STATISTICS

For B.S Classes

By:

Muhammad Naeem Bashir

MAJEED BOOK DEPOT

22-Urdu Bazar, Lahore. Ph: 042-37311484

composed by ARFAN ALI MATHEMATICIAN 03148100197

PREFACE TO SECOND EDITION

I feel pleasure in presenting the Second Edition of the book, "**Statistic for BS Classes**". The book is completely revised according to the latest syllabi of various universities of Pakistan.

Some new chapter on probability distributions, analysis and variance and experimental design have also been added in this edition. Moreover new examples and exercises drawn from past examination of various universities have been included. I hope that the present edition will be more useful for the students.

Thanks are due to ***Mr. Yousaf Mukhtar*** **M/S of Majeed Book Depot Urdu Bazar Lahore** for bringing out the book in a neat presentation.

Suggestion for further improvement of the book will be welcome.

Muhammad Naeem Bashir
0321-7937660

September, 2020

CONTENTS

STATISTICS

CHAPTER # 1 Nature of Statistics and Statistical Data	**1 – 14**
CHAPTER # 2 Presentation of Data	**15 – 54**
CHAPTER # 3 Measure of Central Tendency	**55 – 110**
CHAPTER # 4 Measure of Dispersion	**111 – 149**
CHAPTER # 5 Regression and Correlation	**150 – 174**
CHAPTER # 6 Multiple Regression and Correlation	**175 – 186**
CHAPTER # 7 Probability Theory	**187 – 228**

CHAPTER # 8 Random Variable	**229 – 250**
CHAPTER # 9 Probability Distributions	**251 – 274**
CHAPTER # 10 Sampling and Sampling Distributions	**275 – 311**
CHAPTER # 11 Hypothesis Testing	**312 – 348**
CHAPTER # 12 Association of Attributes and Chi Square	**349 – 362**
CHAPTER # 13 Analysis of Variance	**363 – 385**
CHAPTER # 14 Experimental Design	**386 – 409**
➲ ANSWERS of EXERCISES	**410 – 423**
➲ STATISTICAL TABLE	**424 – 429**

CHAPTER # 1

NATURE OF STATISTICS AND STATISTICAL DATA

1.1 MEANING OF STATISTICS

The word statistics have been derived from Latin word "*Status*" or the Italian word *Statista*. Both these words mean a political state or Government. Government has always been interested in the number of their citizen and in trade figures etc.

The term statistics is now commonly used in three meanings.

1. **Statistics in Plural Sense**

In plural sense statistics are aggregates of facts expressed in numerical form. *e.g.* Statistics of population, statistics of roadside accidents, statistics of births and deaths, imports and exports, male and female etc.

2. **Statistics in Singular Sense**

In singular sense it denotes the methods adopted in the collection, presentation and analysis of numerical facts.

3. **Statistics as Plural of Statistic**

The word statistics is used as the plural of the word statistic, which means a numerical quantity calculated from sample observations.

The Comprehensive definition of statistics is given below.

Statistics is the science of systematic collection, presentation, analysis and interpretation of the numerical data relating to an aggregate of facts on any type of inquiry.

1.1.1 Characteristics of Statistics

(i) **Statistics are aggregates of facts.**

(ii) **Statistics are numerically expressed.**

(iii) **Statistics are collected for a pre-determined purpose.**

(iv) **Statistics are estimated according to reasonable standard of accuracy.**

(v) **Statistics are collected in a systematic manner.**

1.2 TYPES OF STATISTICS

1. Descriptive Statistics

That branch of statistics, which deals with collection, presentation and analysis of numerical data, is called descriptive statistics.

2. Inferential Statistics

That branch of statistics, which deals with procedures of drawing inferences (conclusions) about the population's parameter on the basis of sample data, is called inferential statistics.

3. Applied Statistics

Applied statistics is a branch of statistics that make use of statistical methods and general rules in the investigation of a specific problem.

It is applied in the fields of agriculture, education, administration, economics, physics, banking and health etc.

1.2.1 Limitation of Statistics

(i) Statistical laws are valid or true on the average.

(ii) Statistics deals only with quantitative data, qualitative data *e.g.* beauty, honesty, poverty etc can not be studied directly.

(iii) Statistics deals with the aggregate of facts. Single observation is not statistics.

(iv) Good understanding of the subject is required. Only an expert can use it properly.

1.3 SUBJECT MATTER

Population

Total number of objects under consideration is called population. *e.g.* Number of students in a college etc.

Sample

It is the small part of population, which represents the whole population.

Observation

In statistics, an observation means any sort of numerically recording of information.

Parameter

A numerical quantity, which is calculated from the population, is called parameter, For example population means "u".

Statistic

A numerical quantity, which is calculated from the sample is called statistic. For example sample mean $\bar{X}$, sample variance S^2 etc.

Constant

Any value, which does not change, but remains fixed, is called constant. *e.g.* $\pi = 3.1415$, $e = 2.718$ etc.

Variable

A characteristic that varies with an individual or an object is called variable. For example height is a variable because it varies from person to person.

1.4 TYPES OF VARIABLES

Quantitative Variable

A variable is called quantitative if it can assume numerical values such as weight, height and number of children.

Qualitative Variable

A variable which can assume non numerical values is called qualitative variable *e.g.* sex, beauty, hair colour etc.

Types of Quantitative Variable

A quantitative variable may be classified as discrete and continuous.

Discrete Variable

A variable is called discrete if it can take values in the form of whole numbers *e.g.*, Number of children, the number of roadside accidents etc.

Continuous Variable

A variable is called continuous variable if it can take any value within a given interval. *e.g.*, height of a person, age of the students etc.

1.5 FIELDS (FUNCTIONS) OF STATISTICS AND THE TYPE OF QUESTIONS IT CAN ANSWER

(i) Statistics helps in simplifying the larger set of data in a form that is easily understandable.

(ii) Statistics helps in policy forming.

(iii) Statistics is widely used in survey and research.

(iv) Statistics plays an important role in Banking and Insurance.

(v) It helps in prediction.

(vi) Statistics tests the laws of other sciences.

(vii) Population census and agricultural census without statistics cannot be imagined.

(viii) Statistics studies relationship among different facts.

(ix) Statistics presents facts in a numerical form. Numerical facts give more accurate information than facts expressed in general terms.

1.6 IMPORTANCE OF STATISTICS IN DIFFERENT FIELDS

Statistics plays an important role in almost every field of life. Births and deaths are called vital events in statistics. These events are recorded as statistical events in all over the world. Following are the important uses of statistics in different fields.

Statistics and Economics and Business

Statistical methods are widely used in economics and business. The relationship between supply and demand is studied with the help of statistics. The imports and exports and inflation rate are the problems, which require good knowledge of statistics. It helps the businessman about the prices of different items in future.

Statistics and Insurance

The whole structure of insurance is based on statistics. The amount of premium for the insurance policies is based on life expectancy and accident rates.

Statistics and Banks

All types of banks make use of statistics for a number of purposes. The banks do the business with the help of deposits of the people. The bankers apply the statistical approach based on probability to estimate the number of depositors and demands of with drawls for different periods of time.

Statistics and Research

Statistics is the back bone of research. Most of the advancement has taken place due to experiments conducted with the help of statistical methods.

Statistics and Population Census

Population Census is impossible without statistics, because the whole structure of census is based on data. In population census the researcher covers each and every element concerning with the understudy population.

Statistics and other Sciences

Statistics methods are widely used in physics, chemistry, biology, astronomy, mathematics and medical sciences.

Statistics and Meteorology

Statistics plays a significant role in the field of meteorology for weather forecasting.

Statistics and Agriculture

Statistics is the back bone for agriculture. Experiments on different types of soil, various varieties of seeds and fertilizes are performed with the help of experimental designs. Experimental design is an important branch of statistics. Due to its importance a paper on statistics had been made compulsory for all the students in the agricultural university, Faisalabad. Our country is an agricultural country, its economy depends very much on agricultural products. This object cannot be achieved without complete and accurate knowledge of agricultural statistics.

Statistics and Psychology and Education

Statistics plays an important role in psychology and education. Statistical methods are applied for measurement of intelligence quotients and to determine the aptitude of a student.

1.7 STATISTICAL DATA

The set of observations relating to the experiment is called statistical data. The first step in any investigation is the collection of data. The data may be collected for the whole population or for a sample only. It is mostly collected on sample basis. Statistical data may be classified as primary and secondary.

Primary Data

Primary data are those which are collected for the first time and are original in character.

Secondary Data

Data that have undergone any sort of statistical treatment at least once *i.e.* the data have been collected, classified or tabulated are called secondary data.

1.7.1 Collection of Primary Data

Following methods are used to collect primary data.

(i) Direct Personal Investigation

In this method, an investigator collects the information personally form the individual concerned. Information obtained by this method is quite accurate and reliable; but this method is expensive and time consuming.

(ii) Indirect Personal Investigation

Some times the informants hesitate to give the correct information *e.g.* true income or production etc. In such a case information is collected from the third party. For example, businessman is reluctant to disclose their income to income tax department. In this case official can get desired information by interviewing those people who are not directly involved that business.

(iii) Collection Through Questionnaires

Questionnaire is a list of questions relating to the field of inquiry. This list of questions is sent by mail to the informants with a request that they send it back after answering the questions. This method is cheap. But the difficulty is that the majority of respondent does not care to fill the questionnaires. It is important to note that the questionnaire should be simple and concise.

(iv) Collection Through Enumerators

In this method information is collected through trained enumerators. The enumerators approach the people, ask them the relevant questions and note their answers in the questionnaire. This method gives the reliable information, but this is very costly and can be carried out by the government only.

(v) Collection Through Local Sources

In this method there is no formal collection of data. Local agents or correspondents are directed to collect and send the required information using their own judgment. This method is widely used in crop estimation.

1.7.2 Collection of Secondary Data

The secondary data may be obtained from the following sources.

(i) Official

All the government departments maintain their data and publish it annually as official statistics. *e.g.*, Federal Bureaus of statistics, provincial Bureaus of statistics, Economic Survey of Pakistan, Ministries of Finance, Commerce, Food and Agriculture, Education, Health etc.

(ii) Semi-Official

For example, State Bank of Pakistan, Railway, Wapda, Pakistan International Air Lines (PIA), District Councils etc.

(iii) Private Sources

For example, Publications of trade associations, Chambers of Commerce, Stock markets etc.

(iv) Research Organizations

For example, Publications of universities, medical colleges, Pakistan Agriculture Research Council etc.

1.8 MEASUREMENT SCALES

By measurement we usually mean the assiging of numbers to observations and scaling is a process of measuring. The four types of scales are nominal scale, ordinal scale, interval scale and ratio scale.

Nominal Scale

The classification or grouping of observations into mutually exclusive qualitative categories is called nominal scale. For example gender can be classified into two groups; male and female, this scale can also be used to represent yes/no, like/dislike etc.

Ordinal or Ranking Scale

It is known as ranking scale which posses the property of order. It includes the characteristics of a nominal scale. For example performance of students is rated as excellent, goods or poor or grades like A, B, C, D and F.

Interval Scale

This scale has the properties of order and distance. It has arbitrary zero point and constant unit of measurement. The number on the scale are placed at equidistance.

The centigrade thermometer and Fahrenheit thermometers are some examples of interval scale. In centigrade thermometer, the minimum number is OC° and maximum number is 100C°. Between these two numbers the numbers are placed at equal distance. In Fahrenheit thermometer, the minimum number is 32F° and the maximum number is 212F°. The relationship between these the systems of scales is shown below

$$C = \frac{5}{9}(F - 32)$$

It can not be said that a temperature of 50 degrees is twice as hot as a temperature of 25 degrees i.e 50/25 ratio has no meanings.

Ratio Scale

It is a special kind of an interval scale where the scale of measurement has a true zero point as its origin. Such scales are used in physical sciences to measure length, breadth, weight, volume etc. A person who is 40 years old is twice as old as one who is 20 years old.

EXERCISES

1.1 Explain the different meanings of statistics.

1.2 Define statistics and discuss the type of question it can answer.

1.3 Explain the importance of statistics in different fields.

1.4 Define data. What are the main types of statistical data?

1.5 Differentiate between primary data and secondary data. Which are the main sources of these data?

1.6 Define statistics. Discuss its importance in detail. What are the different methods of collection of primary data?

1.7 Define statistics. Distinguish between primary and secondary data.

1.8 **Explain the difference between the following.**

(i) Discrete and continuous variable.

(ii) Population and Sample.

(iii) Qualitative and Quantitative variable.

(iv) Parameter and Statistic.

(v) Descriptive and Inferential statistics.

(vi) Variable and Constant.

1.9 **Explain the importance of statistics in the following fields.**

(i) Education and psychology

(ii) Agriculture

(iii) Banks

(iv) Economics and Business.

1.10 **Classify the following variables as Qualitative or Quantitative.**

(i) Number of children in a family.

(ii) Colour of roses in a garden.

(iii) Martial status of the persons working in an office.

(iv) Height of the students.

(v) Liking or disliking of 20 people.

(vi) The distance traveled between Lahore and Islamabad.

(vii) The average number of eggs lay by hen in a year.

(viii) The occupations of the persons in a locality.

(ix) Income of a government servant in a city.

(x) Religion of people of a country.

(xi) Wages of workers in a factory.

(xii) The beauty of different persons.

1.11 Classify the following variables as discrete or continuous.

(i) Weight of a student.

(ii) The number of students in a class.

(iii) The number of deaths in an accident.

(iv) Temperature of a place.

(v) The amount of milk produced by a cow.

(vi) The number of children in a family.

(vii) The lifetime of electric tube light.

(viii) The number of passengers in a train.

(ix) Number of books in a library.

(x) The distance between Faisalabad and Lahore.

(xi) Monthly income of an officer.

(xii) The age of a business man.

(xiii) The number of dots, when a single die is rolled.

(xiv) Number of districts in Pakistan.

(xv) Amount of water in a jug.

(xvi) Number of mistakes on a page.

(xvii) Number of petals on a flower.

