\int_0^x \frac{f(x)}{(Z-x)^2}\, dx = \frac{1}{2kb}$ ...(23)

Hence, the value of Z can be determined with the help of Eq. (23).

Case 2 (**When Z < X**): In this case, it can be shown that the minimum of Z cannot occur and therefore, his optimal value can only be determined in Case (*i*).

Example 11: *A piece of equipment can either completely fail, in which case it has to be scrapped (no salvage value), or may suffer a minor defect which can be repaired. The probability that it will not have to be scrapped before age t is f(t). The conditional probability that it will need a repair in the instant t + dt knowing that it was in running order at t, is r(t)dt. The probability of a repair or complete failure is dependent only on the age of the equipment, and not on the previous repair history. Each repair costs ₹ C, and complete replacement costs ₹ K. For some considerable time, the policy has been to replace only on failure.*

(a) *Derive a formula for the expected cost per unit time of the present policy of replacing only on failure.*

(b) *It has been suggested that it might be cheaper to scrap equipment at some fixed age T, thus avoiding the higher risk of repairs with advancing age. Show that the expected cost per unit time of such policy is*

$$\frac{\left[C\int_0^T f(u)\, rdu + K\right]}{\int_0^T f(u)\, du}$$

Solution: The probability that the equipment will have to be scrapped before age *t* is *f*(*t*). Therefore, the equipment will fail at sometime and is given by:

$$\int_0^\infty f(t)\, dt = 1$$

NOTES

Further, the probability that the equipment will need renewal in the interval t and $t + dt$, knowing that it was in running order at time t is given by $r(t)\,dt$.

(a) Probability that equipment needs repair between age u and $u + du$ is $f(u)\,du$. Thus, the expected cost and total expected cost of repair will be:

$$\text{Expected cost} = C\int_0^{\infty} f(u)\,r(u)\,du$$

$$\text{Total expect cost} = \frac{\left[K + C\int_0^{\infty} f(u)\,r(u)\,du\right]}{\int_0^{T} f(u)\,du}$$

$$= K + C\int_0^{\infty} f(u)\,r(u)\,du\;;\; \int_0^{\infty} f(u)\,du = 1$$

(b) Under the policy of scrapping at the age T, the total expected cost of repair will be:

$$K + C\int_0^{T} f(u)\,r(u)\,du$$

Hence, the expected cost per unit of time will be:

$$E(T) = \frac{\left[K + C\int_0^{T} f(u)\,r(u)\,du\right]}{\int_0^{T} f(u)\,du}$$

19.6 INTRODUCTION: SEQUENCING PROBLEMS

Every organization wants to utilize its productive systems effectively and efficiently and wants to maximize its profit by meeting the delivery deadlines. A number of jobs involving many operations have to be performed and there are limited resources in terms of plant and machinery on which the jobs have to be performed. It is necessary that available facilities are optimally utilized and they are loaded, scheduled and sequenced properly.

A sequence is the order in which different jobs are to be performed. When there is a choice that a number of tasks can be performed in different orders, then the problem of sequencing arises. Such situations are very often encountered by manufacturing units, overhauling of equipments or aircraft engines, maintenance schedule of a large variety of equipment used in a factory, customers in a bank or car servicing garage and so on.

The basic concept behind sequencing is to use the available facilities in such a manner that the cost (and time) is minimized. The sequencing theory has been developed to

solve difficult problems of using limited number of facilities in an optimal manner to get the best production and minimum costs.

Terms Commonly Used

1. ***Job:*** These have to be sequenced, hence there should be a particular number of jobs (groups of tasks to be performed) say *n* to be processed.
2. ***Machine:*** Jobs have to be performed or processed on machines. It is a facility which has some processing capability.
3. ***Loading:*** Assigning of jobs to facilities and committing of facilities to jobs without specifying' the time and sequence.
4. ***Scheduling:*** When the time and sequence of performing the job is specified, it is called *scheduling*.
5. ***Sequencing:*** Sequencing of operations refers to a systematic procedure of determining the order in which a series of jobs will be processed in a definite number, say k, facilities of machines.
6. ***Processing time:*** Every operation that is required to be performed requires definite amount of time at each facility or machine when processing time is definite and certain, scheduling is easier as compared to the situation in which it is not known.
7. ***Total Elapsed time:*** It is the time that lapses between the starting of first job and the completion of the last one.
8. ***Idle Time:*** The time for which the facilities or machine are not utilized during the total elapsed time.
9. ***Technological order:*** It is the order which must be followed for completing a job. The requirement of the job dictates in which order various operations have to be performed, for example, painting cannot be done before welding.
10. ***Passing not allowed:*** If '*n*' jobs have to be processed through '*m*' machines in a particular order of M_1, M_2, M_3 then each job will go to machine M_1 first and then to M_2 and finally to M_1. This order cannot be passed.
11. ***Static arrival pattern:*** If all the jobs to be done are received at the facilities simultaneously.
12. ***Dynamic arrival pattern:*** Here the jobs keep arriving continuously.

Assumptions

In sequencing problems, the following assumptions are made:

(i) All machines can process only one job at a time.

(ii) No time is wasted in shifting a job from one machine to other.

(iii) Processing time of job on a machine has no relation with the order in which the job is processed.

(iv) All machines have different capability and capacity.

(v) All jobs are ready for processing.

(vi) Each job when put on the machine is completed.

(vii) All jobs are processed in specified order as soon as possible.

Check Your Progress

Choose the correct option for the following statements:

1. Group replacement policy is suitable for large number of identical ..... that are likely to fail with age.
   (a) high cost items (b) low cost items
   (c) group of items (d) given items
2. Future lifetime of an item is considered to be a .....
   (a) continuous variable (b) finite variable
   (c) infinite variable (d) random variable
3. ..... is a facility that has some processing capability.
   (a) Job (b) Machine
   (c) Equipment (d) Customer
4. The basic concept behind sequencing is to use the available facilities in such a way that the cost is .....
   (a) maximized (b) minimized
   (c) doubled (d) halved
5. When the time and sequence of performing the job is specified it is called ..... .
   (a) loading (b) scheduling
   (c) processing (d) sequencing

19.7 TYPES OF SEQUENCING PROBLEMS

The following types of sequencing problems will be discussed in this unit:

(a) *n* jobs one machine case

(b) *n* jobs two machines case

(c) *n* jobs '*m*' machine case

(d) Two jobs '*m*' machines case.

The solution of these problems depends on many factors such as :

(a) The number of jobs to be scheduled

(b) The number of machines in the machine shop

(c) Type of manufacturing facility (slow shop or fast shop)

(d) Manner in which jobs arrive at the facility (static or dynamic)

(e) Criterion by which scheduling alternatives are to be evaluated.

As the number of jobs (n) and the number of machines (m) increases, the sequencing problems become more complex. In fact, no exact or optimal solutions exist for sequencing problems with large n and m. Simulation seems to be a better solution technique for real life scheduling problems.

n-*Jobs One Machine Case*

This case of a number of jobs to be processed on one facility is very common in real life situations. The number of cars to be serviced in a garage, number of engines to be overhauled in one workshop, number of patients to be treated by one doctor, number of different jobs to be machined on a lathe, etc, are the cases which can be solved by using the method under study. In all such cases we are all used to '*first come first served*' principle to give sense of satisfaction and justice to the waiting jobs. But if this is not the consideration, it is possible to get more favourable results in the interest of effectiveness and efficiency. The following assumptions are applicable:

(a) The job shop is static.

(b) Processing time of the job is known.

The implication of the above assumption that job shop is static will mean that new job arrivals do not disturb the processing of n jobs already being processed and the new job **arrivals** wait to be attended to in next batch.

Shortest Processing Time (SPT) Rule

This rule says that jobs are sequenced in such a way that the job with least processing time is picked up first, followed by the job with the next Smallest Processing Time (SPT) and so on. This is referred to as *shortest processing time sequencing*. However, when the importance of the jobs to be performed varies, a different rule called Weight-Scheduling (Weight Scheduling Process Time) rule is used. Weights are allotted to jobs, greater weight meaning more important job. Let W_i be the weight allotted. By dividing the processing time by the weight factor, the tendency to move important job to an earlier position in the order is achieved.

$$\text{Weighted Mean Flow Time, (WMFT)} = \frac{\sum_{i=1}^{n} W_i f_i}{\sum_{i=1}^{n} W_i}$$

where f_i = flow time of job $i = W_i + t_i$

t_i = processing time of job i

WSPT rule for minimizing Weighted Mean-Flow Time (WMFT) puts n jobs in a sequence such that

$$\frac{t[1]}{W[1]} \leq \frac{t[2]}{W[2]} \leq \ldots \leq \frac{t[n]}{W[n]}$$

The numbers in brackets above define the position of the jobs in the optimal sequence.

NOTES

Example 12: *Consider the 8 jobs with processing times, due dates and importance weights as shown below.*

8 jobs one machine case data

Task (*i*)	Processing time (t_i)	Due data (d_i)	Importance weight (W_i)	$\frac{t_i}{W_i}$
1	5	15	1	5.0
2	8	10	2	4.0
3	6	15	3	2.0
4	3	25	1	3.0
5	10	20	2	5.0
6	14	40	3	4.7
7	7	45	2	3.5
8	3	50	1	3.0

From processing time t_i in the table the SPT sequence is 4–8–1–3–7–2–5–6 resulting completion of these jobs at times 3, 6, 14, 20, 27, 36, 46, 60 respectively.

$$\text{WMFT} = \frac{3+6+14+20+27+36+46+60}{8} = 26.5 \text{ hours}$$

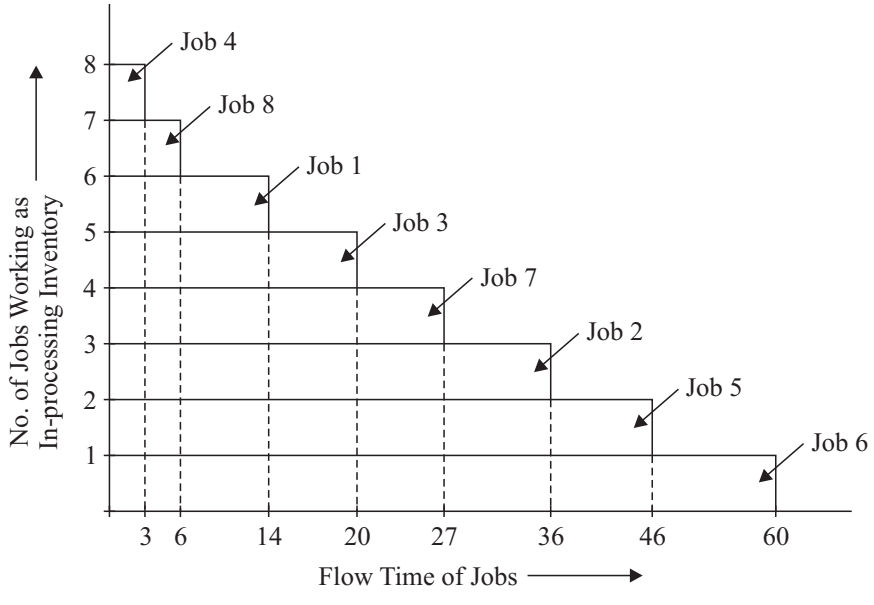


Fig. 19.1

The sequence is shown graphically above from which the number of tasks waiting as in process inventory is seen to be 8 during 0–3, 7 during 3–6, 6 during