1.12 SELECT THE CORRECT ANSWER

(i) The word Statistics has

(a) two meanings (b) three meanings

(c) four meanings (d) five meanings

(ii) Sample is a representative part of

(a) Statistic (b) Proportion

(c) Parameter (d) Population

(iii) A numerical quantity computed from a population is called

(a) Statistic (b) Parameter

(c) Experiment (d) Observation

(iv) Statistics are

(a) aggregate of facts (b) always true

(c) always continuous (d) always qualitative

(v) **A characteristic that varies from object to object is called**

(a) Constant (b) Sample

(c) Variable (d) Population

(vi) **A variable which can assume numerical values is called**

(a) qualitative variable (b) quantitative variable

(c) discrete variable (d) continuous variable

(vii) **Date which is collected for the first time and are original in character is called**

(a) Primary data (b) Secondary data

(c) Qualitative data (d) Quantitative data

(viii) **Census returns are**

(a) Primary data (b) Secondary data

(c) Published data (d) Qualitative data

(ix) **Data published by Government sources are**

(a) Primary data (b) Secondary data

(c) First hand data (d) Raw data

(x) **The group data is also called**

(a) Raw data (b) Primary data

(c) Secondary data (d) Qualitative data

(xi) **Any value which does not change but remains fixed is called**

(a) Parameter (b) Constant

(c) Variable (d) Statistic

(xii) **Total number of objects under consideration is called**

(a) Sample (b) Statistic

(c) Population (d) Parameter

(xiii) **A variable which can assume non numerical values is called**

(a) Variable (b) Quantitative variable

(c) Qualitative variable (d) Discrete variable

(xiv) **Data which have undergone any sort of statistical treatment is known as**

(a) Discrete data (b) Primary data

(c) Continuous data (d) Secondary data

(xv) The lifetime of T.V tube is:

(a) Discrete data (b) Continuous data
(c) Constant (d) None of these

(xvi) The sources of primary data are primary:

(a) Surveys (b) Areas
(c) Units (d) Factors

(xvii) Data classified by attributes is called:

(a) Qualitative (b) Quantitative
(c) Discrete (d) Continuous

(xviii) A variable which can take all possible values in an interval is called:

(a) Discrete variable (b) Continuous variable
(c) Qualitative variable (d) Finite variable

(xix) A discrete data has:

(a) No class boundaries (b) Class boundaries
(c) Fractions (d) None of these

(xx) The sum of random error is equal to:

(a) 1 (b) 2
(c) 3 (d) Zero

(xxi) Questionnaire method is used in collection of:

(a) Secondary data (b) Grouped data
(c) Primary data (d) None of these

(xxii) Any data given in your book for you is:

(a) Primary (b) Raw
(c) Secondary (d) False

ANSWERS: 1.12:

(i)	(b)	**(ii)**	(d)	**(iii)**	(b)	**(iv)**	(a)	**(v)**	(c)
(vi)	(b)	**(vii)**	(a)	**(viii)**	(a)	**(ix)**	(b)	**(x)**	(c)
(xi)	(b)	**(xii)**	(c)	**(xiii)**	(c)	**(xiv)**	(d)	**(xv)**	(b)
(xvi)	(a)	**(xvii)**	(a)	**(xviii)**	(b)	**(xix)**	(a)	**(xx)**	(d)
(xxi)	(c)	**(xxii)**	(c)						

SHORT QUESTION AND ANSWERS

WRITE THE SHORT ANSWERS TO THE FOLLOWING QUESTIONS.

1. **Define Statistics.**

Ans. Statistics is the science of systematic collection, presentation, analysis and interpretation of the numerical data.

2. **Write down the different meaning of Statistics.**

Ans. (a) **Statistics in Plural Sense**

In plural sense statistics are aggregates of facts expressed in numerical form. *e.g.* statistics of accidents, births and deaths etc.

(b) **Statistics in Singular Sense**

In singular sense it denotes the methods adopted in the collection, presentation and analysis of numerical facts.

(c) **Statistics as Plural of Statistic**

The word Statistics is used as the plural of word statistic which means a numerical quantity calculated from sample observations.

3. **What are the characteristics of Statistics?**

Ans. (a) Statistics are aggregates of facts.

(b) Statistics are collected for a pre-determined purpose.

(c) Statistics are numerically expressed.

(d) Statistics are collected in a systematic manner.

4. **What is inferential Statistics?**

Ans. That branch of Statistics, which deals with procedures of drawing inferences (conclusions) about the population's parameter on the basis of sample data.

5. **Define descriptive statistics**

Ans. That branch of statistics which deals with the collection, presentation and analysis of numerical data is called descriptive statistics.

6. **Define applied statistics.**

Ans. Applied statistics is a branch of statistics that make use of statistical methods and general rules in the investigation of a specific problem.

7. **Define Population.**

Ans. Total number of objects under consideration is called population. *e.g.* No. of students in a college etc.

8. Define Sample.

Ans. It is the small part of population, which represents the whole population. For example, if we select 10 students from a class of 100 students, 10 selected students constitute a sample.

9. What is the difference between constant and variable?

Ans. Constant

Any value, which does not change, but remains fixed, is called constant. *e.g.* $\pi = 3.14$, $e = 2.718$ etc.

Variable

A characteristic that varies with an object is called variable. For example, height, weight etc.

10. Define Statistic and Parameter.

Ans. Statistic

A numerical quantity which is calculated from the sample is called Statistic *e.g.* sample mean ($\bar{X}$).

Parameter

A numerical quantity which is calculated from the population is called parameter *e.g.* population mean (μ).

11. What is quantitative variable?

Ans. A variable which can assume numerical values is called quantitative variable *e.g.* weight, height etc.

12. What is the difference between discrete and continuous variable?

Ans. Discrete Variable

A variable is said to be discrete if it can take only discrete values (whole numbers). *e.g.* number of children, family size, books in a library etc.

Continuous Variable

A variable is called continuous if it can take any value within a given interval *e.g.*, height, weight, age etc.

13. What is the importance of Statistics in different fields? Give at least three lines.

Ans.

1. Statistical methods are widely used in economics and business.
2. The whole structure of insurance is based on Statistics.
3. Statistics is the back bone for agriculture and research.
4. Statistical methods are widely used in physics, mathematics, biology and social sciences.

14. Define statistical data.

Ans. The set of observations relating to the experiment is called data.

15. Distinguish between Primary and Secondary data.

Ans. **Primary Data**

Data which is collected for the first time and are original in character is called primary data.

Secondary Data

Data that have under gone any sort of statistical treatment at least once is called secondary data.

16. Write the names of different methods for collection of Primary data.

Ans. 1. Direct personal investigation.

2. Indirect Personal investigation.

3. Collection through questionnaire.

4. Collection through enumerators.

5. Collection through local sources.

17. Write down the names of different methods for the collection of Secondary data.

Ans. The secondary data may be obtained from the following sources.

(i) Official (ii) Semi official

(iii) Private sources (iv) Research organizations

CHAPTER # 2

PRESENTATION OF DATA

2.1 INTRODUCTION

We have discussed the collection of data in the previous chapter. When the data have been collected, it must be presented in a form which is easy to understand. For this purpose following methods are applied.

1. Classification
2. Tabulation
3. Diagrammatic and Graphic representation

2.2. CLASSIFICATION

The process of arranging the data into homogenous classes according to their resemblances and similarities is called classification of data. *e.g.*, the process of sorting letters in a post office. The letters are classified according to cities first then further arranged according to towns, sectors and streets.

2.2.1 The Basis of Classification

The basis of classification are

(i) Geographical or Spatial
(ii) Chronological or Temporal
(iii) Qualitative
(iv) Quantitative

(i) Geographical or Spatial Classification

In geographical classification, the data are classified by geographical region or locations. For example, population of a country may be classified by province division, district and Tehsils etc.

Population size by Province (1998 Census)

Administration Unit	Population (in thousands)
Punjab	72585
Sindh	29991
N.W.F.P.	17555
Balochistan	6511
Fata	3138
Islamabad	799
Pakistan	130580

Sources : Population Census Report, 1998.

(ii) Chronological (Temporal) Classification

When the data are classified according to the time of its occurrence such as year, months, weeks, days, hours etc. It is called chronological classification. An arrangement of data by their time of occurrence is called a time series.

Population of Punjab Province Since 1951

Years	Population
1951	20,556,800
1961	25,499,876
1972	37,611,668
1981	47,292,441
1998	72,585,430

Source : Population Census Report, 1998.

(iii) Qualitative Classification

In qualitative classification, the data are classified on the basis of some quality or attribute, such as religion, sex, beauty, martial status, colour etc.

(iv) Quantitative Classification

In this type of classification data are classified according to some quantitative measurements *e.g.* heights, income, age, sales, temperature etc.

2.3 TABULATION

The systematic arrangement of data in to rows and columns is called tabulation.

2.3.1 The Main Parts of a Table

The main parts of table is

(i) Title
(ii) Prefatory notes
(iii) Column captions and box head
(iv) Row captions and Stub.
(v) Body of a table
(vi) Foot Notes
(vii) Source Notes

(i) Title

A tile is the main heading of the table. Title should be in capital letters and must be written on the top of the table. The title should be clear, brief and self-explanatory.

(ii) Prefatory Notes

It is a statement, given below the title and enclosed in brackets usually describing the units of measurements.

(iii) Column Captions and Box Head

The heading of each column is called a column caption, while the portion of a table that contains the column caption is a called box head. The headings should be clear. Only the first word in each column should be in capital letters.

(iv) Row Caption and Stub.

The heading of each row is called a row caption, while the section of a table that contains row captions is called stub.

(v) Body of a Table

This is the main part of the table, which contains the numerical information.

(vi) Foot Notes

It is given below the body of the table. Footnotes are used to clarify any information in the table. It provides also additional information about the table.

(vii) Source Notes

The source note is given at the end of the table. It indicates the source from where information have been taken. It includes the information about compiling agency, publication and page number etc.

GENERAL SKETCH OF A TABLE

← TITLE →

← Prefatory notes

Box Head Stub ↓ →	COLUMN CAPTIONS ↓							
ROW CAPTIONS →								
			BODY					

Foot notes : _____

Source : _____

2.3.2 General Rules of Tabulation

(i) A table should be simple and easy to understand.

(ii) A table should be complete and self-explanatory.

(iii) The row and column captions should be arranged in a systematic manner.

(iv) Do not use ditto marks. If a figure is repeated show it each time.

(v) Abbreviations should be avoided, especially in the title and headings.

2.3.3 Types of Tabulation

1. One way Tabulation

When the data are tabulated according to one characteristic, then it is called one way tabulation. *e.g.* Division wise population of Punjab according to 1998 census.

Division Wise Population of Punjab According to 1998 Census

DIVISION	Population (in thousands)
Bahawalpur	7518
D.G. Khan	6402
Faisalabad	9735
Gujranwala	11115
Lahore	13985
Multan	11428
Rawalpindi	6780
Sargodha	5622

Source : Population Census Report, 1998.

Two Way Tabulation

When the data are tabulated according to two characteristics / criteria, then it is called two way tabulation. *e.g.*, Division wise population of Punjab by Sex.

Division Wise Population of Punjab by Sex (1998 Census)

Division	Population (in thousands)		
	Male	Female	Total
Bahawalpur	3918	3600	7518
D.G. Khan	3330	3072	6402
Faisalabad	5053	4682	9735
Gujranwala	5694	5421	11115
Lahore	7321	6664	13985
Multan	5935	5493	11428
Rawalpindi	3396	3384	6780
Sargodha	2863	2759	5622

Source : Population Census Report, 1998.

Three Way Tabulation

When the data are tabulated according to three characteristics at a time, then it is called three-way tabulation.

2.4 FREQUENCY DISTRIBUTION

The arrangement of data according to size and magnitude is called frequency distribution.

Group Data

Data presented in the form of frequency distribution are called group data.

Ungroup Data

The data, which are collected for the first time and original in character, are called ungroup data.

2.4.1 Formation of Frequency Distribution

There are two types of frequency distribution.

1. Discrete Frequency Distribution
2. Continuous Frequency Distribution

2.4.2 Formation of Discrete Frequency Distribution

To prepare the discrete frequency distribution following steps are taken.

(i) Find the largest and smallest value of the given data.

(ii) Place all possible values of the variable from the smallest to largest in the first column with the title of the variable.

(iii) The second column is the tally column. In this column a vertical bar called tally bar is put against the particular value to which it relates.

(iv) Last column is the frequency column. Count all the bars and write these numbers in the frequency column. The total of the frequency column must be equal to the total number of observations.

When the discrete data are sufficiently large, then they are treated in the same way as continuous frequency distribution.

EXAMPLE 2.1

The numbers of children per family of 25 families are given below.

2, 3, 0, 5, 4, 1, 2, 4, 6, 3, 3, 5, 6, 4, 7, 1, 4, 3, 5, 3, 1, 2, 3, 5, 4

Represent the data in the form of discrete frequency distribution.

composed by ARFAN ALI MATHEMATICIAN 03148100197

SOLUTION

Frequency Distribution of No. of Children

No. of Children "x"	Tally bar	No. of families "f"
0	|	1
1	|||	3
2	|||	3
3	~~||||~~ |	6
4	~~||||~~	5
5	||||	4
6	||	2
7	|	1
Total	—	25

Terms Related with Continuous Frequency Distribution

Class Limits

The values of the classes are called class limits. Every class has two values. The smaller value is the lower class limit and the larger value is the upper class limit.

Mid Point or Class Mark

The average value of the lower and upper limit is called mid point or class mark.

$$\text{Mid points} = \frac{\text{Lower class limit} + \text{upper class limit}}{2}$$

Class Boundaries

The true or precise values, which describe the actual or true class limits of a class are called class boundaries. The upper class boundary of a class coincide with the lower class boundary of the next class.

Class Interval

The width of the class is called the class interval. It is also defined as the difference between the upper class boundary and the lower class boundary. It is denoted by h.

Class Frequency

The number of values in any class is the class frequency of that class. It is denoted by f.

2.4.3 Construction of Continuous Frequency Distribution

Following steps are involved in the construction of continuous frequency distribution.

1. **Determine the Range**

First of all find the range of the data. Range is the difference between the largest and the smallest value in the data.

2. **Decide the Number of Classes**

There is no hard and fast rule for this purpose. Number of classes should neither be very large nor very small. When the data are sufficiently large, the number of classes should lie between 5 and 20. H.A sturges has given a formula for determining the number of classes *i.e.*,

$$K = 1 + 3.3 \log N$$

Where

K = Number of classes

N = Total No. of observations

For example if there are 50 observations, then

$$K = 1 + 3.3 \log 50$$

$$K = 1 + 3.3\,(1.6990)$$

$$= 1 + 5.607$$

$$K = 6.6067 \approx 7$$

i.e. 7 classes

But this rule is rarely used in practice.

3. **Decide the size of the Class Interval**

The size of the class interval depends upon the range and the number of classes. Sturges has also given a formula for this purpose.

$$\text{Class interval} = \frac{\text{Range}}{K}$$

In case of fractional results, the next higher whole number is usually taken as the size of the class interval.

4. Deciding the Starting Point

The lowest class usually starts with a number that is a multiple of class interval such that it should cover the smallest value of the data.

5. Determine the Remaining Class Limits

The remaining lower and upper class limits may be obtained by adding the size of class interval.

6. Distribute the data into Appropriate Classes

All the observations are put into respective classes. This may be done by using "Tally Bar". The frequency column is obtained by counting the tally bars of each class.

7. Total of the Frequency Column

The total of the frequency column must be equal to the total number of observations to see that all the data have been accounting for.

EXAMPLE 2.2

Marks obtained by 60 students of a class are given below.

60, 50, 46, 28, 58, 64, 36, 20, 50, 18, 42, 56, 20, 38, 40, 34, 24, 64, 64, 42, 46, 52, 50, 44, 36, 0, 24, 30, 46, 40, 64, 40, 36, 14, 36, 8, 56, 40, 30, 36, 24, 22, 36, 50, 58, 16, 40, 34, 0, 42, 42, 0, 36, 18, 18, 68, 30, 46, 38, 16.

Make a frequency distribution using appropriate class interval.

SOLUTION

First of all we find the range

Range = Largest value − smallest value

= 68 − 0 Largest value = 68

= 68 Smallest value = 0

Now we determine the No. of classes.

$K = 1 + 3.3 \log N$

$K = 1 + 3.3 \log (60)$

$= 1 + 3.3 (1.7782)$

$= 1 + 5.87 \Rightarrow K = 6.87 \approx 7$ classes

Now we decide the size of the class interval

$$h = \frac{\text{Range}}{K} = \frac{68}{7} = 9.71 \qquad i.e., h = 10$$

Size of class interval = 10

Frequency distribution of the marks of students

Marks	Tally Bars	Frequency
0 – 9	||||	4
10 – 19	卌 |	6
20 – 29	卌 ||	7
30 – 39	卌 卌 ||||	14
40 – 49	卌 卌 ||||	14
50 – 59	卌 ||||	9
60 – 69	卌 |	6
Total	--	60

EXAMPLE 2.3

Form a frequency distribution from the following data taking 0.05 as the class interval and 1.19 as the lowest class limit.

1.35, 1.46, 1.50, 1.32, 1.45, 1.24, 1.49, 1.64, 1.47, 1.59, 1.41, 1.48, 1.36, 1.48, 1.51, 1.45, 1.26, 1.38, 1.76, 1.63, 1.19, 1.56, 1.65, 1.54, 1.61, 1.73, 1.60, 1.50, 1.45, 1.76, 1.67, 1.35, 1.55, 1.68, 1.46, 1.40, 1.32, 1.47, 1.64, 1.45

SOLUTION

FREQUENCY DISTRIBUTION

Class Limits	Tally Bars	Frequency (f)
1.19 – 1.23	|	1
1.24 – 1.28	||	2
1.29 – 1.33	||	2
1.34 – 1.38	||||	4
1.39 – 1.43	||	2
1.44 – 1.48	~~||||~~ ~~||||~~	10
1.49 – 1.53	||||	4
1.54 – 1.58	|||	3
1.59 – 1.63	||||	4
1.64 – 1.68	~~||||~~	5
1.69 – 1.73	|	1
1.74 – 1.78	||	2
Total	--	**40**

Cumulative Frequency

The total frequency of all classes less than the upper class boundary of a given class is called cumulative frequency of that class.

Less than Cumulative Frequency

If the Cumulative process is from the lowest value to the highest, it is called less than cumulative frequency.

More than Cumulative Frequency

If the cumulative process is from the highest value to the lowest value, it is called more than cumulative frequency.

Relative frequency

The frequency of a class divided by the total frequency is called relative frequency of that class.

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Total frequency}}$$

EXAMPLE 2.4

The following table gives the frequency distribution of marks obtained by 60 students of a class. (See Example 2.2)

Marks	0 – 9	10–19	20–29	30–39	40–49	50–59	60–69
No. of Students	4	6	7	14	14	9	6

Construct the following.

(i) Class boundaries and mid points.

(ii) A less than cumulative frequency.

(iii) A more than cumulative frequency.

(iv) Relative frequency.

SOLUTION

(i) **Class boundaries and Mid Points**

Class Limits	Frequency	Class Boundaries	Mid Point
0 – 9	4	0 – 9.5	4.5
10 – 19	6	9.5 – 19.5	14.5
20 – 29	7	19.5 – 29.5	24.5
30 – 39	14	29.5 – 39.5	34.5
40 – 49	14	39.5 – 49.5	44.5
50 – 59	9	49.5 – 59.5	54.5
60 – 69	6	59.5 – 69.5	64.5

(ii) Less than Cumulative Frequency

End values (Upper C.B.)	f	Cumulative Frequency
Less than 0	0	0
Less than 9.5	4	4
Less than 19.5	6	4 + 6 = 10
Less than 29.5	7	10 + 7 = 17
Less than 39.5	14	17 + 14 = 31
Less than 49.5	14	31 + 14 = 45
Less than 59.5	9	45 + 9 = 54
Less than 69.5	6	54 + 6 = 60

(iii) More than Cumulative Frequency

End values Lower C.B.)	f	Cumulative Frequency
More than 0	4	56 + 4 = 60
More than 9.5	6	50 + 6 = 56
More than 19.5	7	43 + 7 = 50
More than 29.5	14	29 + 14 = 43
More than 39.5	14	15 + 14 = 29
More than 49.5	9	6 + 9 = 15
More than 59.5	6	6
More than 69.9	0	0

(iv) Relative Frequency

Marks	f	Relative Frequency
0 – 9	4	$\frac{4}{60} = 0.07$
10 – 19	6	$\frac{6}{60} = 0.10$
20 – 29	7	$\frac{7}{60} = 0.12$
30 – 39	14	$\frac{14}{60} = 0.23$
40 – 49	14	$\frac{14}{60} = 0.23$
50 – 59	9	$\frac{9}{60} = 0.15$
60 – 69	6	$\frac{6}{60} = 0.10$
Total	60	1.00

2.5 GRAPHIC AND DIAGRAMMATIC REPRESENTATION OF DATA

Visual representation of data in the form of points, lines, symbols and areas are called Graphic representation. Such visual representation of data can be divided into two main groups namely diagrams and graphs.

2.5.1 Diagrams

There are different types of diagrams. Commonly used diagrams are

(a) One dimensional diagram

(b) Aerial or two dimensional diagrams

(c) Pie Diagrams

2.6 ONE DIMENSIONAL DIAGRAMS

These are the diagrams in which only the length is taken into account but are of equal width. One dimensional diagrams are further divided into

(i) Simple Bar Diagram

(ii) Multiple Bar Diagram

(iii) Sub divided or Component Bar Diagrams

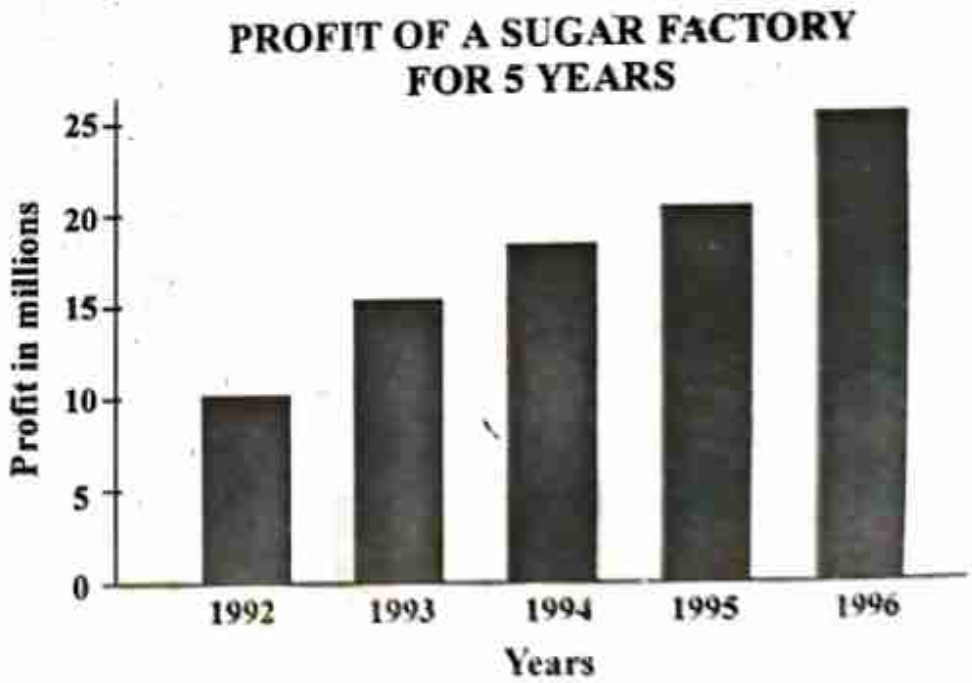
2.6.1 Simple Bar Diagrams

A simple bar diagram is used when the data consist of single factor, such as production, profit, yield etc. It consists of horizontal or vertical bars of equal width The length of the bar is in proportion to the actual data. The space between the bars should not exceed than the width of the bar. The bars should neither very tall nor very small. The data should be arranged in ascending or descending order if the time is not given.

EXAMPLE 2.5

Draw a simple bar diagram to represent the profits of a sugar factory for 5 years.

Year	1992	1993	1994	1995	1996
Profit (Million Rs.)	10	15	18	20	25



2.6.2 Multiple Bar Diagram

Multiple bar diagram is an extension of simple bar diagram, which is applicable to only one factor. But in multiple bar diagrams two or more factors are

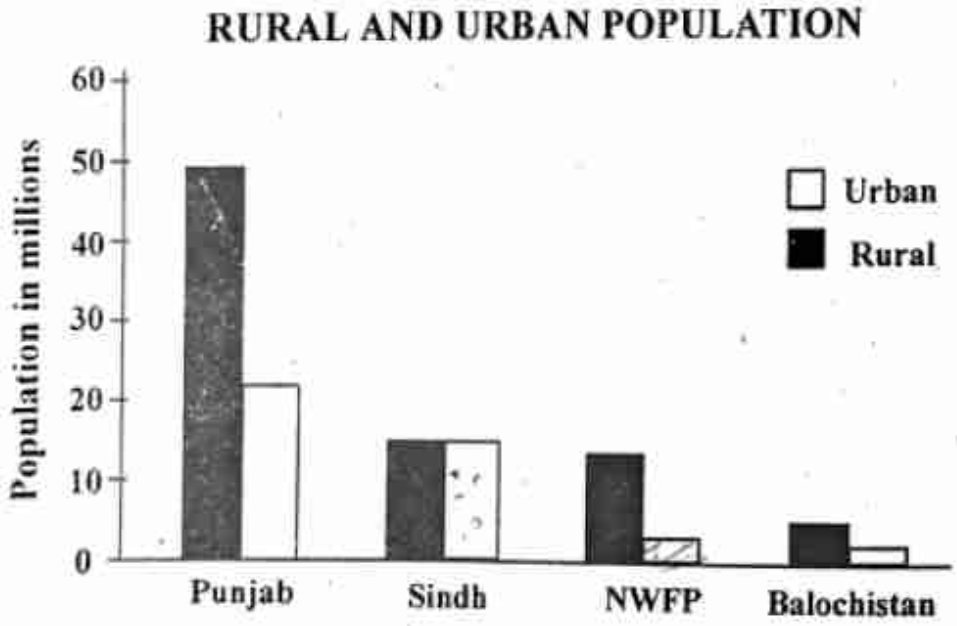
taken into account. In this diagram, grouped bars are used to represent related sets of data. Each bar is shaded or coloured differently for identification. It is a good device for the comparison of two or more than two kind of information. For example, exports and imports, birth and death rates of a country can be compared.

EXAMPLE 2.6

Draw a multiple bar diagram to represent the Rural and Urban population of four provinces of Pakistan in 1998.

Provinces	Population in Millions	
	Rural	Urban
Punjab	49.9	22.7
Sindh	15.3	14.7
N.W.F.P	14.6	3.0
Balochistan	5.0	1.5

Source : Population Census Report, 1998.



2.6.3 Component Bar Diagram

Component bar diagrams are those in which each bar represent the total of the components. Then it is divided into two or more sections in the ratio of their

Teeme

components. The components parts are coloured or shaded differently for identification. They are also known as sub-divided bars.

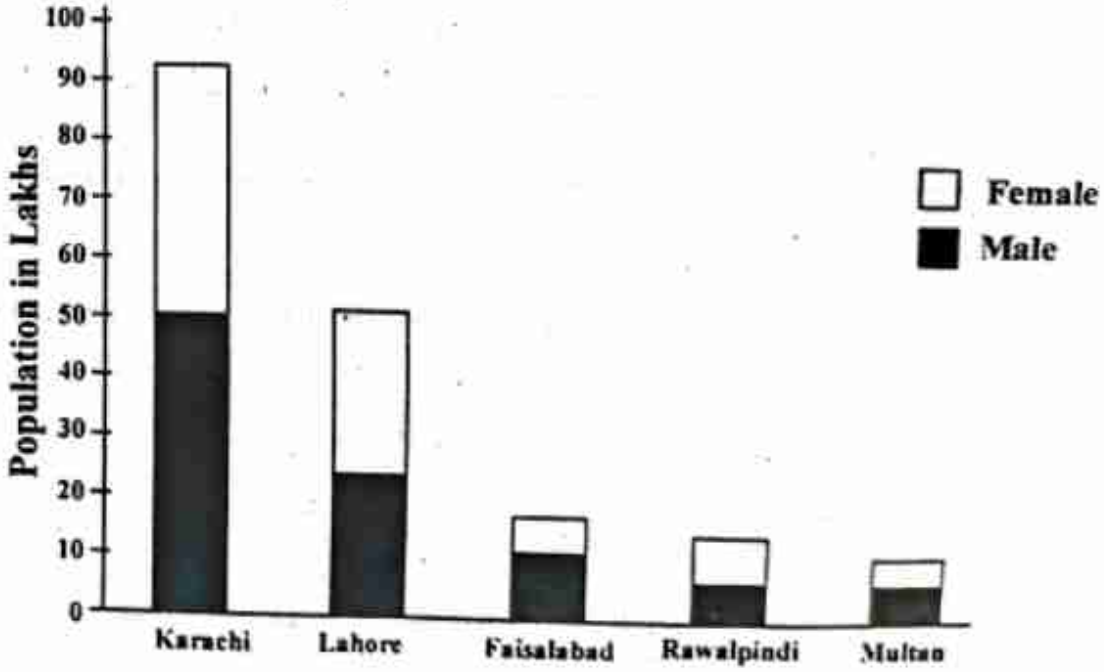
EXAMPLE 2.7

Draw a component bar diagram to represent the male and female population of 5 big cities in Pakistan in 1998.

Cities	Population in Lakhs		
	Male	Female	Total
Karachi	50	43	93
Lahore	26.6	24	50.6
Faisalabad	10.4	9.4	19.8
Rawalpindi	7.5	6.6	14.1
Multan	6.3	5.5	11.8

Source : Pakistan Census Report, Govt. of Pakistan, July 1998.

MALE AND FEMALE POPULATION OF FIVE BIG CITIES IN PAKISTAN IN 1998



2.7 AERIAL OR TWO DIMENSIONAL DIAGRAMS

Two dimensional diagrams are based on the area. The area is the product of length and breadth. Hence in two dimensional diagrams, both length and breadth are taken into account. They are in the form of rectangles, subdivided rectangles and squares.

2.7.1. Rectangular Diagrams

When the data have moderately large variations, rectangular diagrams can be used. The area of the rectangle is equal to the product of its length and breadth. There are two methods of drawing rectangles.

(i) The lengths of the rectangle are kept constant and breadths are taken proportional to the size of data.