6–14, 5 during 14–20, 4 during 20–27, 3 during 27–36, 2 during 36–46 and one during 46–60, Thus, the average inventory is given by

$$\text{Average inventory} = \frac{8\times3+7\times3+6\times8+5\times6+4\times7+3\times9+2\times10+1\times14}{60}$$

$$= \frac{24+21+48+30+28+27+20+14}{60} = \frac{212}{60} = 3.53 \text{ jobs.}$$

Weight Scheduling Process Time

If the important weights W_i were to be considered the WSPT could be used to minimize the Weighted Mean Flow Time (WMFT) to yield the sequence 3–8–8–2–7–6–5–1. This results by first choosing job with minimum $\frac{t_i}{W_i}$ in the table. The respective flow time of jobs in this sequence are 6, 9, 12, 21, 28, 42, 52, 58. Mean flow time is hours

$$\text{WMFT} = \frac{6\times3+9\times1+12\times1+21\times3+28\times2+42\times3+52\times2+58\times1}{3+1+1+3+2+3+2+1}$$

$$= \frac{18+9+12+63+56+126+104+58}{16} = \frac{446}{16} = 27.85 \text{ hours}$$

Example 13: *Eight jobs A, B, C, D, E, F, G and H arrive at one time to be processed on a single machine. Find out the optimal job sequence, when their operation time is given in the table below.*

Job (n)	*Operation time in minutes*
A	*16*
B	*12*
C	*10*
D	*8*
E	*7*
F	*4*
G	*2*
H	*1*

Solution: For determining the optimal sequence, the jobs are selected in a non-descending operation time as follows:

Non-decreasing operation time sequence is $H \to G \to F \to E \to D \to C \to B \to A$.

Total processing time

H = 1

G = 1 + 2 = 3

F = 1 + 2 + 4 = 7

E = 1 + 2 + 4 + 7 = 14

D = 1 + 2 + 4 + 7 + 8 = 22

C = 1 + 2 + 4 + 7 + 8 +10 = 32

B = 1 + 2 + 4 + 7 + 8 + 10 + 12 = 44

A = 1 + 2 + 4 + 7 + 8 + 10 + 12 + 16 = 60

Average processing time = Total time/number of jobs = 183/8 = 23 minutes

In case the jobs are processed in the order of their arrival, *i.e.*, A → B → C → D → E → F → G → H the total processing time would have been as follows :

A = 16

B = 16 + 12 = 28

C = 16 + 12 + 10 = 38

D = 16 + 12 + 10 + 8 = 46

E = 16 + 12 + 10 + 8 + 7 = 53

F = 16 + 12 + 10 + 8 + 7 + 4 = 57

G = 16 + 12 + 10 + 8 + 7 + 4 + 2 = 59

H = 16 + 12 + 10 + 8 + 7 + 4 + 2 + 1 = 60

Average processing time = 357/8 = 44.6, which is much more than the previous time.

Priority Sequencing Rules

The following priority sequencing rules are generally followed in production/service system.

1. **First Come First Served** (FCFS)**:** As explained earlier, it is followed to avoid any heart burne and avoidable controversies.
2. **Earliest Due Date** (EDD)**:** In this rule, top priority is allotted to the waiting job, which has the earliest due/delivery date. In this case the order of arrival of the job and processing time it takes is ignored.
3. **Least Slack Rule** (LS)**:** It gives top priority to the waiting job whose slack time is the least. Slack time is the difference between the length of time remaining until the job is due and the length of its operation time.
4. **Average Number of Jobs in the System:** It is defined as the average number of jobs remaining in the system (waiting or being processed) from the beginning of sequence through the time when the last job is finished.
5. **Average Job Lateness:** Jobs lateness is defined as the difference between the actual completion time of the job and its due date. Average job lateness

is sum of lateness of all jobs divided by the number of jobs in the system. This is also called *Average job Tardiness.*

6. **Average Earliness of Jobs:** If a job is completed before its due date, the lateness value is negative and the magnitude is referred as earliness of job. Mean earliness of the job is the sum of earliness of all jobs divided by the number of jobs in the system.
7. **Number of Tardy Jobs:** It is the number of jobs which are completed after the due date.

Sequencing *n* Jobs Through Two Machines

The sequencing algorithm for this case was developed by Johnson and is called *Johnson's Algorithm.* In this situation *n* jobs must be processed through machines M_1 and M_2. The processing time of all the *n* jobs on M_1 and M_2 is known and it is required to find the sequence, which minimizes the time to complete all the jobs.

Johnson's algorithm is based on the following assumptions:

(i) There are only two machines and the processing of all the jobs is done on both the machines in the same order, *i.e.*, first on M_1 and then on M_2.

(ii) All jobs arrive at the same time (static arrival pattern) have no priority for job completion.

Johnson's algorithm involves following steps:

1. List operation time for each job on machine M_1 and M_2.
2. Select the shortest operation or processing time in the above list.
3. If minimum-processing time is on M_1, place the corresponding job first in the sequence. If it is on M_2, place the corresponding job last in the sequence. In case of tie in shortest processing time, it can be broken arbitrarily.
4. Eliminate the jobs which have already been sequenced as result of step 3.
5. Repeat steps 2 and 3 until all the jobs are sequenced.

Example 14: *Six jobs are to be sequenced, which require processing on two machines M_1 and M_2. The processing time in minutes for each of the six jobs on machines M_1 and M_2 is given below. All the jobs have to be processed in sequence M_1, M_2. Determine the optimum sequence for processing the jobs so that the total time of all the jobs is minimum. Use Johnson's algorithm.*

Jobs		*1*	*2*	*3*	*4*	*5*	*6*
Processing	*Machine M_1*	*30*	*30*	*60*	*20*	*35*	*45*
Time	*Machine M_2*	*45*	*15*	*40*	*25*	*30*	*70*

NOTES

Solution:

Step I. Operation time or processing time for each jobs on M_1 and M_2 is provided in the question.

Step II. The shortest processing time is 15 for job 2 on M_2.

Step III. As the minimum processing time is on M_2, job 2 has to be kept last as follows:

					2

Step IV. We ignore job 2 and find out the shortest processing time of rest of jobs. Now the least processing time is 20 minutes on machine M_1 for job 4. Since it is on M_1, it is to be placed first as follows:

4					2

The next minimum processing time is 30 minutes for job 5 on M_2 and Job 1 on M_1. So, job 5 will be placed at the end. Job 1 will be sequenced earlier as shown below.

4	1			5	2

The next minimum processing time is 40 minutes for job 3 on M_2, hence it is sequenced as follows:

4	1		3	5	2

Job 6 has to be sequenced in the gap or vacant space. The complete sequencing of the jobs is as follows.

4	1	6	3	5	2

The minimum time for six jobs on machine M_1 and M_2 can be shown with the help of a Gantt chart as shown below.

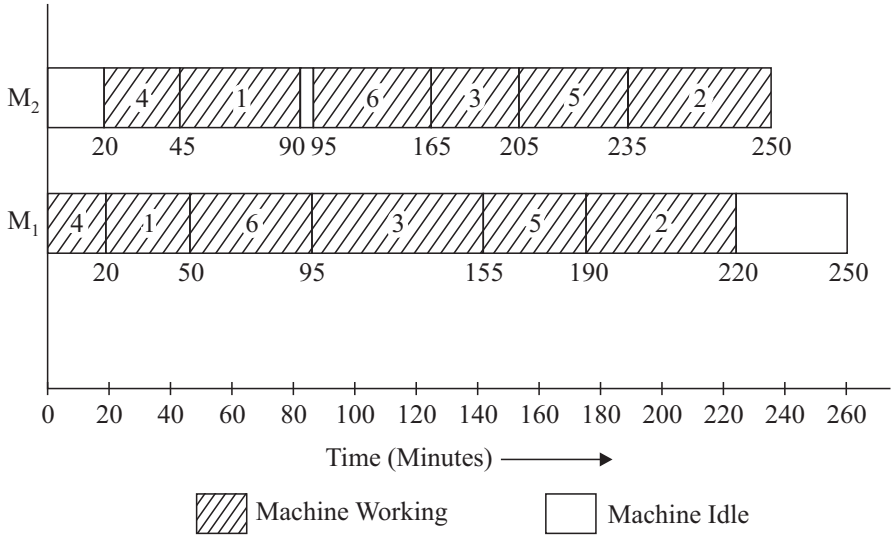


Fig. 19.2

The above figure shows idle time for M_1 (30 minutes) after the last job (2) has been processed. Idle time for M_2 is 20 minutes before job 4 is started and 5 minutes before processing 6 and finishing job 1. The percentage utilization of M_1 = 250 – 30/250 = 88% and M_2 = 250 – 25/250 = 90%.

'n' Jobs 'm' Machine Case

Let there be '*n*' jobs 1, 2, 3, ... *n* and '*m*' machine M_1, M_2, M_3 ..... *m.* The order of processing is M_1, M_2, M_3 ..... *m* and no passing is permitted. The processing time for the machine is shown below.

Job	M_1	M_2	M_3	**m**
1	a_1	b_1	c_1	M_1
2	a_2	b_2	c_2	M_2
3	a_3	b_3	c_3	M_3
: :	: :	: :	: :	: :
n	a_n	b_n	c_n	M_n

If the following conditions are used, we can replace '*m*' machines by an equivalent of two machines problem:

(a) Minimum $a_i \geq$ maximum of M_2, M_3 ..... $(m-1)$

(b) Minimum $m \geq$ maximum M_2, M_3 ..... $(m-1)$

when $M_1' = a + b_i + c_i + \ldots\ldots + (m-1)_i$

$M_2' = b_i + c_i + \ldots\ldots + (m-1)_i + m_i$

Example 15: *Determine the optimal sequence of performing 5 jobs on 4 machines. The machines are used in the order M_1, M_2, M_3 and M_4 and the processing time is given below:*

Job	M_1	M_2	M_3	M_4
1	*8*	*3*	*4*	*7*
2	*9*	*2*	*6*	*5*
3	*10*	*6*	*6*	*8*
4	*12*	*4*	*1*	*9*
5	*7*	*5*	*2*	*3*

Solution:

Step I. Let us find out if any of the conditions stipulated is satisfied or not.

Condition 1. Minimum $a_i \geq$ maximum of M_2 and M_3.

Minimum $a_i = 7$

Maximum $b_i = 6$

Maximum $c_i = 6$

Hence, the condition is satisfied.

Step II. Let us form the matrix of new processing time by creating two fictitious machines M_1' and M_2'.

Job	$M_1' = a_i + b_i + c_i$	$M_2' = b_i + c_i + d_i$
1	15 (8 + 3 + 4)	14 (3 + 4 + 7)
2	17 (9 + 2 + 6)	13 (2 + 6 + 5)
3	22 (10 + 6 + 6)	20 (6 + 6 + 8)
4	17 (12 + 4 + 1)	14 (4 + 1 + 9)
5	14 (7 + 5 + 2)	10 (5 + 2 + 3)

Step III. Now solve 5 jobs 2 machine problem.

Minimum time of processing is for job 5 on machine M_2' so it will be sequenced last

				5

Next minimum time is 13 for jobs 2 on machine M_2' so it will be sequenced as shown

			2	5

Next minimum time is for jobs 1 and 4 on machine M_2' so it will be sequenced as shown.

	1	4	2	5

Next minimum time is 20 for job 3 on machine M_2'.

3	1	4	2	5

Two Jobs 'm' Machines Case

1. Two axis to represent job 1 and 2 are drawn at right angles to each other. Same scale is used for X and Y-axes. X-axis represents the processing time and sequence of job 1 and Y-axis represents the processing time and sequence of job 2. The processing time on machines are laid out in the technological order of the problem.

2. The area which represents processing times of jobs 1 and 2 and is common to both the jobs is shaded. As processing of both the jobs on it machine is not feasible, the shaded area represents the unfeasible region in the graph.
3. The processing of both the jobs 1 and 2 are represented by a continued path which consists of horizontal, vertical and 45 degree diagonal region. The path starts at the lower left corner and stops at upper right corner and the shaded area is avoided. The path is not allowed to pass through shaded area which as brought out in step II represents both the jobs being processed simultaneously on the same machine.

   Any vertical movement represents that job 2 is in progress and job 1 is waiting to be processed. Horizontal movement along the path indicates that job 1 is in progress and job 2 is idle waiting to be processed. The diagonal movement of the path indicates that both the jobs are being processed on different machines simultaneously.
4. A feasible path maximizes the diagonal movement minimizes the total processing time.
5. Minimum elapsed time for any job = processing time of the job + idle time of the same job.

Example 16: *The operation time of two jobs 1 and 2 on 5 machines M_1, M_2, M_3, M_4 and M_5 is given is the following table. Find out the optimum sequence in which the jobs should be processed so that the total time used is minimum. The technological order of use of machine for job 1 is M_1, M_2, M_3, M_4 and M_5 for job 2 is M_3, M_1, M_4, M_5 and M_2.*

Time Hours

Job	M_1	M_2	M_3	M_4	M_5
1	1	2	3	5	1
Job	M_3	M_1	M_4	M_5	M_2
2	3	4	2	1	5

Job 1 preceeds job 2 on machine M_1, job 1 preceeds job 2 on machine M_2, job 2 preceeds job 1 on machine M_3, job 1 preceeds job 2 on M_4 and job 2 preceeds job 1 on M_5.

The minimum processing time for jobs 1 and 2, total processing time for job 1 + idle time for Job 1 = 12 + 3 = 15 hours

Total processing time for job 2 + idle time for job 2 = 15 + 0 = 15 hours.

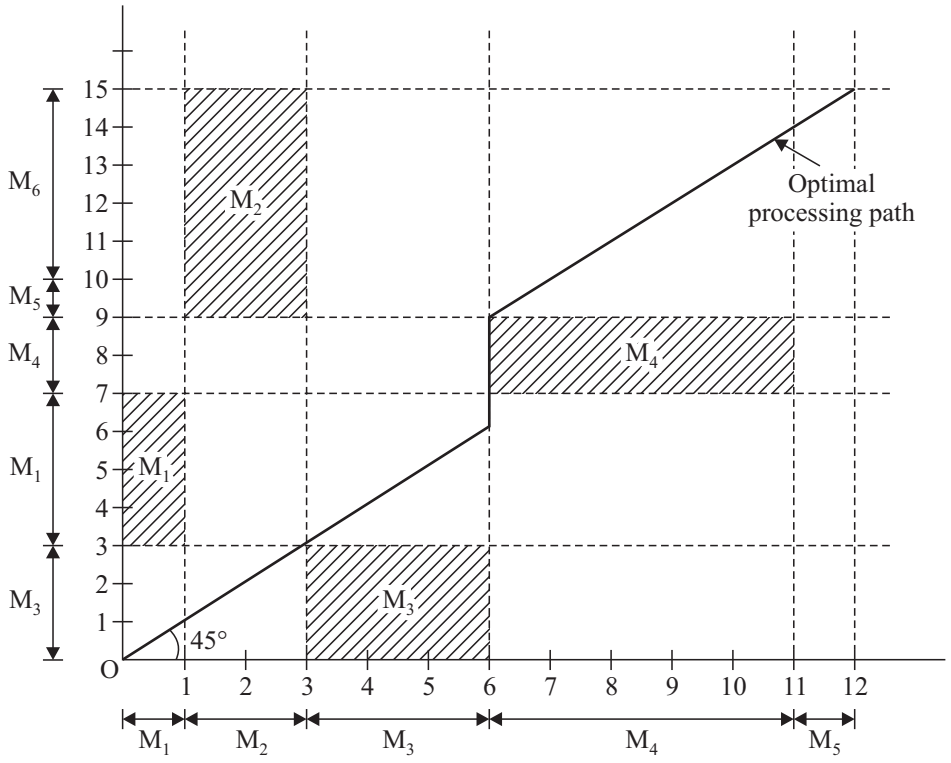


Fig. 19.3

Example 17: *Two parts A ana B for a product need processing of their operations through six machines at stations S_1, S_2, S_3, S_4, S_5 and S_6. The technological order of these parts and the manufacturing time on the machines are as given below.*

Part A	*Technological order*	S_3	S_1	S_5	S_6	S_4	S_2
	Time (hours)	*2*	*3*	*4*	*5*	*6*	*1*
Part B	*Technological order*	S_2	S_1	S_5	S_6	S_3	S_4
	Time (hours)	*3*	*2*	*5*	*3*	*2*	*3*

Determine the optimal sequencing order to minimize the total processing time for part A and B.

Solution: Let us construct the two-dimensional graph let X-axis represent job A and Y-axis represent Job B.

Total elapsed time = 23

Part A = 21 + 2 (idle time) = 23

Part B = 18 + 2 + 2 + 1 (idle time) = 23

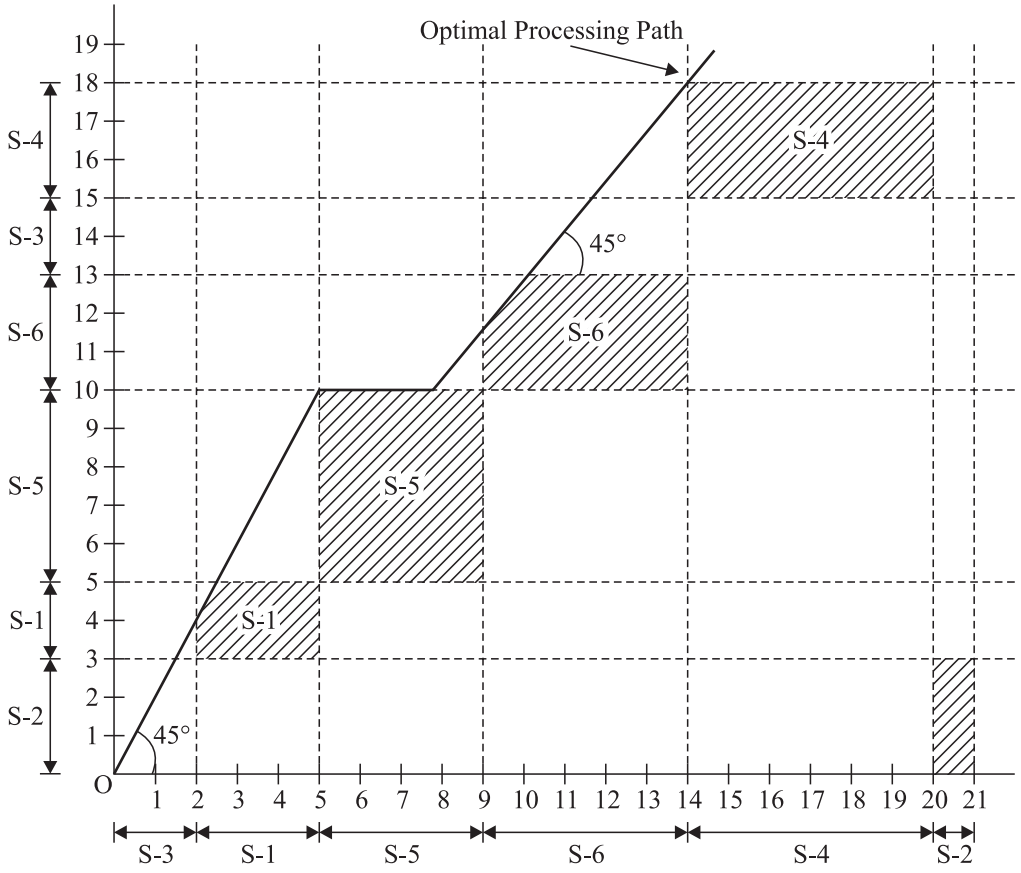


Fig. 19.4

19.8 SUMMARY

- Gradual failure is progressive in nature. That is, as the life of an item increases, its operational efficiency also deteriorates.
- This type of failure occurs in items after some period of desired service rather than deterioration while in service.
- When operational efficiency of an item deteriorates with time (gradual failure), it is economical to replace the same with a new one.
- If the effect of the time-value of money is to be considered, then replacement decision must be based upon an equivalent annual cost.
- The discounted value is the amount of money required to build up funds at compound interest that is sufficient to pay the required cost when due.
- The running cost of an equipment that deteriorates over a period of time increases and the value of the money decreases with a constant rate. If r is the interest rate, then:

$$Pwf = (1 + r)^{-n}$$

is called the *present worth factor* (*Pwf*) or present value of one rupee spent in *n* years from time now onwards.