(ii) Secondly breadths are kept constant and length are taken proportional to the size of data.

EXAMPLE 2.8

Draw rectangles to represent the are of four provinces of Pakistan.

Provinces	Area (Thousand square km)
Balochistan	347
Punjab	205
Sindh	141
K.P.K	75

SOLUTION

In order to keep the breadths constant. We split the total into two portions as

Area of Balochistan $= 347 = 10 \times 34.7$

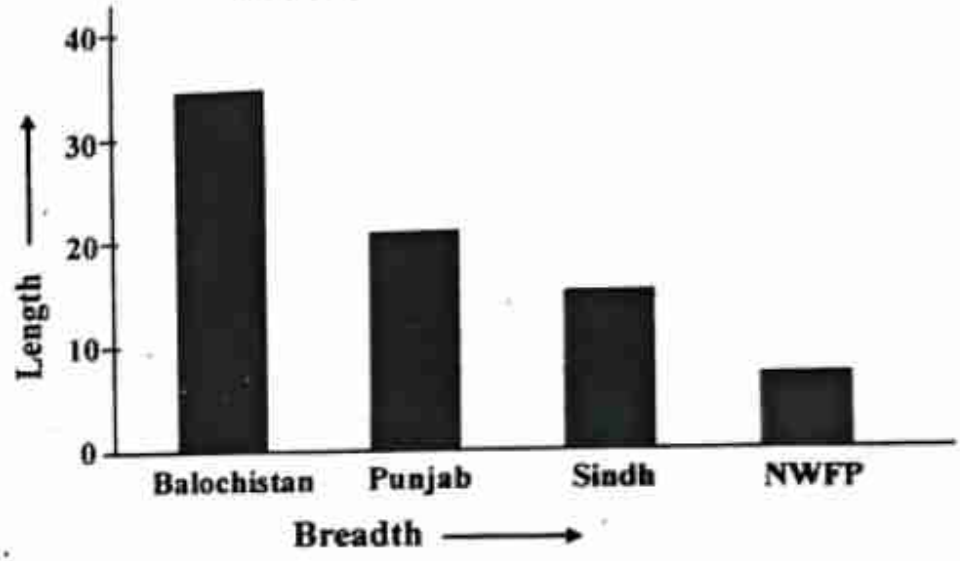
Area of Punjab $= 205 = 10 \times 20.5$

Area of Sindh $= 14.1 = 10 \times 14.1$

Area of K.P.K $= 75 = 10 \times 7.5$

In this question, the breadth of each rectangle is equal to 10 unit, while the length of rectangles are 34.7, 20.5, 14.1 and 7.5 units for Balochistan, Punjab, Sindh and KPK respectively.

RECTANGULAR DIAGRAM



2.7.2. Sub-divided Rectangles

Sub divided rectangular diagrams is used to represent the data where the quantities along with their components are to be compared. These diagrams are generally used to compare the budgets of different families.

In this diagram, rectangles are drawn with length equal to 100 units and breadth proportional to the size of the values. The component parts are expressed as percentage of their corresponding totals. Each rectangle is divided into different sections. The component parts are coloured or shaded differently to increase the effectiveness of the diagram.

EXAMPLE 2.9

Draw a sub-divided rectangular diagram to compare the budgets of two families A and B.

Items of Expenditure	Family A	Family B
Food	480	1200
Clothing	240	850
Fuel & Lighting	80	250
Housing	180	600
Service	100	400
Misc.	120	300

SOLUTION

The necessary calculations are given below.

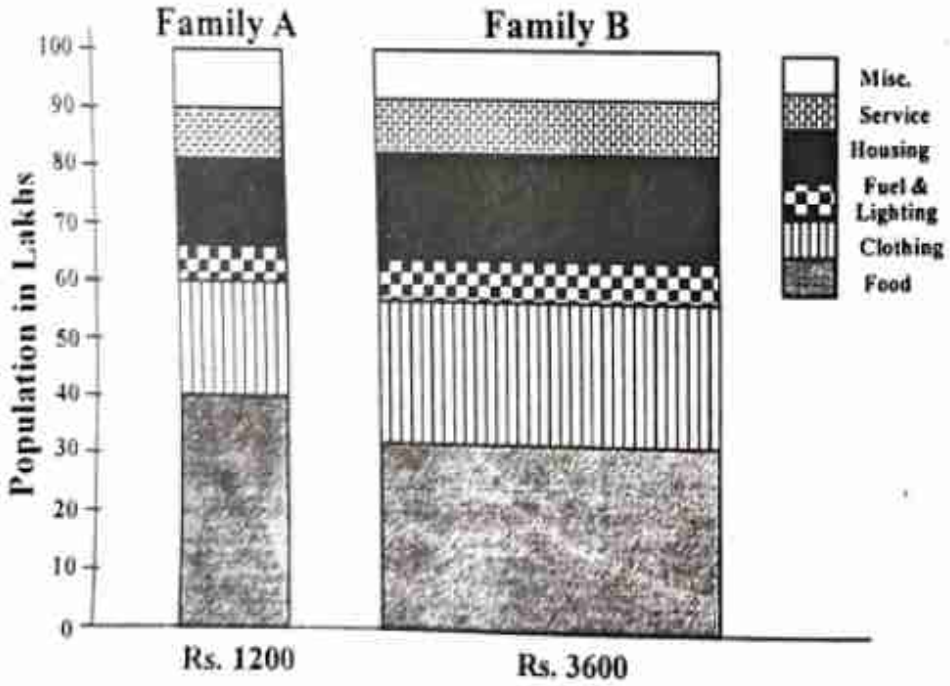
Items of Expenditure	Family A			Family B		
	Expend	% Expend	Cum %	Expend	% Expend	Cum %
Food	480	40.00	40.00	1200	33.33	33.33
Clothing	240	20.00	60.00	850	23.61	56.94
Fuel & Light	80	6.67	66.67	250	6.94	63.88
Housing	180	15.00	81.67	600	16.67	80.55
Service	100	8.33	90.00	400	11.11	91.66
Misc.	120	10.00	100.00	300	8.34	100.00
Total	1200	100.00	—	3600	100.00	—

Length of the rectangles = 100 units

The breadth of the rectangles are in proportion = 1200 : 3600

i.e 1 : 3

SUB-DIVIDED RECTANGULAR DIAGRAM



2.7.3 Pie Diagrams

Like the component bar diagram, pie diagrams show the relationship between the whole and its components. However, there is an important difference. In component bar diagrams, the length of bars are compared, where as in pie diagrams, areas of the sectors are compared.

To construct a pie diagram, draw a circle of any suitable radius. A circle consists of 360°. To show the component parts by sectors, we calculate the angles for each sector by using the relation.

$$\text{Angle} = \frac{\text{Component Part}}{\text{Whole quantity}} \times 360$$

Then divide the circle into different sectors by constructing angles at the center by means of a protractor. The different sectors are shaded differently for identification.

EXAMPLE 2.10

Draw a pie diagram for the following data.

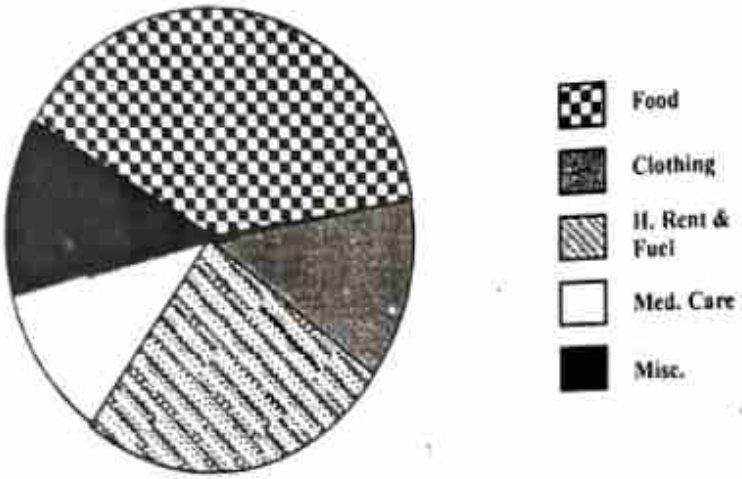
Items	Food	Clothing	H. Rent & Fuel	Med. Care	Misc.
Exp.	90	30	50	40	30

SOLUTION

The necessary calculations are given below.

Items	Expenditure (in Rs.)	Angle of Sectors (in Degrees)
Food	90	$\frac{90}{240} \times 360 = 135$
Clothing	30	$\frac{30}{240} \times 360 = 45$
H. Rent & Fuel	50	$\frac{50}{240} \times 360 = 75$
Medical Care	40	$\frac{40}{240} \times 360 = 60$
Misc.	30	$\frac{30}{240} \times 360 = 45$
Total	240	360

PIE DIAGRAM



2.8 GRAPHIC REPRESENTATION

Graphic representation means the visual representation of the numerical data. The movement of data can be presented very effectively by means of a graph. It is the simplest and most widely used method of presenting the data. A graph shows the relationship between two variables, for example, the amount of sales and the period of time when the sales were made. Graphs provide an overall picture of a statistical series. Graphs can also be used to make predictions and forecast. But graphs are less accurate and convey limited information. That is the only disadvantage of a graph.

2.8.1 Construction of Graphs

The first step in the construction of a graph is the drawing of two lines at right angles. The hoizontal line is called x-axis or Abscissa and the vertical line is known as y-axis or ordinate. These two lines together are known as co-ordinate axis. the point of inter section of the two axis is called the origin which is denoted by O. Some suitable scales are selected along x-axis and y-axis independent variable is taken along x-axis and dependent variable along y-axis. Points are plotted along both axis and joined to get the required graph.

2.9 TYPES OF GRAPHS

Graphs can be divided into two main groups, namely;

(a) Graphs of Time Series (Historigram)

(b) Graphs of Frequency distribution(Histogram)

2.9.1 Graph of Time Series (Historigram)

A curve showing the changes in the value of one or more than one items over a period of time is called the graph of a time series. A graph of time series is called

Historigram. The time is taken along x-axis and the other variable is along y-axis. Points are plotted and joined by means of a straight line to get the historigram.

EXAMPLE 2.11

The price of wheat per 40 kg during the years 1993 to 2000 are given below. Draw a Historigram.

Years	1993	1994	1995	1996	1997	1998	1999	2000
Price	130	160	160	173	240	240	240	300

HISTORIGRAM

2.9.2 Graphs of Frequency Distribution

The important graphs of frequency distribution are

(i) Histogram

(ii) Frequency Polygon

(iii) Frequency Curves

(iv) Cumulative frequency polygon (OGIVE)

2.9.3 Histogram

A frequency distribution can be shown in the form of a diagram. To construct histogram class boundaries are taken along x-axis and the frequency along y-axis. When the class intervals are equal, the rectangles are completed with class interval as the width and heights proportional to the frequencies. The rectangles are adjacent to each other and there is no gap between the rectangles. It is used to locate mode graphically.

If the class intervals are unequal, then the heights of the rectangles are adjusted. Each class frequency is divided by its interval size, then rectangles are completed with class interval as a width and heights proportional to the ratio of the frequencies.

EXAMPLE 2.12

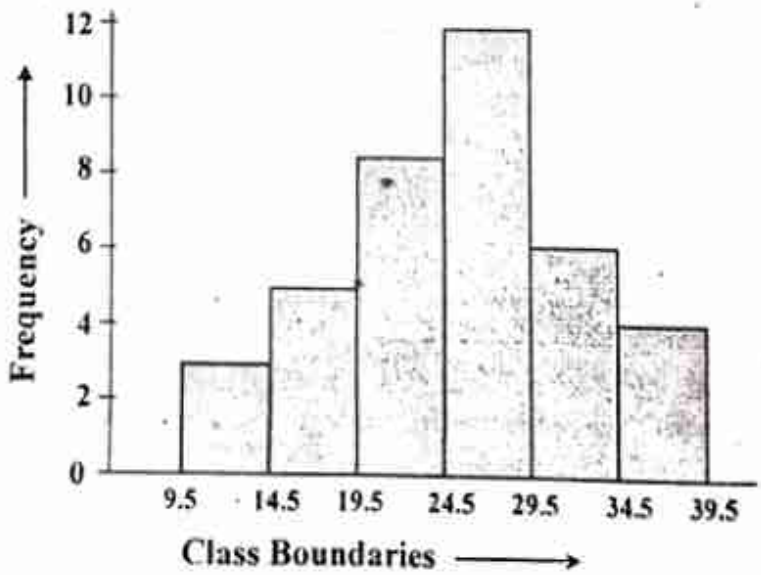
Construct Histogram for the following frequency distribution.

Groups	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39
Freq.	3	5	9	12	6	4

SOLUTION

Groups	f	Class Boundaries
10 – 14	3	9.5 – 14.5
15 – 19	5	14.5 – 19.5
20 – 24	9	19.5 – 24.5
25 – 29	12	24.5 – 29.5
30 – 34	6	29.5 – 34.5
35 – 39	4	34.5 – 39.5

HISTOGRAM



EXAMPLE 2.13

Construct a Histogram for the following frequency distribution.

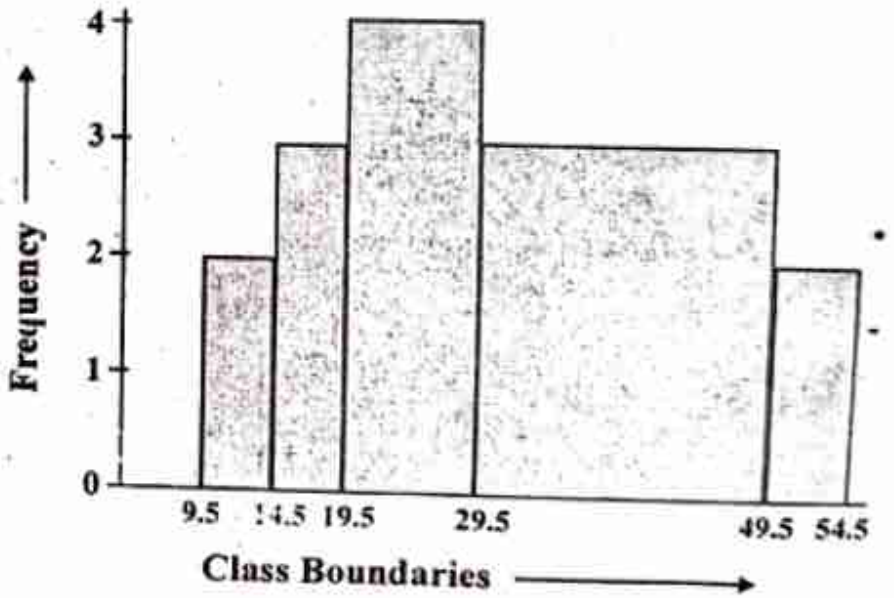
Classes	10–14	15–19	20–29	30–49	50–54
Frequency	10	15	40	60	10

SOLUTION

Here the class interval are unequal. The heights of the rectangles may be adjusted. Each class frequency is divided by its class interval size. The necessary calculations are given below.