- The replacement of items on the basis of present worth factor (*Pwf*) includes the present worth of al1 future expenditure and revenues for each replacement alternatives.
- ***Mortality Tables:*** These tables are used to derive the probability distribution of life span of an equipment in question.
- The probability that an equipment has survived to an age $(t - 1)$, and will fail during the interval $(t - 1)$ to t can be defined as the conditional probability of failure.
- Under this policy, an item (machine or equipment) is replaced individually as when it failed. This ensures smooth running of the system.
- Sometime the immediate replacement on failure of the item(s) is costly. In such cases a *group replacement policy* is preferred.
- The group replacement policy is suitable for a large number of identical low cost items that are likely to fail with age and for which it is difficult as well as not justified to keep the record of their individual ages.
- The principles of replacement may also be applied to formulate some useful recruitment and promotion policies for the staff working in an organization.
- A sequence is the order in which different jobs are to be performed. When there is a choice that a number of tasks can be performed in different orders, then the problem of sequencing arises.
- The basic concept behind sequencing is to use the available facilities in such a manner that the cost (and time) is minimized.
- This rule says that jobs are sequenced in such a way that the job with least processing time is picked up first followed by the job with the next Smallest Processing Time (SPT) and so on.
- ***Least Slack Rule (LS):*** It gives top priority to the waiting job whose slack time is the least. Slack time is the difference between the length of time remaining until the job is due and the length of its operation time.

19.9 GLOSSARY

- ***Progressive Failure:*** If the probability of failure of an item increases with the increase in its life, then such a failure is called progressive failure.
- ***Retrogressive Failure:*** If the probability of failure in the beginning of the life of an item is more but as time passes the chances of its failure becomes less, then such failure is said to be retrogressive.

- ***Random Failure:*** In this type of failure, the constant probability of failure is associated with items that fail from random causes such as physical shocks not related to age.
- ***Depriciation Ratio:*** The value of money that decreases with constant rate is known as its deprieciation ratio.
- ***Renewal Rate:*** The probability that an item will need a renewal in the interval t to $t + dt$ is called the renewal rate, at time t, provided it is in running order at age t.
- ***Loading:*** Assuming of jobs of facilities and committing of facilities to jobs without specifying the time and sequence.
- ***Static Arrival Time:*** If all the jobs to be done are received at the facilities simultaneously.
- ***Dynamic Arrival Time:*** Here the jobs keep arriving continuously.

19.10 ANSWERS TO CHECK YOUR PROGRESS

1. progressive
2. progressive failure
3. random failure
4. operational efficiency
5. gradual failure
6. True
7. False
8. True
9. False
10. True
11. (b)
12. (d)
13. (b)
14. (b)
15. (b)

NOTES

16.11 TERMINAL AND MODEL QUESTIONS

1. What is replacement? Describe some important replacement situations.
2. Suppose the cost of maintenance of a machine increases with time and its scrap value is constant.
   (a) If time is measured in continuous units, then the average annual cost will be minimized by replacing the machine when the average cost till date becomes equal to the current maintenance cost.
   (b) If time is measured in discrete units, then the average annual cost will be minimized by replacing the machine when the next period's maintenance cost becomes greater than the current average cost.
3. Describe the problem of replacement of items whose maintenance cost increase with time. Assume that the value of money remains constant.
4. What are situations that make the replacement of items necessary?
5. Explain, with examples, the failure mechanism of items.

Model I

6. The cost of a machine is ₹ 6,100 and its scrap value is ₹ 100. The maintenance costs found from experience are as follows:

Year	:	1	2	3	4	5	6	7	8
Maintenance cost (₹)	:	100	250	400	600	900	1,200	1,600	2,000

   When should the machine be replaced?

7. A truck owner finds, from his past records, that the maintenance costs per year of a truck whose purchase price is ₹ 8,000 are as given below:

Year	:	1	2	3	4	5	6	7	8
Maintenance cost (₹)	:	1,000	1,300	1,700	2,000	2,900	3,800	4,800	6,000
Resale price (₹)	:	4,000	2,000	1,200	600	500	400	400	400

   Determine what time would it be profitable to replace the truck.

8. A fleet owner finds, from his past records, that the cost per year of running a vehicle, whose purchase price is ₹ 50,000, is:

Year	:	1	2	3	4	5	6	7
Running cost (₹)	:	5,000	6,000	7,000	9,000	11,500	16,000	18,000
Resale value (₹)	:	30,000	15,000	7,500	3,750	2,000	2,000	2,000

   Thereafter, the running cost increases by ₹ 2,000, but the resale value remains constant at ₹ 2,000. At what age is a replacement due?

9. A plant manager is considering the replacement policy for a new machine. He estimates the following costs (all costs in rupees):

Year	:	1	2	3	4	5	6
Replacement cost at the beginning of year	:	100	110	125	140	160	190
Resale value at the end of year	:	60	50	40	25	10	0
Operating costs	:	25	30	40	50	65	80

Find an optimal replacement policy and its corresponding minimum cost.

10. A new tempo costs ₹ 80,000 and may be sold at the end of any year at the following prices:

Year (end)	:	1	2	3	4	5	6
Selling price (₹) (*at present value*)	:	50,000	33,000	2,000	1,100	6,000	1,000

The corresponding annual operating costs are:

Year (end)	:	1	2	3	4	5	6
Cost/year (₹) (*at present value*)	:	10,000	12.000	15,000	20,000	30,000	50,000

It is not only possible to sell the tempo after use but also to buy a second-hand tempo.

It may be cheaper to do so than to replace it with a new tempo.

Age of tempo	:	0	1	2	3	4	5
Purchase price (₹) (*at present value*)	:	80,000	58,000	40,000	26,000	16,000	10,000

Determine the time at which it is profitable to sell the tempo and to minimize its average annual cost?

Model II

11. The cost of a new machine is ₹ 5,000. The maintenance cost of *n*th year is given by $C_n = 500(n - 1)$; $n = 1, 2, \ldots$. Suppose that the discount rate per year is 0.5. After how many years will it be economical to replace the machine by a new one?

12. A manufacturer is offered two machines A and B. Machine A is priced at ₹ 5,000 and its running costs are estimated at ₹ 800 for each of the first five years increasing by ₹ 200 per year in the sixth and subsequent years. Machine B that has the same capacity as A, costs ₹ 2,500 but would have running costs of ₹ 1,200 per year for six years, increasing by ₹ 200 per year thereafter.

    If money is worth 10 per cent per year, which machine should be purchased? (Assume that the machine will eventually be sold for scrap at a negligible price.)

13. Assume that the present value of one rupee to be spent in a year's time is ₹ 0.9 and C = ₹ 3,000, capital cost of equipment. The running costs are given in the table below.

Year	:	1	2	3	4	5	6	7
Running cost (₹)	:	500	600	800	1,000	1,300	1,600	2,000

When should the machine be replaced?

14. If Mr X wishes to have a minimum rate of return of 10 per cent per annum on his investment, which out of the following two plans should be prefer?

		Plan A	*Plan B*
First cost	:	₹ 75,000	₹ 75,000
Estimated scrap value after 20 years	:	₹ 37,500	₹ 6,000
Receipts over annual disbursement	:	₹ 7,500	₹ 9,000

Pwfs at 10 per cent for 20 years = 8.514

Pwfs at 10 per cent for 20 years = 0.2472

15. A manual stamper currently valued at ₹ 1,000 is expected to last two years. It costs ₹ 4,000 per year to operate. An automatic stamper which can be purchased for ₹ 3,000 will last four years and can be operated at an annual cost of ₹ 3,000. If money carries a rate of interest of 10 per cent per annum, determine which stamper should be purchased.

16. An engineering company is offered two types of material handling equipments A and B. A is priced at ₹ 60,000, which includes the cost of installation. The costs of operation and maintenance are estimated to be ₹ 10,000 for each of the first five years, increasing every year by ₹ 3,000 per year in the sixth and subsequent years. Equipment B with rated capacity same as A, requires an initial investment of ₹ 30,000 but in terms of operation and maintenance costs more than A. These costs for B are estimated to be ₹ 13,000 per year for the first six years, increasing every year by ₹ 4,000 from the seventh year onwards. The company expects a return of 10 per cent on all its investments. Neglecting the scrap value of the equipment at the end of its economic life, determine which equipment should the company buy.

17. Find the cost per period of individual replacement of installation of 300 light bulbs, given the following:

    (a) Cost of replacing individual bulb is ₹ 3.

(b) Conditional probability of failure is given below:

Week number	:	0	1	2	3	4
Conditional probability of failure	:	0	1/10	1/3	2/3	1

18. The following failure rates have been observed for a certain type of light bulbs:

End of week	:	1	2	3	4	5	6	7	8
Prob. of failure to date	:	0.05	0.13	0.25	0.43	0.68	0.88	0.96	1.00

The cost of replacing an individual bulb is ₹ 2.25, the decision is made to replace all bulbs simultaneously at fixed intervals, and also to replace individual bulbs as they fail in service. If the cost of group replacement is 60 paise per bulb and the total number of bulbs is 1,000, what is the best interval between group replacements?

19. The following mortality rates have been observed for a special type of light bulbs:

Month	:	1	2	3	4	5
Per cent failing at the end of month	:	10	25	50	80	100

In an industrial unit there are 1,000 special type of bulbs in use It costs ₹ 10 to replace an individual bulb that has burnt out. If all bulbs were replaced simultaneously it would cost ₹ 2.50 per bulb. It is proposed to replace all bulbs at fixed intervals, whether or not they have burnt out, and to continue replacing burnt out bulbs as they fail. At what intervals of time should the manager replace all the bulbs?

20. A computer has 20,000 resistors. When any of the resistors fail, It is replaced. The cost of replacing a resistor individually is ₹ 1. If all the resistors are replaced at the same time the cost per resistor is reduced to ₹ 0.40. The percentage surviving at the end of month *t*, and the probability of failure during the month, are given below:

		0	1	2	3	4	5	6
Percentage surviving at the end of *t*	:	100	96	90	65	35	20	0
Probability of failure during month *t*	:	—	0.04	0.06	0.25	0.30	0.15	0.20

What is the optimum replacement plan?

21. Calculate the probability of a staff resignation in each year from the following survival table:

Year	Number of original staff in service at the end of year
0	1,000
1	940
2	820
3	580
4	400
5	280
6	190
7	130
8	70
9	30
10	0

22. An airline, whose staff are subject to the same survival rates as in the previous problem, currently has a staff whose ages are distributed in the following table It is estimated that for the next two years staff requirements would increase by 10 per cent per year. If women are to be recruited at the age of 21, how many should be recruited for the next year and at what age will promotions take place? How many should be recruited for the following year and at what age would promotions take place?