Classes	Class interval	f	Proportional height	Class Boundaries
10 – 14	5	10	$\frac{10}{5} = 2$	9.5 – 14.5
15 – 19	5	15	$\frac{15}{5} = 3$	14.5 – 19.5
20 – 29	10	40	$\frac{40}{10} = 4$	19.5 – 29.5
30 – 49	20	60	$\frac{60}{20} = 3$	29.5 – 49.5
50 – 54	5	10	$\frac{10}{5} = 2$	49.5 – 54.5

HISTOGRAM FOR UNEQUAL CLASS INTERVAL



2.9.4 Frequency Polygon

Frequency polygon is obtained by joining the mid points of the tops of the rectangles of a histogram by means of a straight line. As we know that polygon is a many sided closed figure, therefore, we add extra classes at both ends of the frequency distribution with zero frequencies. By doing so, the polygon forms a closed figure.

EXAMPLE 2.14

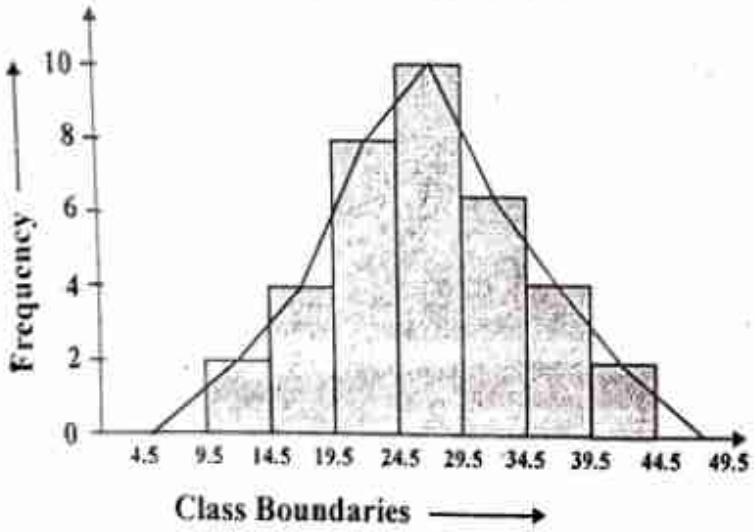
Draw a frequency polygon for the following frequency distribution.

Groups	10–14	15–19	20–24	25–29	30–34	35–39	40–44
Freq.	2	4	8	10	7	4	2

SOLUTION

Groups	f	Class boundaries
10 – 14	2	9.5 – 14.5
15 – 19	4	14.5 – 19.5
20 – 24	8	19.5 – 24.5
25 – 29	10	24.5 – 29.5
30 – 34	7	29.5 – 34.5
35 – 39	4	34.5 – 39.5
40 – 44	2	39.5 – 44.5

FREQUENCY POLYGON



2.9.5 Frequency Curves

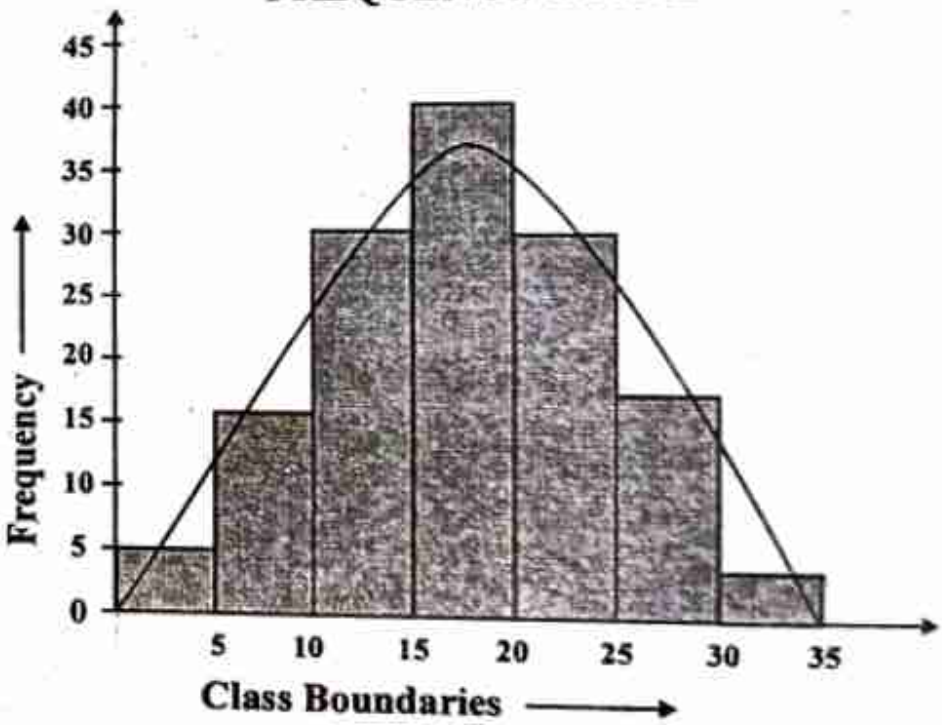
If the class intervals are very small and the number of classes is very large in a frequency distribution, then the histogram and frequency polygon approach a smooth curve *i.e.* there is no ups and downs in the curve. That smooth curve is called frequency curve. In the frequency curve the mid points of tops of rectangles of histogram are not joined together by straight lines. The free hand drawing method of drawing curve is used.

EXAMPLE 2.15

Draw a frequency curve of the following data.

Marks	0–5	5–10	10–15	15–20	20–25	25–30	30–35
No. of Students.	5	16	30	40	30	18	4

FREQUENCY CURVE



2.9.6 Cumulative Frequency Polygon (Ogive)

In cumulative frequency polygon, the cumulative frequencies are plotted against the upper class boundaries. The upper class boundaries are taken along x-axis and the cumulative frequency along y-axis. The plotted points are joined by a straight line. This curve is known as cumulative frequency polygon. It is also called Ogive. It is used to locate median, Quartiles, Deciles and Percentiles.

EXAMPLE 2.16

Draw cumulative frequency polygon or Ogive from the following data.

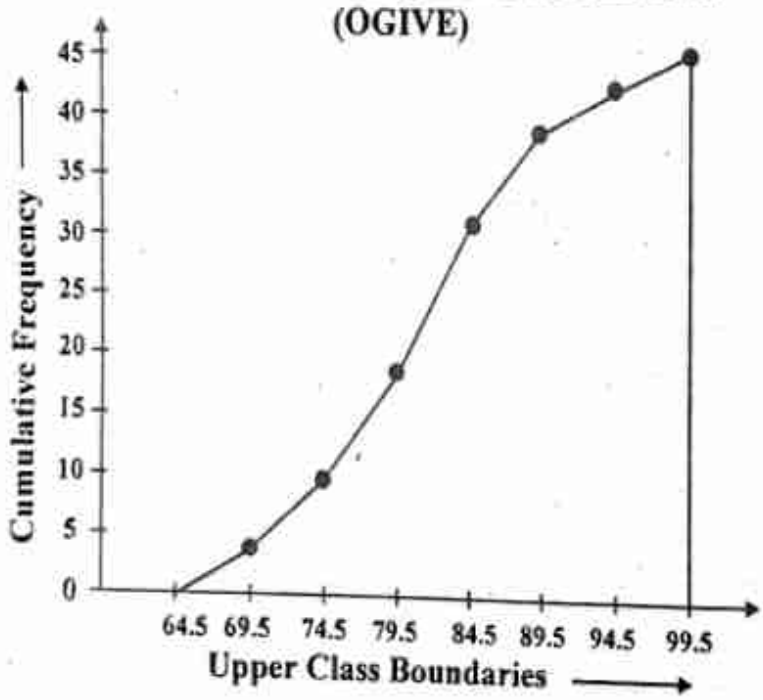
Daily wages	65 - 69	70 - 74	75 - 79	80 - 84	85 - 89	90 - 94	95 - 99
No. of Persons	4	6	9	12	8	4	2

SOLUTION

The necessary calculations are given below.

Classes	Class Boundaries	f	C.f.	U.C.B.
65 – 69	64.5 – 69.5	4	4	69.5
70 – 74	69.5 – 74.5	6	10	74.5
75 – 79	74.5 – 79.5	9	19	79.5
80 – 84	79.5 – 84.5	12	31	84.5
85 – 89	84.5 – 89.5	8	39	89.5
90 – 94	89.5 – 94.5	4	43	94.5
95 – 99	94.5 – 99.5	2	45	99.5

CUMULATIVE FREQUENCY POLYGON (OGIVE)



Difference between Histogram and Historigram

Histogram is the graph of frequency distribution, while Historigram is the graph displaying the variation in the time series dealing with prices, production, export and import etc.

EXERCISES

2.1 What is meant by classification? Write down the basis of classification and explain it.

2.2 Define the term tabulation. Discuss the main parts of a table.

2.3 (a) Differentiate between classification and tabulation.

(b) Distinguish between simple and compound tables.

2.4 **Define the following terms.**

(i) Frequency distribution

(ii) Class-interval

(iii) Class limits

(iv) Class marks / mid points

(v) Class boundaries

2.5 (a) What is meant by frequency distribution. Describe briefly the main steps involved in the preparation of frequency distribution.

(b) The following data shows the number of road side accidents per day.

2,5,1,2,0,6,3,4,5.1,2,5,3,0,6,4,0,6,3,2,4,3,6,7,2

Make a discrete frequency distribution.

The number of Persons in 20 families are given below:

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

Construct an ungrouped frequency distribution.

2.6 The number of children born to 30 women are given below.

5,3,3,2,0,6,1,4,1,6,3,4,8,7,2,4,5,2,4,6,4,3,3,0,5,2,1,4,5,3

Represent the data in the form of frequency distribution taking one as a class interval.

2.7. The weights in pounds of 30 college students are given below.

130, 133, 124, 121, 115, 139, 137, 144, 142, 133, 133, 128, 129, 132, 131, 128, 126, 132, 134, 135, 138, 130, 141, 136, 135, 141, 123, 126, 118, 134

Prepare a frequency distribution taking a class interval of size 5.

2.8 Tabulate the following marks into frequency distribution taking 10 as the class interval and 45 as the lower limit.

109, 74, 49, 103, 95, 90, 118, 52, 88, 101, 96, 72, 56, 64, 110, 97, 59, 62, 96, 82, 65, 85, 105, 116, 91, 83, 99, 52, 76, 84, 89, 77, 104

2.9 The following data give the index number of commodities in a certain year. Form a frequency distribution using suitable class interval.

70, 99, 114, 110, 88, 108, 87, 101, 71, 117, 110, 153, 145, 127, 98, 156, 127, 132, 113, 115, 100, 117, 73, 134, 108, 102, 123, 106, 119, 104, 129, 117, 106, 119, 116

2.10 Marks of students in English out of 100 are given below.

62, 59, 58, 42, 35, 60, 5, 12, 29, 52, 42, 38, 39, 45, 39, 28, 35, 37, 41, 51, 37, 39, 45, 16, 9, 17, 42, 59, 0, 23, 45, 33, 12, 40, 60

Construct

(i). A frequency distribution by using a class interval of 5.

(ii) A less than cumulative frequency.

(iii) A more than cumulative frequency.

(iv) Relative frequency.

2.11 (a) In an experiment measuring the percentage of shrinkage on dying 40 plastic clay test specimens gave the following results.

19.3, 16.9, 17.9, 17.3, 15.8, 18.5, 17.1, 19.5, 20.4, 18.7, 22.3, 17.5, 18.4, 13.9, 18.8, 16.8, 14.9, 19.5, 19.4, 16.3, 17.8, 23.4, 17.4, 19.0, 21.8, 18.8, 18.5, 18.2, 16.1, 18.3, 17.5, 17.4, 18.6, 16.9, 16.5, 18.2, 20.5, 20.5, 17.5, 19.1

Group these data into frequency distribution taking 1.00 as class interval. *e.g.* 13.5 – 14.4, 14.5 – 15.4 etc.

(b) Calculate class boundaries, mid point cumulative frequency and relative frequency of part (a) in question 2.11.

2.12 (a) What do you understand by diagrammatic representation of statistical data?

(b) Explain the following diagrams.

(i) Simple Bar Diagrams

(ii) Multiple Bar Diagrams

(iii) Sub-divided Rectangles

(iv) Pie Diagrams

2.13 The population of 5 big cities in Pakistan is given below. Draw a simple bar diagram.

Cities	Karachi	Lahore	Faisalabad	Rawalpindi	Multan
Pop. (in lakhs)	92.7	50.6	19.8	14.1	11.8

Source : Population Census Report, 1998.

2.14 **(a)** What is meant by tabulation.

(b) For the following data construct simple bar chart / diagram.

Years	1962	1963	1964	1965	1966	1967	1968	1969
Production.	1050	1200	1250	1370	1450	1500	1585	1620

2.15 **(a)** Male and female population of four provinces of Pakistan in 1998 are given below.

Provinces	Population (in millions)		
	Males	Female	Total
Punjab	37.5	35.1	72.6
Sindh	15.8	14.2	30.0
N.W.F.P	9.0	8.6	17.6
Balochistan	3.5	3.0	6.5

Construct

(i) Component Bar Diagram

(ii) Multiple Bar Diagram

(b) Draw a component bar chart for the following data:

(Population in Lakhs)

Division	Both Sexes	Male	Female
Peshawar	64	33	31
Rawalpindi	40	21	19
Sargodha	60	32	28
Lahore	65	35	30

2.16 **(a)** Draw sub-divided rectangles to compare the budgets of two **families** A and B.

Items of Expenditure	Family - A	Family - B
Food	240	600
Clothing	120	425
Fuel	40	125
Housing	90	300
Services	50	200
Misc.	60	150

(b) The following table gives the monthly budgets of two **families.** Represents the data by sub-divided rectangles.

Item of Expenditure	Family A	Family B
Food	6000	8000
Clothing	1000	1000
House Rent	4000	5000
Fuel & Lighting	1000	1000
Miscellaneous	3000	5000
Total	Rs. 15000	Rs. 20000

2.17 **(a)** Draw a Pie Diagram for the following data.

Items	Expenditure in Rs.
Food	95
Clothing	32
House Rent	50
Medical care	13
Misc.	50

(b) Represent the following data by a pie diagram.

Items	Expenditure in Rs.
Food	75
Clothing	50
House rent	30
Education	25
Miscellaneous	20

2.18 (a) What do you meant by graphic representation of data?

(b) Write down the various types of graphs. Explain it briefly.

2.19. (a) What is histogram. How does it differ from historigram?

(b) The heights of the college students are given below.

Heights (in inches)	57–59	60–62	63–65	66–68	69–71	72–74
No. of Students	8	15	27	18	9	3

Draw a histogram and frequency polygon.

2.20 (a) Explain the method of constructing histograms when the class intervals are unequal.

(b) Draw a histogram from the following data.