Assistant

Age	:	21	22	23	24	25				
Number	:	90	50	30	20	10				(Total 200)

Hostesses

Age	:	26	27	28	29	30	31	32	33	34
Number	:	40	35	35	30	28	26	20	18	16
Age	:	35	36	37	38	39	40	41		
Number	:	12	10	8	—	8	8	6		(Total 300)

Supervisors

Age	:	42	43	44	45	46	47	48	49	50
Number	:	5	4	5	3	3	3	6	2	—
Age	:	51	52	53	54	55	56	57	58	59
Number	:	—	4	3	5	—	3	2	—	2

(Total 50)

23. It is required to find the optimum replacement time of a certain type of equipment. The initial cost of equipment is C. Salvage value and repair costs are given by S(t) and R(t), respectively. The cost of capital is r per cent and T is the time period of replacement cycle.

    (i) Show that the present value of all future costs associated with a policy of equipment after T is:

$$\left(\frac{1}{1-e^{-rt}}\right)\left[C - S(t)\,e^{-rt} + \int_0^T R(t)e^{-rt}\,dt\right]$$

    (ii) The optimal value of T is given by, $R(t) - S(T) + S(t)r = r^k/(1 - e^{-t})$

    where k is the present value of the cycle.

24. What is no passing rule in a sequencing algorithm?

25. Explain the four elements that characterize a sequencing problem.

26. Explain the principal assumptions made while dealing with sequencing problems.

27. Describe the method of processing 'n' jobs through two machines.

28. Give Johnson's procedure for determining an optimal sequence for processing n items on two machines. Give justification of the rules used in the procedure.

29. Explain the method of processing *'m'* jobs on three machines A, B, C in the order ABC.

30. Explain the graphical method to solve the two jobs m-machines sequencing problem with given technological ordering for each job. What are the limitations of the method?

31. A Company has 8 large machines, which receive preventive maintenance. The maintenance team is divided into two crews A and B. Crew A takes the machine 'Power' and replaces parts according to a given maintenance schedule. The second crew resets the machine and puts it back into operation. At all times 'no passing' rule is considered to be in effect. The servicing times for each machine are given below.

Machine	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
Crew A	5	4	22	16	15	11	9	4
Crew B	6	10	12	8	20	7	2	21

    Determine the optimal sequence of scheduling the factory maintenance crew to minimize their idle time and represent it on a chart.

32. Use graphical method to find the minimum elapsed total time sequence of 2 jobs and 5 machines, when we are given the following information:

Job 1	Sequence	A	B	C	D	E
	Time (hours)	2	3	4	6	2
Job 2	Sequence	C	A	D	E	B
	Time (hours)	4	5	3	2	6

19.12 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitative Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi

UNIT 20: PERT AND CPM

Structure

20.0 INTRODUCTION

In the previous unit you have learnt about Replacement theory and sequencing problems. In this unit, you will learn about various techniques of Project management like Program. Evaluation Review Technique (PERT) and Critical Path Method (CPM). These techniques decomposes the project into a number of activities, represent the precedence relationships among activities through a network and then determine a critical path through the network.

20.1 UNIT OBJECTIVES

After going through reading this unit, you will be able to:

- Define various terms associated with PERT and CPM
- Explain the use of Network for project
- Develop simple network diagrams

- Identify critical path and project duration
- Explain various elements of crashing of projects

NOTES

20.2 PROJECT MANAGEMENT

Let us define 'Project'.

A project can be considered to be any series of activities and tasks that

(i) have a specific objective to be completed within certain specifications.

(ii) have defined start and end dates.

(iii) have funding limits and consume resources.

A number of techniques have been developed to assist in planning, scheduling and control of projects. The most popular methods are the Critical Path Method (CPM) and the Program Evaluation and Review Technique (PERT). These techniques decompose the project into a number of activities, represent the precedence relationships among activities through a network and then determine a critical path through the network.

The basic concepts are described below:

(a) **Activity**

An activity is an item of work to be done that consumes time, effort, money or other resources. It is represented by an arrow. Tail represents start and head represents end of that activity.

Fig. 20.1

(b) **Event/Node**

It represents a point time signifying the completion of an activity and the beginning of another new activity. Here beginning of an activity represents tail event and end of an activity represents head event.

(c) **Predecessor Activity**

This is an activity that must be completed immediately before the start of another activity.

(d) **Successor Activity**

Activity, which cannot be started until one or more activities are completed but immediately succeeds them is called successor activity of a project.

(e) **Dummy Activity**

This shows only precedence relationship and they do not represent any real activity and is represented by a dashed line arrow or dotted line arrow and does not consume any time. *e.g.*,

Fig. 20.2

(f) **Rules for Construction of a Network**

1. Each activity is shown by one and only one arrow.
2. There will be only one beginning node/event and only one end node/event.
3. No two activities can be identified by the same head and tail events.
4. All events/node should be numbered distinctly.
5. Time flows from left to right.

(g) **Common Errors in Network**

1. ***Loops:***

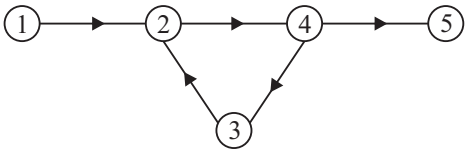


Fig. 20.3

This situation can be avoided by checking the precedence relationship of the activities and by numbering them in a logical order.

2. ***Dangling:***

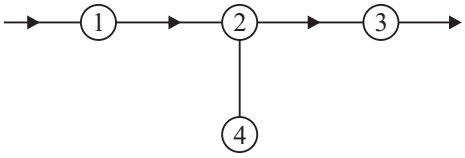


Fig. 20.4

This situation can be avoided by keeping in mind that all events except the starting and ending event of the whole project must have at least one entering and one leaving activity. A dummy activity can be introduced to avoid this dangling.

(3) ***Redundancy:***

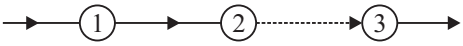


Fig. 20.5

The dummy activity is redundant and can be eliminated.

(h) ***Critical Path***

It is the longest path in the project network. Any activity on this path is said to be critical in the sense that any delay of that activity will delay the completion time of the project.

20.3 TIME CALCULATIONS IN NETWORK

Let t_{ij} be the duration of an activity (i, j).

(a) ***Earliest Start Time (ES):*** This is the earliest occurrence time of the event from which the activity emanates.

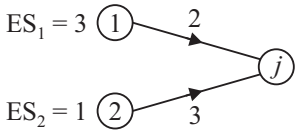
For the beginning event, $ES_1 = 0$ and let ES_i = ES of all the activities emanating from node i. Then

$$ES_j = \underset{i}{\text{Max}}\{ES_i + t_{ij}\}$$

(b) ***Earliest Finish/Completion Time (EF):*** This is the ES plus the activity duration

$$EF_i = ES_i + t_{ij}.$$

For example, Consider a part of the network.

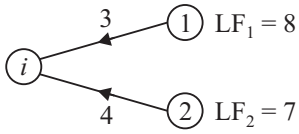


$$ES_j = \text{Max. } \{ES_1 + t_{1j}, ES_2 + t_{2j}\}$$
$$= \text{Max. } \{3 + 2, 1 + 3\} = 5$$
$$EF_1 = 3 + 2 = 5, EF_2 = 1 + 3 = 4.$$

(c) ***Latest Finish/Completion Time (LF):*** This is the latest occurrence time of the event at which the activity terminates.

$$LF_i = \underset{j}{\text{Min}}\{LF_j - t_{ij}\}$$

For example, consider a part of the network



Then

$$LF_i = \text{Min. } \{LF_1 - t_{i1}, LF_2 - t_{i2}\}$$
$$= \text{Min. } \{8 - 3, 7 - 4\} = 3.$$

(d) ***Latest Start Time (LS_i):*** This is the last time at which the event can occur without delaying the completing of the project.

(e) ***Total Floats (TF):*** It is a time duration in which an activity can be delayed without affecting the project completion time.

$$\therefore \quad TF_{ij} = LF_j - ES_i - t_{ij}$$
$$= LF_j - (ES_i + t_{ij})$$
$$= LF_j - EF_{ij}.$$

Also
$$TF_{ij} = LS_{ij} - ES_i$$
$$= (LF_j - t_{ij}) - ES_i.$$

(f) ***Free Floats (FF):*** It is a time duration in which the activity completion time can be delayed without affecting the earliest start time of immediate successor activities in the network.

$$EF_{ij} = ES_j - ES_i - t_{ij}$$
$$= ES_j - (ES_i + t_{ij})$$
$$= ES_j - EF_{ij}.$$

An activity (i, j) is said to be critical if all the following conditions are satisfied:

$$ES_i = LF_i,\ ES_j = LF_j,\ ES_j - ES_i = LF_j - LF_i = t_{ij}.$$

Thus any critical activity will have zero total float and zero free float.

(g) ***Independent Floats:*** It is defined as the difference between the free float and the tail slack.

Note: Slack is with reference to an event and float is with respect to an activity. Slack is generally used with PERT and float with CPM, but they may be used interchangeably used.

Check Your Progress

Fill in the blanks:

1. Dummy activity is represented by ..... or ..... and do not consume any time.
2. In an activity which is represented by an arrow, tail represents ..... and head represents ..... of that activity.
3. Critical path is ..... in the project network.
4. ..... is the latest occurrence time of the event at which the activity terminates.
5. Independent float is the difference between the ..... and .....

NOTES

20.4 CRITICAL PATH METHOD (CPM)

CPM was developed by E.I. duPont in 1957 and was first applied to construction and maintenance of chemical plants. Since then, the use of CPM has grown at a rapid rate. There are computer programs to perform the calculations.

Let the project network be drawn. Then this method consists of two phases calculations. In Phase 1, which is also called *forward pass*, Earliest start times (ES) of all the nodes are calculated.

In Phase 2, which is also called *backward pass*, Latest finish time (LF) of all the nodes are calculated.

These two calculations are displayed in the network diagram in a two chamber boxes. Upper chamber represents LF and the lower one as ES.

The critical activities (*i.e.,* ES = LF) are identified. The critical path is obtained by joining them using double arrow.

Example 1: *A project schedule has the following characteristics:*

Activity	*Time*	*Activity*	*Time*
1–2	*3*	*5–6*	*5*
1–3	*1*	*5–7*	*8*
2–4	*1*	*6–8*	*1*
2–5	*1*	*7–9*	*2*
3–5	*5*	*8–10*	*4*
4–9	*6*	*9–10*	*6*

Draw the project network and find the critical path. Also calculate the total floats and free floats.

Solution:

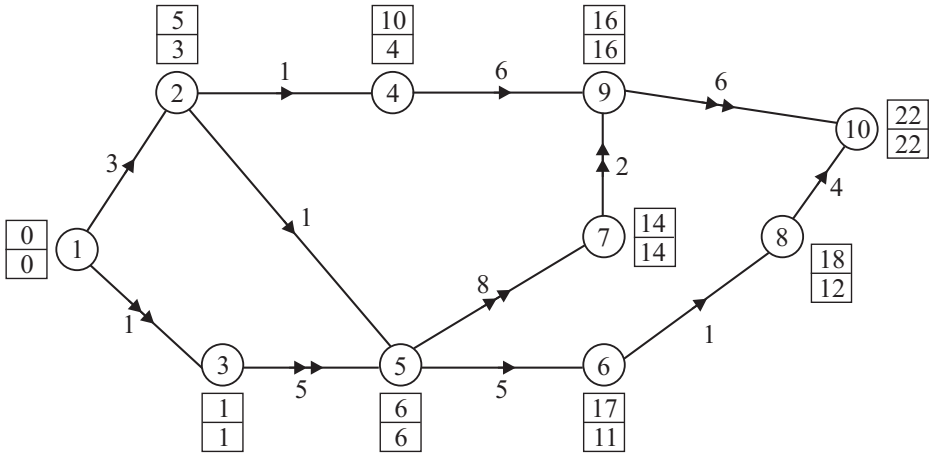


Fig. 20.6

Set $ES_1 = 0$

$ES_2 = ES_1 + t_{12} = 0 + 3 = 3$

$ES_3 = ES_1 + t_{13} = 0 + 1 = 1$

$ES_4 = ES_2 + t_{24} = 4$

$ES_5 = \text{Max. } \{ES_3 + t_{35}, ES_2 + t_{25}\} = \text{Max. } \{6, 4) = 6$

$ES_6 = ES_5 + t_{56} = 11$

$ES_7 = ES_5 + t_{57} = 14$

$ES_8 = ES_6 + t_{68} = 12$

$ES_9 = \text{Max. } \{ES_4 + t_{49}, ES_7 + t_{79}\} = \text{Max. } \{10, 16) = 16$

$ES_{10} = \text{Max. } \{ES_9 + t_{910}, ES_8 + t_{810}\} = \text{Max. } \{22, 16) = 22$

Set. $LF_{10} = ES_{10} = 22$

$LF_9 = LF_{10} - t_{910} = 22 - 6 = 16$

$LF_8 = LF_{10} - t_{810} = 22 - 4 = 18$

$LF_7 = LF_9 - t_{79} = 16 - 2 = 14$

$LF_6 = LF_8 - t_{68} = 17$

$LF_5 = \text{Min. } \{LF_7 - t_{57}, LF_6 - t_{56}\} = \text{Min. } \{6, 12\} = 6.$

$LF_4 = LF_9 - t_{49} = 10$

$LF_3 = LF_5 - t_{35} = 1$

$LF_2 = \text{Min. } \{LF_4 - t_{24}, LF_5 - t_{25}\} = \text{Min. } \{9, 5\} = 5.$

$LF_1 = \text{Min. } \{LF_3 - t_{13}, LF_2 - t_{12}\} = \text{Min. } \{0, 2\} = 0.$

Activity (*i, j*)	**Duration** t_{ij}	**Total float** TF_{ij}	**Free float** FF_{ij}
1 – 2	3	2	0
1 – 3	1	0	0
2 – 4	1	6	0
2 – 5	1	2	2
3 – 5	5	0	0
4 – 9	6	6	6
5 – 6	5	6	0
5 – 7	8	0	0
6 – 8	1	6	0
7 – 9	2	0	0
8 – 10	4	6	6
9 – 10	6	0	0

The critical path is 1—3—5—7—9—10.

Example 2: *Consider the following informations:*

Activity	*Immediate predecessors*	*Duration*
A	*None*	*2*
B	*None*	*3*
C	*A*	*1*
D	*B*	*4*
E	*C, D*	*3*
F	*D*	*1*
G	*E*	*2*
H	*F*	*3*

Draw the project network and find the critical path.

Solution: The network is drawn below:

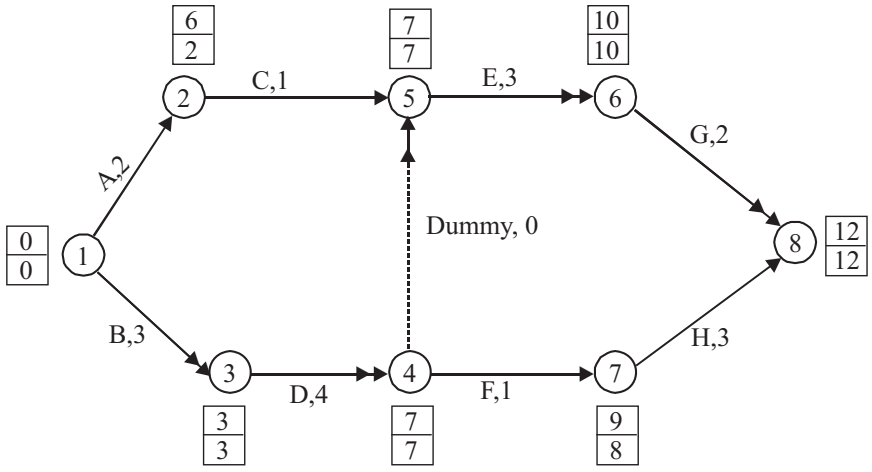


Fig. 20.7

Set $ES_1 = 0$

$ES_2 = ES_1 + t_{12} = 0 + 2 = 2$

$ES_3 = ES_1 + t_{13} = 0 + 3 = 3$

$ES_4 = ES_3 + t_{34} = 3 + 4 = 7$

$ES_5 = \text{Max. } \{ES_2 + t_{25}, ES_4 + t_{45}\}$

$= \text{Max } \{2 + 1, 7 + 0\} = 7$

$ES_6 = ES_5 + t_{56} = 10$

$ES_7 = ES_4 + t_{47} = 8$

$$ES_8 = \text{Max. } \{ES_6 + t_{68}, ES_7 + t_{78}\}$$

$$= \text{Max } \{10 + 2, 8 + 3\} = 12$$

Set $LF_8 = ES_8 = 12$

$$LF_7 = LF_8 - t_{78} = 12 - 3 = 9$$

$$LF_6 = LF_8 - t_{68} = 12 - 2 = 10$$

$$LF_5 = LF_6 - t_{56} = 10 - 3 = 7$$

$$LF_4 = \text{Min. } \{LF_5 - t_{54}, LF_7 - t_{47}\}$$

$$= \text{Min. } \{7, 8\} = 7$$

$$LF_3 = LF_4 - t_{34} = 7 - 4 = 3$$

$$LF_2 = LF_5 - t_{25} = 7 - 1 = 6$$

$$LF_1 = \text{Min. } \{LF_2 - t_{12}, LF_3 - t_{13}\}$$

$$= \text{Min. } \{4, 0\} = 0.$$

Thus the critical path is B—D—(dummy)—E—G.

20.5 PROGRAM EVALUATION AND REVIEW TECHNIQUE (PERT)

PERT was originally developed in 1958 to 1959 as part of the Polaris Fleet Ballistic Missile Program of the United States' Navy.

The primary difference between PERT and CPM is that PERT takes explicit account of the uncertainty in the activity duration estimates. CPM is activity oriented whereas PERT is event oriented. CPM gives emphasis on time and cost whereas PERT is primarily concerned with time.

In PERT, the probability distribution is specified by three estimates of the activity duration— a most likely duration (t_m), an optimistic duration (t_0) and a pessimistic duration (t_p). This type of activity duration is assumed to follow the beta distribution with

$$\text{Mean} = \frac{t_0 + 4t_m + t_p}{6}$$

and

$$\text{Variance} = \left(\frac{t_p - t_0}{6}\right)^2$$

The network construction phase of PERT is identical to that of CPM. Furthermore, once mean and variance are computed for each activity, the critical path determination

NOTES

is identical to CPM. The earliest and latest event times for the network are random variables. Once the critical path is determined, probability statements may be made about the total project duration and about the slack at any event.

Example 3: *A project consists of the following activities and different time estimates:*

Activity	t_0	t_m	t_p
1–2	*3*	*5*	*8*
1–3	*2*	*4*	*8*
1–4	*6*	*8*	*12*
2–5	*5*	*9*	*12*
3–5	*3*	*5*	*9*
4–6	*3*	*6*	*10*
5–6	*2*	*4*	*8*

(*a*) *Draw the network.*

(*b*) *Determine the expected time and variance for each activity.*

(*c*) *Find the critical path and the project variance.*

(*d*) *What is the probability that the project will be completed by 22 days?*

Solution: (a) Using the given information the resulting network is drawn as follows:

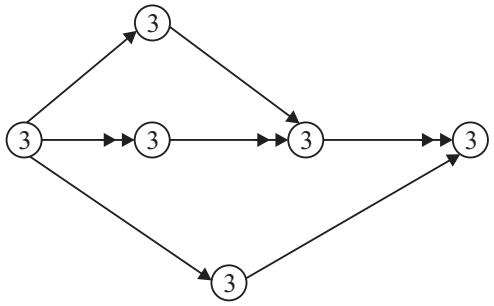


Fig. 20.8

(b) $$\text{Expected time} = \frac{t_0 + 4t_m + t_p}{6} = \bar{t}_{ij}$$

$\bar{t}_{12} = 5.17$ $\bar{t}_{25} = 8.83$ $\bar{t}_{56} = 4.33$

$\bar{t}_{13} = 4.33$ $\bar{t}_{35} = 5.33$

$\bar{t}_{14} = 8.33$ $\bar{t}_{46} = 6.17$

$$\text{Variance} = \left(\frac{t_p - t_0}{6}\right)^2.$$

$\sigma_{12}^2 = 0.694$ $\sigma_{25}^2 = 1.361$ $\sigma_{56}^2 = 1$

$\sigma_{13}^2 = 1$ $\sigma_{35}^2 = 1$

$\sigma_{14}^2 = 1$ $\sigma_{46}^2 = 1.361$

(c) Set $ES_1 = 0$

Then $ES_2 = ES_1 + \bar{t}_{12} = 5.17$

$ES_3 = ES_1 + \bar{t}_{13} = 4.33$

$ES_4 = ES_1 + \bar{t}_{14} = 8.33$

$ES_5 = \text{Max.}\{ES_3 + \bar{t}_{35}, ES_2 + \bar{t}_{25},\}$

$= \text{Max. } \{9.66, 14\} = 14$

$ES_6 = \text{Max.}\{ES_5 + \bar{t}_{56}, ES_4 + \bar{t}_{46},\}$

$= \text{Max. } \{18.33, 14.5\} = 18.33$

Set $LF_6 = ES_6 = 18.33$

Then $LF_5 = LF_6 - \bar{t}_{56} = 14$

$LF_4 = LF_6 - \bar{t}_{46} = 12.16$

$LF_3 = LF_5 - \bar{t}_{35} = 8.67$

$LF_2 = LF_5 - \bar{t}_{25} = 5.17$

$LF_1 = \text{Min.}\{LF_3 - \bar{t}_{13}, LF_2 - \bar{t}_{12}, LF_4 - \bar{t}_{14}\}$

$= \text{Min. } \{4.34, 0, 3.83\} = 0.$

Hence the critical path is (1)—(2)—(5)—(6)

Project variance $= \sigma_{12}^2 + \sigma_{25}^2 + \sigma_{56}^2$

$= 0.694 + 1.361 + 1 = 3.055.$

(d) Here mean project length is 18.33.