Classes	12–14	15–19	20–29	30–34	35–38	39–40
Freq.	12	25	70	25	16	6

2.21 (a) What is meant by classification.

(b) Prepare a frequency distribution from the following data taking class interval as

2.2 – 2.7, 2.8 – 3.3, ...

4.1, 3.5, 3.2, 4.2, 3.6, 3.5, 4.2, 4.8, 4.1, 4.3, 4.4, 4.3, 3.8, 4.2, 4.7, 2.8, 3.7, 4.9, 5.6, 3.3, 3.7, 2.6, 2.7, 4.7, 4.1, 4.2, 4.6, 4.1, 4.9, 3.7, 4.5, 3.7

(c) Draw a histogram for the distribution you obtain in part (b) above.

2.22 (a) Differentiate between histogram and historigram.

(b) Population of Pakistan (in millions) from 1965 - 90 is given below.

Years	1965	1970	1975	1980	1985	1990
Population (in Millions)	52.6	60.6	71.0	82.6	96.2	112.4

Source : Pakistan population welfare program, Report 1995, P-56.

Draw a historigram.

(c) For the following data draw a historigram.

Year	1962	1963	1964	1965	1966	1967	1968
Production	1050	1200	1250	1370	1450	1500	1600

2.23 (a) Define frequency polygon and cumulative frequency polygon or Ogive.

(b) Daily wages of factory workers are given below.

Wages	75–79	80–84	85–89	90–94	95–99	100–104
No. of Workers	2	4	8	11	13	7

Draw a cumulative frequency polygon or Ogive.

(c) Construct an ogive from the following table.

Weight	118-126	127-135	136-144	145-153	154-162	163-171	172-180
Frequency	3	5	9	12	5	4	2

2.24 Make a frequency distribution taking classes are; 1.20 – 1.49, 1.50 – 1.79 ,..... 3.20, 3.17, 2.87, 1,45, 1.49, 2.37, 2.86, 2.50, 1.67, 2.66, 3.18, 3.06, 2.56, 1.86, 1.99, 2.06, 2.45, 2.22, 3.10, 1.72, 2.04, 2.15, 2.45, 2.68, 2.75, 2.89, 1.44, 1.60, 1.54, 1.48.

2.25 Make a frequency distribution of the following marks obtained by students of a class in the subject of mathematics:

48, 88, 70, 72, 73, 49, 5, 17, 38, 40, 56, 49, 78, 98, 89, 14, 68, 87, 98, 99, 79, 49, 40, 52, 14, 34, 78, 29, 17, 83, 70, 68, 82, 55, 48, 31, 72, 5, 12, 9, 11, 48, 59, 71, 80, 72, 62, 66, 49, 48.

Taking class limits 1 – 20, 21 – 40 and so on.

2.26. SELECT THE CORRECT ANSWER

(i) The process of arranging the data into homogenous classes according to their similarities is called

(a) average (b) classification
(c) tabulation (d) data array

(ii) The basis of classification are

(a) two (b) three
(c) four (d) five

(iii) A table has at least

(a) one part (b) two parts
(c) three parts (d) four parts

(iv) In table, column caption is also called

(a) box head (b) stub
(c) body (d) title

(v) Total angle of the Pie chart is

(a) 180° (b) 270°
(c) 360° (d) 380°

(vi) In pie chart the arrangement of angles of different sectors are generally

(a) clockwise (b) anti clockwise
(c) random (d) none of these

(vii) The grouped data is also called

(a) raw data (b) primary data
(c) secondary data (d) simple values

(viii) The data related to export and import of a country should be presented by

(a) simple bar chart (b) multiple bar chart
(c) pie chart (d) aerial chart

(ix) When the classes are 1 – 3, 4 – 6, 7 – 9, ... the class interval is

(a) 2 (b) 3
(c) 4 (d) 5

(x) The total of relative frequency is always equal to

(a) 1 (b) 0.5
(c) –1 (d) 100

(xi) The number of values in any class is called

(a) mid point (b) class interval
(c) frequency (d) cumulative frequency

(xii) The difference between maximum and minimum value is called

(a) range (b) class interval

(c) mid point (d) frequency

(xiii) At G.P.O. sorting of letters is a type of

(a) one way tabulation (b) frequency distribution

(c) tabulation (d) classification

(xiv) Graph of the frequency distribution is called

(a) Histogram (b) Historigram

(c) Ogive (d) Frequency polygon

(xv) Graph of the time series is called

(a) Histogram (b) Historigram

(c) Ogive (d) Frequency polygon

(xvi) The number of observations falling in a particular class is called:

(a) Class mark (b) Class interval

(c) Mid point (d) None of these

(xvii) A pie chart is represented by:

(a) Rectangle (b) Square

(c) Circle (d) Triangle

(xviii) A sector diagram is also called:

(a) Bar diagram (b) Histogram

(c) Pie diagram (d) Component bar diagram

Answers 2.26

(i)	(b)	**(ii)**	(c)	**(iii)**	(d)	**(iv)**	(a)	**(v)**	(c)
(vi)	(b)	**(vii)**	(c)	**(viii)**	(b)	**(ix)**	(b)	**(x)**	(a)
(xi)	(c)	**(xii)**	(a)	**(xiii)**	(d)	**(xiv)**	(a)	**(xv)**	(b)
(xvi)	(d)	**(xvii)**	(c)	**(xviii)**	(c)				

SHORT QUESTION ANSWERS

1. **What is classification?**

Ans. The process of arranging the data into homogenous classes according to their resemblances and similarities is called classification *e.g.*, the process of sorting letters.

2. **What are the basis of classification?**

Ans. The basis of classification are

(a) Geographical or spatial

(b) Chronological or temporal

(c) Qualitative

(d) Quantitative

3. **Define Geographical or spatial classification.**

Ans. In geographical classification, the data are classified by geographical region or locations. For example, population of a country may be classified by provinces, divisions, districts and tehsils etc.

4. **Define chronological or temporal classification.**

Ans. When the data are classified according to the time of its occurrence such as years, months weeks, days, hours etc.

5. **Define qualitative classification.**

Ans. In qualitative classification, the data are classified on the basis of some quality or attribute, such as religion, sex, beauty, colour etc.

6. **What do you mean by quantitative classification?**

Ans. In this type data are classified according to some quantitative measurements *e.g.* height, income, age, sale etc.

7. **Define range.**

Ans. The difference between maximum and minimum value is called range.

i.e. $R = X_m - X_o$

8. **Define tabulation.**

Ans. The systematic arrangement of data into rows and columns is called tabulation.

9. What are the main parts of a table?

Ans. The main parts of the table is

(a) Title

(b) Prefatory notes

(c) Column captions and Box head

(d) Row captions and stub

(e) Body

(f) Foot notes

(g) Source notes

10. Define one way tabulation.

Ans. When the data are tabulated according to one characteristics is called one way tabulation. *e.g.* division wise population of Punjab.

11. Define two way tabulation.

Ans. When the data are tabulated according to two characteristics, then it is called two way tabulation. *e.g.* Division wise population of the Punjab by sex. (Male and Female)

12. Define frequency distribution.

Ans. The arrangement of data according to size and magnitude is called frequency distribution.

13. Write down the steps for construction of continuous frequency distribution.

Ans. Following steps are involved.

(a) Determine the range

(b) Decide the number of classes

(c) Class interval size

(d) Decide the starting point

(e) Determine the remaining class limits.

(f) Distribute the data into appropriate classes.

14. Define midpoint or class mark.

Ans. The average value of the lower and upper class limit is called midpoint or class mark.

15. What is the difference between class limits and class boundaries?

Ans. Class Limit

The values of the classes are called class limits. Every class has two values. The smaller is called lower class limit and the larger is called upper class limit.

Class Boundaries

The true or precise values, which describe the actual or true class limits of a class, are called class boundaries.

16. Define class frequency.

Ans. The number of values in any class is the class frequency of that class.

17. What do you mean by class interval?

Ans. The width of the class is called class interval. It is denoted by h.

18. Define cumulative frequency.

Ans. The total frequency of all classes less than the upper class boundary of a given class is called cumulative frequency of that class.

19. Define relative frequency.

Ans. The frequency of a class divided by the total frequency is called relative frequency of that class.

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Total frequency}}$$

20. What do you know about graphic representation?

Ans. Visual representation of data in the form of points, lines, symbols or areas are called Graphic representation.

21. What do you mean by pie chart?

Ans. Pie chart/diagram show the relationship between the whole and its components. In pie chart areas of the sectors are compared.

To construct a pie chart, draw a circle of suitable radius. To show the components parts by sectors, we calculate the angles for each sector by using the relation.

$$\text{Angle} = \frac{\text{Component Part}}{\text{Whole quantity}} \times 360^\circ$$

22. What do you know about cumulative frequency polygon or ogive.

Ans. It is used to locate median. In cumulative frequency polygon, the cumulative frequencies are plotted against the upper class boundaries. The upper class boundaries are taken along x-axis and the cumulative frequency along y-axis. The plotted points are joined by a straight line. This curve is known as cumulative frequency curve.

23. What is the difference between Histogram and Historigram?

Ans. Histogram is the graph of frequency distribution, while historigram is the graph of time series.

CHAPTER # 3

MEASURE OF CENTRAL TENDENCY

3.1 INTRODUCTION

An average is a single value, which represents the set of data as whole. Since the average tends to lie in the center of distribution they are also called measure of central tendency.

3.1.1 Qualities of Good Average

An average that possesses all or most of the following qualities is considered a good average.

1. It should be rigidly defined.
2. It should be easy to understand.
3. It should be easy to calculate.
4. It should be based on all the observations of the data.
5. It should be unaffected by extreme observations.
6. It should have sampling stability.

3.1.2 Types of Averages

The commonly used averages are

(i) Arithmetic mean
(ii) The Median
(iii) The Mode
(iv) Geometric mean
(v) Harmonic mean

3.2 ARITHMETIC MEAN

It is defined as the sum of all the observations divided by the number of observations. It is denoted by $\bar{X}$.

Let $X_1, X_2, \ldots, X_n$ be the 'n' observations then arithmetic mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + \ldots\ldots X_n}{n}$$

$$\bar{X} = \frac{\Sigma X}{n}$$

EXAMPLE 3.1

The marks obtained by 5 students are given below.

70, 50, 40, 35, 55

Calculate the arithmetic mean.

SOLUTION

$$\bar{X} = \frac{\Sigma X}{n}$$

$$= \frac{70 + 50 + 40 + 35 + 55}{5}$$

$$= \frac{250}{5} = 50 \text{ Marks}$$

3.2.1 Mean from Group Data

When the number of observations are very large, they are grouped into a frequency distribution.

Let $X_1, X_2, \ldots X_n$ be the mid points of different class intervals and let $f_1, f_2, \ldots, f_n$ be their corresponding frequencies, then arithmetic mean is given by

$$A.M. = \frac{f_1X_1 + f_2X_2 + \ldots\ldots + f_nX_n}{f_1 + f_2 + \ldots\ldots + f_n}$$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

EXAMPLE 3.2

The marks of 100 students in statistics are given below. Calculate the arithmetic mean.

Marks	30-35	35-40	40-45	45-50	50-55	55-60
No. of Stud.	14	16	18	23	18	11

SOLUTION

Marks	No. of Students (f)	Mid Points (x)	fx
30 - 35	14	32.5	455.00
35 - 40	16	37.5	600.00
40 - 45	18	42.5	765.00
45 - 50	23	47.5	1092.50
50 - 55	18	52.5	945.00
55 - 60	11	57.5	623.50
Σ	100	—	4490

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

$$= \frac{4490}{100} = 44.90$$

$$\bar{X} = 45 \text{ marks}$$

3.2.2 Mean from Frequency Distribution (Discrete Data)

In discrete series the values of each of the observation is multiplied by the corresponding frequency. These products are added. This sum is divided by the total frequency.

Let $X_1, X_2, \ldots\ldots X_n$ be the different values and let $f_1, f_2, \ldots\ldots f_n$ are corresponding frequencies, then mean is defined as

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \ldots\ldots + f_nX_n}{f_1 + f_2 + \ldots\ldots + f_n}$$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

EXAMPLE 3.3

Find the arithmetic mean from the following frequency distribution.

x	2	3	4	5	6
f	5	7	8	3	2

SOLUTION

x	f	fx
2	5	10
3	7	21
4	8	32
5	3	15
6	2	12
Σ	25	90

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{90}{25} = 3.6 \approx 4$$

EXAMPLE 3.4

The following table gives the marks obtained by a batch of 5 candidates in an examination in History, Statistics and Economics.

Roll No.	History	Statistics	Economics
1	41	46	50
2	35	50	52
3	38	39	41
4	34	50	46
5	30	38	39

In which subject is the level of knowledge highest?

SOLUTION

Roll No.	History	Statistics	Economics
1	41	46	50
2	35	50	52
3	38	39	41
4	34	50	46
5	30	38	39
Σ	178	223	228

Mean marks in History $= \bar{X}_h = \frac{\Sigma X}{n}$

$$= \frac{178}{5} = 35.6$$

Mean marks in Statistics $= \bar{X}_s = \frac{\Sigma X}{n}$

$$= \frac{223}{5} = 44.6$$

Mean marks in Economics $= \bar{X}_e = \frac{\Sigma X}{n}$

$$= \frac{228}{5} = 45.6$$

Mean marks in Economics are higher than Statistics and History. So the level of knowledge in Economics is highest.

EXAMPLE 3.5

Which class is better on the average?

Marks	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Class A	100	125	86	45	18	12
Class B	90	140	75	50	15	10

SOLUTION

The given data is group data first of all we calculate the average marks of class A and B separately.

Average marks of class A

Marks	Frequency (f)	X	fx
10 - 20	100	15	1500
20 - 30	125	25	3125
30 - 40	86	35	3012
40 - 50	45	45	2025
50 - 60	18	55	990
60 - 70	12	65	780
Σ	386	—	11430

$$\bar{X}_a = \frac{\Sigma fX}{\Sigma f}$$

$$= \frac{11430}{386} = 29.61$$

Average Marks of Class B

Marks	Frequency (f)	X	fx
10 - 20	90	15	1350
20 - 30	140	25	3500
30 - 40	75	35	2625
40 - 50	50	45	2250
50 - 60	15	55	825
60 - 70	10	65	650
Σ	**380**	—	**11200**

$$\bar{X}_b = \frac{\Sigma fX}{\Sigma f}$$

$$= \frac{11200}{380} = 29.47$$

Average marks of class A is more than Class B. So class A is better on the average.

3.2.3 Shortcut Method for Computing Mean

$$\bar{X} = A + \frac{\Sigma D}{n} \qquad \text{(For Discrete Values)}$$

Where $D = X - A$

A = Assumed mean

$$\bar{X} = A + \frac{\Sigma fD}{\Sigma f} \qquad \text{(Group Data)}$$

Where $D = X - A$

Σf = Total frequency

EXAMPLE 3.6

Calculate arithmetic mean by using shortcut method.

20, 25, 35, 45, 60

SOLUTION

Let $D = X - A$

$D = X - 35$

X	D = X – A (X – 35)
20	–15
25	–10
	–25
35	0
45	10
60	25
	35
Σ	10

$$\bar{X} = A + \frac{\Sigma D}{n}$$

$$= 35 + \frac{10}{5}$$

$$= 35 + 2 = 37$$

EXAMPLE 3.7

The scores of students in a cricket tournament are given below. Calculate the arithmetic mean by using shortcut method.

Scores	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49
No. of Students	5	15	12	10	8

SOLUTION

Scores	f	X	D = X – A (x – 24.5)	f.D
0 - 9	5	4.5	–20	–100
10 - 19	15	14.5	–10	–150
				–250
20 - 29	12	24.5	0	0
30 - 39	10	34.5	10	100
40 - 49	8	44.5	20	160
				260
Total	50	—	0	10

$$\bar{X} = A + \frac{\Sigma fD}{\Sigma f} \qquad \text{Let } A = 24.5$$

$$= 24.5 + \frac{10}{50} = 24.5 + 0.20$$

$$= 24.70 \approx 25 \text{ scores}$$

3.2.4 Step Deviation or Coding Method for Computing Mean

It is a very short method and should always be used for group data where class interval sizes are equal. The formula for computing mean is given below.

$$\bar{X} = a + \frac{\Sigma fU}{\Sigma f} \times h$$

Where a = Assumed mean

$$U = \frac{X - a}{h}$$

h = Class interval size.

EXAMPLE 3.8

Calculate A.M. from the following distribution by coding method.

Groups	1.5-2.0	2.0-2.5	2.5-3.0	3.0-3.5	3.5-4.0	4.0-4.5
Freq.	3	7	10	15	9	6

SOLUTION

Group	f	X	$u = \frac{X - 2.75}{0.5}$	fu
1.5 - 2.0	3	1.75	–2	–6
2.0 - 2.5	7	2.25	–1	–7
				–13
2.5 - 3.0	10	2.75	0	0
3.0 - 3.5	15	3.25	1	15
3.5 - 4.0	9	3.75	2	18
4.0 - 4.5	6	4.25	3	18
				51
Total	**50**	—	—	**38**

$$\bar{X} = a + \frac{\Sigma fu}{\Sigma f} \times h \qquad \text{Let } u = \frac{X - a}{h}$$

$$= 2.75 + \frac{38}{50} \times 0.5 \qquad a = 2.75$$

$$= 2.75 + \frac{19}{50} \qquad h = 0.5$$

$$= 2.75 + 0.38 = 3.13$$

3.2.5 Weighted Arithmetic Mean

Simple arithmetic mean gives equal importance to all the observations of the date, when the observations are not of equal importance, we assign them weights according to their relative importance.

Let $x_1, x_2, \ldots, x_n$ be 'n' observations with corresponding weights $w_1, w_2, \ldots, w_n$ respectively, then weighted mean is denoted by $\bar{X}_w$ and defined as

$$\bar{X}_w = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{\Sigma wx}{\Sigma w}$$

EXAMPLE 3.9

A student obtained the following marks at a certain examination English = 52, Urdu = 73, Physics = 84, Chemistry = 65 and Biology = 79. Find the weighted mean if weights of 2, 1, 3, 3 and 4 respectively are allotted to the subjects.

SOLUTION

Let X = Marks obtained

Subjects	x	w	wx
English	52	2	104
Urdu	73	1	73
Physics	84	3	252
Chemistry	65	3	195
Biology	79	4	316
Total	—	**13**	**940**

$$\bar{X}_w = \frac{\Sigma wx}{\Sigma w}$$

$$= \frac{940}{13} = 72.3 \approx 72 \text{ marks}$$

3.2.6 Properties of Arithmetic Mean

(i) The sum of deviations of the values X_i from their mean $\bar{X}$ is zero.

$\Sigma(X_i - \bar{X}) = 0$ (Ungroup Data)

$\Sigma f(X_i - \bar{X}) = 0$ (Group Data)

(ii) The sum of the squared deviations of values X_i from their mean is minimum.

i.e. $\Sigma(X_i - \bar{X})^2 \leq \Sigma(X_i - A)^2$ (Ungroup Data)

$\Sigma f(X_i - \bar{X})^2 \leq \Sigma f(X_i - A)^2$ (Group Data)

Where A is any value other than mean.

This is called minimal property of mean.

(iii) If $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ be the means of k groups with respective frequencies $n_1, n_2, \ldots, n_k$. Then the combined mean $\bar{X}_c$ for the whole distribution is given by

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{\Sigma n\bar{X}}{\Sigma n}$$

(iv) If $Y_i = ax_i + b \ (i = 1, 2, 3, \ldots, n)$

then $\bar{Y} = a\bar{X} + b$

Where a and b are constants.

EXAMPLE 3.10

A distribution consists of three components with sizes 100, 150 and 200 having their means 16, 19 and 22 respectively. Find the combined mean.

SOLUTION

We have given that

$n_1 = 100, \ n_2 = 150, \ n_3 = 200$

$\bar{X}_1 = 16, \ \bar{X}_2 = 19, \ \bar{X}_3 = 22$

Combined mean is given by

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1 + n_2 + n_3}$$

$$= \frac{100(16) + 150(19) + 200(22)}{100 + 150 + 200}$$

$$= \frac{8850}{450} = 19.67$$

3.2.7 When to Use Arithmetic Mean

We use arithmetic mean, when we are required to study social, economic and commercial problems like production, price, export and import. It helps in getting average income, average price, average production etc.

3.2.8 Advantages and Disadvantages of Arithmetic Mean

Advantages

(i) It is easy to calculate and simple to follow.

(ii) It is based on all the observations.

(iii) It can be determined for almost every kind of data.

(iv) It is commonly used average.

(v) It provides a good basis for comparison.

Disadvantages

(i) It is highly effected by extreme values.

(ii) It can not be accurately calculated for open end frequency distribution.

(iii) It can not be calculated accurately if any observation is missing.

3.3 MEDIAN

Median is the middle most value of a set of data when the data is arranged in order of magnitude.

If the number of observations in the array is odd, then median is the middle value and if the number of observations in the array is even, then median is the average of two middle values. It is denoted by $\tilde{X}$.

Mathematically

$$\text{Median} = \text{Value of}\left(\frac{n+1}{2}\right) \text{ the item.}$$

EXAMPLE 3.11

The scores of a cricket player in 7 matches is given below. Calculate median.

15, 45, 10, 16, 40, 19, 32

SOLUTION

Arranging the score in order we have

10, 15, 16, 19, 32, 40, 45

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

Here $n = 7$, *i.e.* odd therefore

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

$$= \text{value of 4th item}$$

4th item corresponds to 19. Therefore Median = 19.

EXAMPLE 3.12

Compute the median form the following data.

12, 13, 37, 14, 57, 48, 29, 27

SOLUTION

After arranging the data in ascending order we have

12, 13, 14, 27, 29, 37, 48, 57

Here $n = 8$, *i.e.* even, therefore

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

$$= \text{value of}\left(\frac{8+1}{2}\right)\text{th item}$$

$$= \text{value of 4.5 th item}$$

$$= \frac{\text{4th item} + \text{5th item}}{2}$$

$$= \frac{27+29}{2} = \frac{56}{2} = 28$$

Alternative Method

After arranging the data in ascending order, we have

12, 13, 14, 27, 29, 37, 48, 57 $\qquad n = 8$

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

$$= \text{Value of}\left(\frac{8+1}{2}\right)\text{th item.}$$

$$= \text{value of 4.5 th item}$$

= the value of 4th item + 0.5 (The value of 5th item − value of 4th item)

$$\text{Median} = 27 + 0.5\,(29 - 27)$$

$$= 27 + 0.5(2)$$

$$= 27 + 1 = 28$$

3.3.1 Median for Discrete Frequency Distribution

Let $X_1, X_2, \ldots\ldots X_n$ be the different values and let $f_1, f_2, \ldots\ldots, f_n$ are corresponding frequencies. First of all we calculate the cumulative frequencies and then we see the median number $\left(\frac{n+1}{2}\right)$ under the cumulative frequency column. the item, which corresponds to the median number is called median.

Mathematically

$$\text{Median} = \text{Value corresponding to}\left(\frac{n+1}{2}\right)\text{th cumulative frequency.}$$

EXAMPLE 3.13

The following table shows the number of heads in an experiment of 5 coins 100 times.

No. of heads	0	1	2	3	4	5
Freq.	10	25	30	20	10	5

Calculate median.

SOLUTION

X	f	C.f
0	10	10
1	25	25 + 10 = 35
2	30	30 + 25 = 65 → MEDIAN GROUP
3	20	20 + 65 = 85
4	10	10 + 85 = 95
5	5	5 + 95 = 100

$\Sigma f = 100 = n$

$$\text{Median} = \text{Value corresponding} \left(\frac{n+1}{2}\right) \text{th C.f}$$

$$= \text{Value corresponding to} \left(\frac{100+1}{2}\right) \text{th C.f}$$

$$= \text{Value of 50.5 th item}$$

50.5 th item corresponds to 2, therefore Median = 2

3.3.2 Median In Case Of Continuous Frequency Distribution (Group Data)

In case of group data, we form the cumulative frequencies and then calculate the median number $\left(\frac{n}{2}\right)$. The group, which corresponds to median number, is called the median group. The median lies in this group. Before calculating the median we convert the class limits into class boundaries if these are not given.

Median is given by the formula

$$\text{Median} = L + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

Where

$$\text{Median class} = \text{value of}\left(\frac{n}{2}\right)\text{th item}$$

L = Lower class boundary of the median class

h = Size of the class interval

f = Frequency of the median class

n = Σf = Total frequency

c = Cumulative frequency preceding the median class

EXAMPLE 3.14

The heights (in inches of 40 students of II years class are given below. Calculate the median.

Height (inch)	54 - 56	56 - 58	58 - 60	60 - 62	62 - 64	64 - 66
No. of Student	5	7	10	9	6	3

SOLUTION

Height	f	C.f
54 - 56	5	5
56 - 58	7	12
58 - 60	10	22 → Median Class
60 - 62	9	31
62 - 64	6	37
64 - 66	3	40

$$\text{Median} = \text{Value of}\left(\frac{n}{2}\right)\text{th item}$$

$$= \text{Value of}\left(\frac{40}{2}\right)\text{th item}$$

$$= \text{Value of 20th item}$$

The 20th item lies in the class 58 - 60. So 58 - 60 is the median class.

Therefore

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

Where L = 58, h = 2, f = 10, n = 40, c = 12

$$\text{Median} = 58 + \frac{2}{10}\left(\frac{40}{2} - 12\right)$$

$$= 58 + \frac{2}{10}(20 - 12)$$

$$= 58 + \frac{2}{10}(8) = 58 + \frac{16}{10}$$

$$= 58 + 1.6 = 59.6 \text{ inch}$$

EXAMPLE 3.15

Calculate median from the following frequency distribution.

Classes	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69
f	10	15	26	24	13	12

SOLUTION

Classes	C.B.	f	C.f
10 – 19	9.5 – 19.5	10	10
20 – 29	19.5 – 29.5	15	25
30 – 39	29.5 – 39.5	26	51 → **Median Class**
40 – 49	39.5 – 49.5	24	75
50 – 59	49.5 – 59.5	13	88
60 – 69	59.5 – 69.5	12	100

$$\text{Median} = \text{Value of}\left(\frac{n}{2}\right)\text{th item}$$

$$= \text{Value of}\left(\frac{100}{2}\right)\text{th item}$$

$$= \text{Value of 50 the item}$$

The 50th item lies in the group class 29.5 – 39.5. Therefore 29.5 – 39.5 is the median class.

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

Where L = 29.5, h = 10, f = 26, n = 100, c = 25

$$\text{Median} = 29.5 + \frac{10}{26}\left(\frac{100}{2} - 25\right)$$

$$= 29.5 + \frac{10}{26}(50 - 25)$$

$$= 29.5 + \frac{10}{26}(25)$$

$$= 29.5 + \frac{250}{26}$$

$$= 29.5 + 9.62 = 39.12$$

3.3.3 Graphic Location of Median

Median can be located by the graph of a cumulative frequency polygon (Ogive).

First of all a cumulative frequency curve is drawn for the given data. The values of the upper class boundaries are taking along x-axis and cumulative frequency along y-axis. Then the position of the median is located by the formula.

$$\text{Median} = \left(\frac{n}{2}\right)\text{th cumulative frequency}$$

A horizontal line is drawn along x-axis corresponding to the $\left(\frac{n}{2}\right)$ th position. That line intersects the ogive at certain point. Then from that point a perpendicular is drawn along x-axis which touches the x-axis on a certain point. This point on x-axis is the required value of the median.

EXAMPLE 3.16

Draw a cumulative frequency curve (Ogive) and locate median.

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Freq.	5	10	20	10	5

Check your result by actual calculations.

SOLUTION

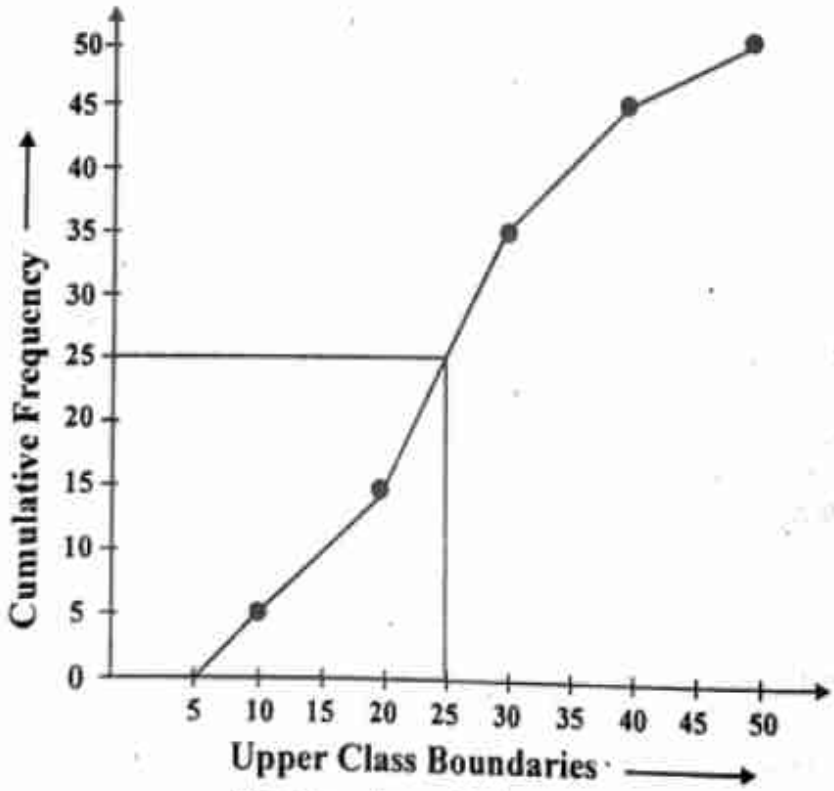
Marks	f	C.f	upper class boundaries
0 - 10	5	5	10
10 - 20	10	15	20
20 - 30	20	35	3
30 - 40	10	45	40
40 - 50	5	50	50

Median = Value corresponding to $\left(\frac{n}{2}\right)$ th c.f

= Value corresponding to $\left(\frac{50}{2}\right)$ th c.f

= Value corresponding to 25th c.f.