Set $z = \dfrac{x - 18.33}{\sqrt{3.055}} \sim N(0, 1)$

For $x = 22, z = 2.1$

NOTES

Then the required probability $= P(X \le 22) = P(z \le 2.1)$

$= 0.5 + 0.4821 = 0.9821$

$\Rightarrow$ there is 98.21% chance that the project will be completed by 22 days.

Example 4: *A PERT network consists of 10 activities. The precedence relationships and expected time and variance of activity times, in days, are given below:*

Activity	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
Immediate predecessor (s)	–	*a*	*a*	–	*b*	*c*	*d*	*d*	*e, f, g*	*h*
Expected activity time	*4*	*2*	*6*	*2*	*3*	*9*	*5*	*7*	*1*	*10*
Variance of activity time	*1*	*1*	*2*	*1*	*1*	*5*	*1*	*8*	*1*	*16*

Construct an arrow diagram. Find the critical path based on expected times. Based on this critical path find the probability of completing the project in 25 days.

Solution: The resulting network is given in Fig. 20.9.

Set

$$ES_1 = 0$$

$$ES_2 = ES_1 + \bar{t}_{12} = 4$$

$$ES_3 = ES_2 + \bar{t}_{23} = 6$$

$$ES_4 = ES_2 + \bar{t}_{24} = 10$$

$$ES_5 = ES_1 + \bar{t}_{15} = 2$$

$$ES_6 = \text{Max.}\left\{ES_3 + \bar{t}_{36}, ES_4 + \bar{t}_{46}, ES_5 + \bar{t}_{56}\right\} = 19$$

$$ES_7 = ES_5 + \bar{t}_{57} = 9$$

$$ES_8 = \text{Max.}\left\{ES_6 + \bar{t}_{68}, ES_7 + \bar{t}_{78}\right\} = 20$$

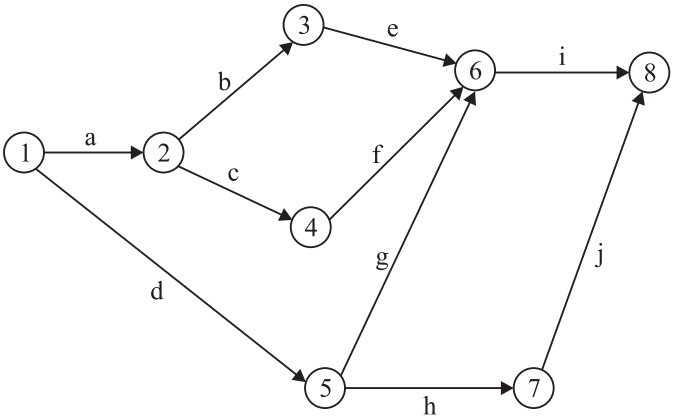


Fig. 20.9

Set
$$LF_8 = 20$$
$$LF_7 = LF_8 - \bar{t}_{78} = 10$$
$$LF_6 = LF_8 - \bar{t}_{68} = 19$$
$$LF_5 = \text{Min.}\left\{LF_7 - \bar{t}_{57}, LF_6 - \bar{t}_{56}\right\} = 3$$
$$LF_4 = LF_6 - \bar{t}_{46} = 10$$
$$LF_3 = LF_6 - \bar{t}_{36} = 16$$
$$LF_2 = \text{Min.}\left\{LF_3 - \bar{t}_{23}, LF_4 - \bar{t}_{24}\right\} = 4$$
$$LF_1 = \text{Min.}\left\{LF_2 - \bar{t}_{12}, LF_5 - \bar{t}_{15}\right\} = 0$$

Hence the critical path is $a \to c \to f \to i$ on which ES = LF

Total expected time = 4 + 6 + 9 + 1 = 20

Project variance = 1 + 2 + 5 + 1 = 9

Set
$$z = \frac{x - 20}{3} \sim N(0, 1)$$

For $x = 25$, $z = 1.67$

Then the required probability = $P(X \leq 25)$
$$= P(z \leq 1.67)$$
$$= 0.5 + \Phi(1.67)$$
$$= 0.5 + 0.4525$$
$$= 0.9525$$

$\Rightarrow$ There is 95.25% chance that the project will be completed by 25 days.

20.6 ELEMENTS OF CRASHING A NETWORK

Every activity may have two types of completion times—normal time and crash time. Accordingly costs are also two types *i.e.,* normal cost and crash cost. Obviously, the crash cost is higher than the normal cost and the normal time is higher than the crash time.

Crashing of a network implies that crashing of activities. During crashing direct cost increases and there is a trade-off between direct cost and indirect cost. So the project can be crashed till the total cost is economical. The following procedures are carried out:

(a) Calculate the critical path (CP) with normal times of the activities.

(b) Calculate the slope as given below of each activity.

$$\text{Slope} = \frac{\text{Crashing cost} - \text{Normal cost}}{\text{Normal time} - \text{Crash time}}$$

(c) Identify the critical activity with lowest slope.

(d) Compress that activity within crash limit. Compression time can also be calculated by taking min. (crash limit, free float limit).

If there are more than one critical path then select a common critical activity with least slope. If there is no such activity then select the critical activity with least slope from each critical path and compress them simultaneously within the crash limit.

(e) Continue crashing until it is not possible to crash any more.

(f) Calculate the total cost (TC) after each crashing as follows:

TC = Previous TC + Increase in direct cost – Decrease in indirect cost.

If the current TC is greater than the previous TC then the crashing is uneconomical and stop. Suggest the previous solution as optimal crashing solution.

Check Your Progress

State whether the following statements are True or False:

1. In critical path method, phase 1 is called backward pass in which latest finish time of all nodes are calculated.
2. In critical activities

   Earliest Start Time (ES) = Latest Finish Time (LF)
3. Program Evaluation Review Technique (PERT) is activity oriented while Critical Path Method (CPM) is event oriented.
4. CPM gives emphasis on time and cost.
5. PERT is primarility concerned with time.

Example 5: *A project consists of six activities with the following times and costs estimates:*

Activity	Normal time (weeks)	Normal cost (₹)	Crash time (weeks)	Crash cost (₹)
1–2	*9*	*400*	*7*	*900*
1–3	*5*	*500*	*3*	*800*
1–4	*10*	*450*	*6*	*1000*
2–5	*8*	*600*	*6*	*1000*
3–5	*7*	*1000*	*5*	*1300*
4–5	*9*	*900*	*6*	*1200*

If the indirect cost per week is ₹ 120, find the optimal crashed project completion time.

Solution: The slope calculations and the crash limit are given in the following table:

Activity	Slope	Crash limit (weeks)
1–2	250	2
1–3	150	2
1–4	137.5	4
2–5	200	2
3–5	150	2
4–5	100	3

Iteration 1

The CP calculations are shown in Fig. 20.10.

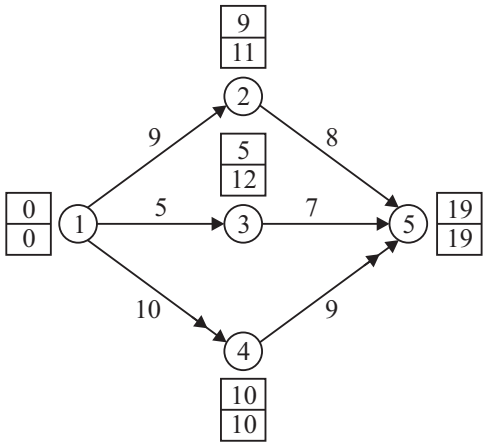


Fig. 20.10

CP: 1–4–5

Normal project duration = 19 weeks

Total direct (*i.e.,* normal) cost = ₹ 3850

Indirect cost = ₹ (19 × 120) = ₹ 2280

Total Cost (TC) = ₹ 3850 + ₹ 2280 = ₹ 6130

The slopes and crash limits of critical activities are summarised below:

Critical activity	Slope	Crash limit (weeks)
1–4	137.5	4
4–5	100*	3

Since 100 is the minimum slope, crash the activity 4–5 by 1 week *i.e.,* from 9 weeks to 8 weeks.

NOTES

Iteration 2

The CP calculations are shown in Fig. 20.11.

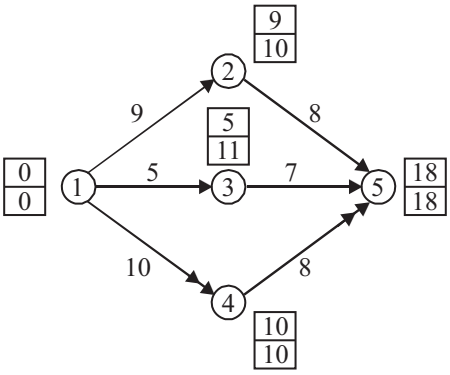


Fig. 20.11

CP: 1–4–5

Under the crashing, the project duration reduces to 18 weeks.

New TC = ₹ (6130 + 100 – 120) = ₹ 6110

Since the new TC is less than the previous TC, the present crashing is economical and proceed for further crashing.

The slopes and crash limits of critical activities are summarised below:

Critical activity	Slope	Crash limit (weeks)
1–4	137.5	4
4–5	100*	2

Crash the activity 4–5 by 1 week *i.e.,* from 8 weeks to 7 weeks.

Iteration 3

The CP calculations are shown in Fig. 20.12.

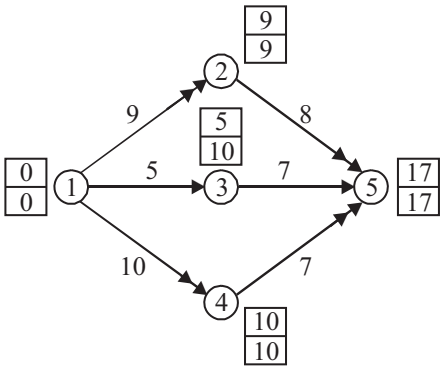


Fig. 20.12

We obtain two CPS: 1–4–5 and 1–2–5.

New TC = ₹ (6110 + 100 – 120) = ₹ 6090.

Since the new TC is less than the previous TC, the present crashing is economical and proceed for further crashing. The slopes and crash limits of critical activities are summarised below:

Critical activity	Slope	Crash limit (weeks)
1–4	137.5	4
4–5	100	1
1–2	250	2
2–5	200	2

Since there is no common critical activity, let us crash 4–5 by 1 week and 2–5 by 1 week.

Iteration 4

The CP calculations are shown in Fig. 20.13.

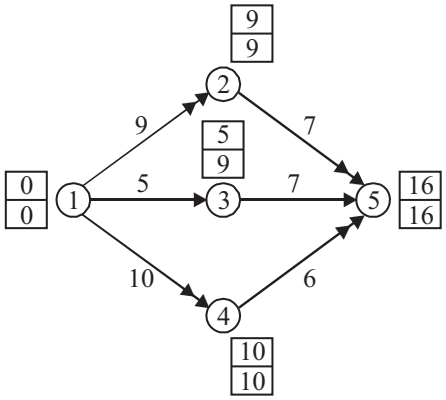


Fig. 20.13

New TC = ₹ 6090 + ₹ 100 + ₹ 200 – ₹ 120 = ₹ 6270.

Since the new TC is greater than previous TC, stop the iteration.

The previous iteration solution is the best for implementation.

Therefore, the final crashed project completion time is 17 weeks and the CPS are 1–2–5 and 1–4–5.

Check Your Progress

Choose the correct option for the following statements:

1. ..... activity cannot be started until one or more activities are completed.
   (a) predecessor activity (b) successor activity
   (c) dummy activity (d) network activity
2. ..... is a time duration in which the activity completion time can be delayed without affecting the earliest start time of immediate successor activities in the network
   (a) total float (b) independent float
   (c) free float (d) none of the above
3. In PERT, the probability distribution is specified by ..... .
   (a) most likely duration (t_m) (b) optimistic duration (t_o)
   (c) pessimistic duration (t_p) (d) all of the above
4. Crash cost is ..... the normal cost.
   (a) higher than (b) lower than
   (c) equal to (d) double
5. Crashing of a network implies ..... .
   (a) addition of activities (b) crashing of activities
   (c) economic activities (d) time bound activities

20.7 SUMMARY

- An activity is an item of work to be done that consumes time, effort, money or other resources. It is represented by an arrow.
- An event represents a point time signifying the completion of an activity and the beginning of another new activity.
- Dummy Activity shows only precedence relationship and they do not represent any real activity and is represented by a dashed line arrow or dotted line arrow and does not consume any time.
- The primary difference between PERT and CPM is that PERT takes explicit account of the uncertainty in the activity duration estimates. CPM is activity oriented whereas PERT is event oriented. CPM gives emphasis on time and cost whereas PERT is primarily concerned with time.

20.8 GLOSSARY

- ***An activity:*** An activity is an item of work to be done that consumes time, effort, money or other resources.
- ***Earliest Start Time (ES):*** This is the earliest occurrence time of the event from which the activity emanates.
- ***Latest Finish/Completion Time (LF):*** This is the latest occurrence time of the event at which the activity terminates.
- ***Latest Start Time (LS_i):*** This is the last time at which the event can occur without delaying the completing of the project.
- ***Total Floats (TF):*** It is a time duration in which an activity can be delayed without affecting the project completion time.
- ***Free Floats (FF):*** It is a time duration in which the activity completion time can be delayed without affecting the earliest start time of immediate successor activities in the network.
- ***Independent Floats:*** It is defined as the difference between the free float and the tail slack.
- ***Crashing of a Network:*** Crashing of a network implies that crashing of activities. During crashing direct cost increases and there is a trade-off between direct cost and indirect cost.

20.9 ANSWERS TO CHECK YOUR PROGRESS

1. dashed line arrow, dotted line arrow
2. start, end
3. longest
4. Latest finish
5. free float and tail slack
6. False
7. True
8. False
9. True
10. True
11. (b)
12. (b)

13. (d)
14. (a)
15. (b)

NOTES

20.10 TERMINAL AND MODEL QUESTIONS

1. For a small project of 12 activities, the details are given below:

Activity	Dependence	Duration (days)
A	-	9
B	-	4
C	-	7
D	B, C	8
E	A	7
F	C	5
G	E	10
H	E	8
I	D, F, H	6
J	E	9
K	I,J	10
L	G	2

(a) Draw the network.

(b) Find the critical path.

2. (a) Draw a network for the following project:

Activities	1–2	1–3	1–4	2–5	2–6	3–6	5–7	6–7	4–7
Time (days)	8	12	4	9	3	6	5	10	5

(b) Determine total slack time for all activities and identify the critical path.

(c) Calculate total float and free floats of each activities.

3. Consider the following informations:

Job	1–2	2–3	2–4	3–4	3–5	3–6	4–5	5–6
Time (days)	10	9	7	6	9	10	6	7

(a) Draw the network.

(b) Find the critical path.

(c) Calculate total floats and free floats of each activities.

4. Draw the network using the given precedence conditions. Calculate the critical path and floats (total and free).

Activity	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***
Immediate Predecessor(s)	–	A	A	A	D	D	E	F, G	C, H	B
Duration (months)	1	4	2	2	3	3	2	1	3	2

5. Develop a network with the following data:

Activity	**Preceded by initial activity**	**Activity Time**
A	NIL	4
B	NIL	6
C	A	10
D	B	15
E	B	10
F	C, D	8
G	E	12

6. An automobile company maufacturing scooters has decided to come up with a scooter specially designed for the women only. The project involves several activities listed in the following table:

Activity	**Description**	**Predecessor activity**
A	Study design of scooters in the market	–
B	Design the new scooter	A
C	Design the marketing programme	A
D	Design new production system	B
E	Select advertising media	C
F	Test prototype	D, E
G	Release scooter in market	F

(Draw a suitable network)

7. Draw a network diagram based on the following project schedule information available:

S. No.	Activity	Immediate Predecessor Activity	Time
1	A	–	2
2	B	–	4
3	C	A	6
4	D	B	5
5	E	C, D	8
6	F	E	3
7	G	F	2

8. The characteristics of a project schedule are given below:

S. No.	Activity	Time	S. No.	Activity	Time
1.	1 – 2	6	2.	1 – 3	4
3.	2 – 4	1	4.	3 – 4	2
5.	3 – 5	5	6.	4 – 7	7
7.	5 – 6	8	8.	6 – 8	4
9.	8 – 7	2	10.	7 – 9	2
11.	8 – 9	1			

Constructruct a suitable network.

9. A project consists of eight activities with the following time estimates:

Activity	Immediate predecessor	Time (days)		
		t_0	t_m	t_p
A	–	1	1	7
B	–	1	4	7
C	–	2	2	8
D	A	1	1	1
E	B	2	5	14
F	C	2	5	8
G	D, E	3	6	15
H	F, G	1	2	3

(a) Draw PERT network.

(b) Find the expected time for each activity.

(c) Determine the critical path.

(d) What is the probability that the project will be completed in (i) 22 days, (ii) 18 days?

(e) What project duration will have 95% chance of completion?

10. Consider the following project:

Activity	Time estimates (in weeks)			Predecessor
	t_0	t_m	t_p	
A	3	6	9	None
B	2	5	8	None
C	2	4	6	A
D	2	3	10	B
E	1	3	11	B
F	4	6	8	C, D
G	1	5	15	E

Find the critical path and its standard deviation. What is the probability that the project will be completed by 18 weeks?

11. A project has the following activities and other characteristics:

Activity	Preceding activity	Time estimates (in weeks)		
		t_0	t_m	t_p
A	—	4	7	16
B	—	1	5	15
C	A	6	12	30
D	A	2	5	8
E	C	5	11	17
F	D	3	6	15
G	B	3	9	27
H	E, F	1	4	7
I	G	4	19	28

(a) Draw the PERT diagram.

(b) Identify the critical path.

(c) Find the probability that the project is completed in 36 weeks.

NOTES

12. A small project is composed of eight activities whose time estimates are given below:

Activity	Time estimates		
	Optimistic	Most likely	Pessimistic
0 – 1	2	3	10
0 – 2	4	5	6
1 – 2	0	3	0
1 – 3	9	7	8
1 – 4	1	5	9
2 – 5	3	5	19
3 – 4	0	0	0
4 – 5	1	3	5

(a) Draw the project network.

(b) Compute the expected duration of each activity.

(c) Compute the variance of each activity.

13. Suppose the computer centre of your institute is planning to organize a national seminar. Consider the possible activities and prepare a PERT network for this seminar.

14. Draw the PERT network diagram using the given precedence conditions. Calculate the expected time, variance of each activity and the critical path.

Activity	Predecessor(s)	Duration (weeks)		
		t_0	t_m	t_p
A	–	1	2	3
B	–	1	2	8
C	A	6	7	8
D	B	1	2	3
E	A	1	4	7
F	C, D	1	5	9
G	C, D, E	1	2	3
H	F	1	2	9

15. Consider the data of a project as shown in the following table:

Activity	Normal time (weeks)	Normal cost (₹)	Crash time (weeks)	Crash cost (₹)
1–2	9	500	8	600
1–3	7	800	6	1100
1–4	8	900	6	1200
2–5	6	850	5	950
3–4	10	1200	8	1400
4–5	4	700	3	870
5–6	5	1000	4	1200

If the indirect cost per week is ₹ 160, find the optimal crashed project completion time.

16. The table below provides the costs and times for a seven activity project:

Activity (*i, j*)	Time estimates (weeks)		Direct cost estimates (₹ '000)	
	Normal	Crash	Normal	Crash
(1, 2)	2	1	10	15
(1, 3)	8	5	15	21
(2, 4)	4	3	20	24
(3, 4)	1	1	7	7
(3, 5)	2	1	8	15
(4, 6)	5	3	10	16
(5, 6)	6	2	12	36

(i) Draw the project network corresponding to normal time.

(ii) Determine the critical path and the normal duration and cost of the project.

(iii) Crash the activities so that the project completion time reduces to 11 weeks irrespective of the costs.

17. A project consists of seven activities with the following times and costs estimates:

Activity	Normal time (weeks)	Normal cost (₹)	Crash time (weeks)	Crash cost (₹)
1–2	12	500	8	900
1–3	6	600	5	700
1–4	8	700	5	850
2–5	11	500	10	820
3–5	7	1000	5	1200
4–6	6	900	4	1000
5–6	10	1200	8	1450

If the indirect cost per week is ₹ 150, find the optimal crashed project completion time.

NOTES

20.11 REFERENCES

1. Bali, N.P. (2008), *A Textbook of Quantitative Techniques*, University Science Press, New Delhi
2. Gupta, Parmanand (2008), *Comprehensive Business Statistics*, Laxmi Publications, New Delhi
3. Ahmad, Dr. Qazi Shoeb (2008), *Topics in Business, Mathematics and Statistics*, Firewall Media, New Delhi