OGIVE



Median from Ogive = 25

Check

$$\text{Median} = \text{Value of the } \left(\frac{n}{2}\right) \text{th item}$$

$$= \frac{50}{2} \text{th item} = 25\text{th item}$$

The 25th item lies in the class 20 – 30. So 20 – 30 is the median class.

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

Where $L = 20, h = 10, f = 20, n = 50, c = 15$

$$\text{Median} = 20 + \frac{10}{20}\left(\frac{50}{2} - 15\right)$$

$$= 20 + \frac{10}{20}(25 - 15)$$

$$= 20 + \frac{10}{20}(10)$$

$$= 20 + \frac{10}{20}$$

$$= 20 + 5 = 25$$

This is as accurate as we can get with the graph.

3.3.4 Quantiles

When the number of observations is sufficiently large, the principle by which a distribution is divided into two equal parts may be extended to divide the distribution into four, ten, or hundred equal parts.

3.3.5 Quartiles

Quartiles are the values which divide the set of data into four equal parts. These are the first quartile Q_1, second quartile Q_2 (median) and third quartile Q_3.

The formula for the position of the quartiles are

$$Q_1 = \frac{n+1}{4} \text{ the value}$$

$$Q_2 = \frac{2(n+1)}{4} \text{ the value}$$

$$Q_3 = \frac{3(n+1)}{4} \text{ th value}$$

For the grouped data the quartiles are calculated as

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

$$Q_2 = l + \frac{h}{f}\left(\frac{2n}{4} - c\right)$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

3.3.6 Deciles

Deciles are the values which divide the set of data into 10 equal parts. These are denoted by $D_1, D_2, D_3, \ldots, D_9$. The formula for the position of deciles is

$$D_1 = \frac{n+1}{10} \text{ th value}$$

$$D_2 = \frac{2(n+1)}{10} \text{ the value}$$

$$D_3 = \frac{3(n+1)}{10} \text{ the value}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$D_9 = \frac{9(n+1)}{10} \text{ the value}$$

In case of grouped data, the deciles are calculated as

$$D_1 = l + \frac{h}{f}\left(\frac{n}{10} - c\right)$$

$$D_2 = l + \frac{h}{f}\left(\frac{2n}{10} - c\right)$$

$$D_3 = l + \frac{h}{f}\left(\frac{3n}{10} - c\right)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$D_9 = l + \frac{h}{f}\left(\frac{9n}{10} - c\right)$$

3.3.7 Percentiles

Percentiles are the value which divides the set of data into 100 equal parts. These are denoted by $P_1, P_2, \ldots, P_{99}$.

The formula for the position value of percentiles are

$$P_1 = \frac{n+1}{100} \text{ th value}$$

$$P_2 = \frac{2(n+1)}{100} \text{ th value}$$

$$P_3 = \frac{3(n+1)}{100} \text{ th value}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$P_{99} = \frac{99(n+1)}{100} \text{ th value}$$

In case of grouped data, the percentiles are calculated as

$$P_1 = l + \frac{h}{f}\left(\frac{n}{100} - c\right)$$

$$P_2 = l + \frac{h}{f}\left(\frac{2n}{100} - c\right)$$

$$P_3 = l + \frac{h}{f}\left(\frac{3n}{100} - c\right)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$P_{99} = l + \frac{h}{f}\left(\frac{99n}{100} - c\right)$$

EXAMPLE 3.17

From the data given below.

25, 40, 32, 62, 56, 38, 23, 63, 56, 42, 35, 50, 39, 47, 53

Find (i) Q_1 (ii) Q_2 (iii) Q_3 (iv) D_3
(v) D_8 (vi) P_{20} (vii) P_{80}

SOLUTION

Arranging the data in an order, we get

23, 25, 32, 35, 38, 39, 40, 42, 47, 50, 53, 56, 56, 62, 63

(i) $Q_1 = \frac{(n+1)}{4}$ the value

$= \frac{(15+1)}{4}$ the value $=$ 4th value

$Q_1 = 35$

(ii) $Q_2 = \frac{2(n+1)}{4}$ the value

$= \frac{2(15+1)}{4}$ the value $= \frac{32}{4}$ th value $=$ 8th value

$Q_2 = 42$

(iiii) $Q_3 = \frac{3(n+1)}{4}$ the value

$= \frac{3(15+1)}{4}$ the value $= \frac{48}{4}$ th value = 12 th value

$Q_3 = 56$

(iv) $D_3 = \frac{3(n+1)}{10}$ the value

$= \frac{3(15+1)}{10}$ the value $= \frac{48}{10}$ th value

$=$ 4th value $+ 0.8\,(5^{th} - 4^{th}$ value)

$= 35 + 0.8\,(38 - 35)$

$= 35 + 0.8(3) = 37.4$

$D_3 = 4.8$th value

(v) $D_8 = \frac{8(n+1)}{10}$ the value

$= \frac{8(15+1)}{10}$ the value $=$ 12.8 th value

$D_8 = 12.8$ th value

$=$ 12th value $+ 0.8\,(13^{th} - 12^{th}$ value)

$= 56 + 0.8\,(56 - 56)$

$= 35 + 0.8(0) = 56$

(vi) $P_{20} = \frac{20(n+1)}{100}$ the value

$= \frac{20(15+1)}{100}$ the value $= \frac{320}{100}$ th value

$= 3.2$th value

$=$ 3rd value $+ 0.2\,(4$th $-$ 3rd value)

$= 32 + 0.2(35 - 32)$

$= 32 + 0.2(3)$

$= 32 + 0.6 = 32.6$

(vii) $P_{80} = \frac{80(n+1)}{100}$ the value

$= \frac{80(15+1)}{100}$ the value

$= 12.8$ th value

$= 12$th value $+ 0.8$ (13th $-$ 12th value)

$= 56 + 0.8(56 - 56) = 56$

EXAMPLE 3.18

Marks of students in Mathematics are given below.

Marks	0 – 9	10 – 19	20 – 29	30 – 39
Frequency	2	5	10	20
Marks	40 – 49	50 – 59	60 – 69	70 – 79
Frequency	15	12	8	3

Calculate

(i) Q_1 (ii) Q_2 (iii) Q_3

(iv) D_3 (v) D_8 (vi) P_{30}

(vii) P_{75}

SOLUTION

Marks	f	Class boundaries	c.f
0 – 9	2	0 – 9.5	2
10 – 19	5	9.5 – 19.5	7
20 – 29	10	19.5 – 29.5	17
30 – 39	20	29.5 – 39.5	37
40 – 49	15	39.5 – 49.5	52
50 – 59	12	49.5 – 59.5	64
60 – 69	8	59.5 – 69.5	72
70 – 79	3	69.5 – 79.5	75
Total	75	—	—

(i) $Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$ ..... (i)

$Q_1 = \frac{n}{4}$ th value $= \frac{75}{4}$ th value $= 18.75$ th value

18.75 value lies in the group 29.5 – 39.50.

Therefore, $l = 29.5$, $h = 10$, $f = 20$, $n = 75$, $c = 17$

Putting the values in (i), we have

$$Q_1 = 29.5 + \frac{10}{20}(18.75 - 17)$$

$$= 29.5 + \frac{1}{2}(1.75) = 30.38$$

(ii) $Q_2 = l + \frac{h}{f}\left(\frac{2n}{4} - c\right) = \text{Median}$ ..... (ii)

$Q_2 = \frac{2n}{4}$ th value $= \frac{2(75)}{4} = 37.5$ th value

$= 37.5$th value

37.5th value lies in the group 39.5 – 49.5.

Therefore $l = 39.5$, $h = 10$, $f = 15$, $c = 37$

Putting the value in (ii), we have

$$Q_2 = 39.5 + \frac{10}{15}(37.5 - 37)$$

$$= 39.5 + \frac{10}{15}(0.5)$$

$$= 39.5 + 0.33 = 39.83$$

(iii) $Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$ ..... (iii)

$Q_3 = \frac{3n}{4}$ th value

$= \frac{3(75)}{4}$ th value $= 56.25$ th value

56.25 th value lies in the group 49.5 – 59.5

So $l = 49.5$, $h = 10$, $f = 12$, $c = 52$

Putting the values in (iii), we get

$$Q_3 = 49.5 + \frac{10}{12}(56.25 - 52)$$

$$= 49.5 + \frac{10}{12}(4.25)$$

$= 49.5 + 3.54$

$= 53.04$

(iv) $D_3 = l + \frac{h}{f}\left(\frac{3n}{10} - c\right)$ ..... (iv)

$D_3 = \frac{3n}{10}$ th value

$= \frac{3 \times 75}{10}$ th value $= 22.5$ th value

22.5 th value lies in the group 29.5 – 39.5

So $l = 29.5,\ h = 10,\ f = 20,\ c = 17$

Putting the values in (iv), we get

$D_3 = 29.5 + \frac{10}{20}(22.5 - 17)$

$= 29.5 + \frac{1}{2}(5.5)$

$= 29.5 + 2.75 = 32.25$

(v) $D_8 = l + \frac{h}{f}\left(\frac{8n}{10} - c\right)$ ..... (v)

$D_8 = \frac{8n}{10}$ th value

$= \frac{8 \times 75}{10}$ th value $= 60$ th value

60 th value lies in the group 49.5 – 59.5

So $l = 49.5,\ h = 10,\ f = 12,\ c = 52$

Putting the values in (v), we get

$D_8 = 49.5 + \frac{10}{12}(60 - 52)$

$= 49.5 + \frac{10}{12}(8)$

$= 49.5 + 6.67 = 56.17$

(vi) $P_{30} = l + \frac{h}{f}\left(\frac{30n}{100} - c\right)$ ..... (vi)

$P_{30} = \frac{30n}{100}$ th value

$= \frac{30 \times 75}{100}$ th value

$= 22.5$ th value

22.5 th value lies in the group 29.5 – 39.5

So $l = 29.5,\ h = 10,\ f = 20,\ c = 17$

$$P_{30} = 29.5 + \frac{10}{20}(22.5 - 17)$$

$$= 29.5 + \frac{1}{2}(5.5)$$

$$= 29.5 + 2.75 = 32.25$$

(vii) $P_{75} = l + \frac{h}{f}\left(\frac{75n}{100} - c\right)$ ..... (vii)

$$P_{75} = \frac{75n}{100} \text{ th value}$$

$$= \frac{75 \times 75}{100} \text{ th value}$$

$$= 56.25 \text{ th value}$$

56.25 th value lies in the group 49.5 – 59.5

So $l = 49.5,\ h = 10,\ f = 12,\ c = 52$

$$P_{75} = 49.5 + \frac{10}{12}(56.25 - 52)$$

$$= 49.5 + \frac{10}{12}(4.25)$$

$$= 49.5 + 3.54 = 53.04$$

3.3.8 When we Apply Median

We apply median to the situations, when the direct measurements of the variables are not possible like poverty, beauty and intelligence etc.

3.3.9 Advantages and Disadvantages of Median

Advantages

(i) It is easy to calculate and understand.

(ii) It is not effected by extreme values.

(iii) It can be computed even in open end frequency distribution.

(iv) It can be used for qualitative data.

(v) It can be located graphically.

Disadvantages

(i) It is not rigorously defined.

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

(iii) It is not suitable for further algebraic treatment.

3.4 THE MODE

Mode is the value, which occurs maximum number of times in a set of data.

Or

It is the value which occurs most frequently in a set of data. A set of data may have more than one mode or no mode.

Mode = $\hat{X}$ = The value appearing maximum no. of times.

EXAMPLE 3.19

Calculate the mode for the following data.

(i) 10, 12, 6, 8, 7, 15, 10

(ii) 2, 7, 5, 2, 6, 9, 10, 6

(iii) 50, 60, 70, 60, 80, 60, 50

(iv) 20, 25, 28, 12, 18, 23

(v) Letters of PAKISTAN.

(vi) The days of the week for marriages in Pakistan.

(vii) Shoe sizes sold at a store

32, 34, 32, 35, 36, 35, 38, 40, 36, 35

SOLUTION

(i) Mode = 10, because it is repeated 2 times.

(ii) There are two modes

i.e. 2 and 6, because both are repeated 2 times.

(iii) Mode = 60, because it is repeated 3 times.

(iv) There is no mode, because all the values occur only once.

(v) Modal letter = A, because it is repeated 2 times.

(vi) Modal day = Sunday, because maximum number of marriages are held on Sunday.

(vii) Mode = 35, because it is repeated 3 times.

EXAMPLE 3.20

Calculate A.M. and Mode from the following values.

48, 34, 69, 25, 22, 34, 56, 68

SOLUTION

$$\bar{X} = \frac{\Sigma x}{n}$$

$$= \frac{48 + 34 + 69 + 25 + 22 + 34 + 56 + 68}{8}$$

$$= \frac{356}{8} = 44.50$$

Mode

Mode = 34, because it is repeated 2 times.

Mode for Discrete Frequency Distribution

Mode = The value corresponding to maximum frequency

EXAMPLE 3.21

The following are the sizes of shoes sold at a Bata shop. Find the modal size of the shoes.

Size of the shoes	4	5	6	7	8	9	10
Freq.	10	12	18	10	25	7	2

SOLUTION

Size of the Shoe	Frequency
4	10
5	12
6	18
7	10
8	25
9	7
10	2

Mode = The value corresponding to maximum frequency

Mode = 8 *i.e.* Modal size of the shoes is 8.

3.4.1 Mode in case of Group Data or Mode for Continuous Frequency Distribution

In case of group data, the mode lies in the class, which has maximum frequency. This class is called modal class. The mode is given by the formula

$$\text{Mode} = L + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \times h$$

Where

L = Lower class boundary of the modal class.

f_m = Frequency of the modal class.

f_1 = Frequency of the class preceding the modal class.

f_2 = Frequency of the class following the modal class

h = Size of the class interval

Modal class = A class which has maximum frequency.

EXAMPLE 3.22

Calculate the mode form the following frequency distribution.

Classes	0 – 8	8 – 16	16 – 24	24 – 32	32 – 40
Frequency	5	7	12	8	6

SOLUTION

Classes	Frequency (f)
0 – 8	5
8 – 16	$7 = f_1$
16 – 24	$12 = f_m$
24 – 32	$8 = f_2$
32 – 40	6

$$\text{Mode} = L + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \times h$$

Modal class is 16 – 24, because it has maximum frequency.

Therefore

$l = 16,\ f_m = 12,\ f_1 = 7,\ f_2 = 8,\ h = 8$

$$\text{Mode} = 16 + \frac{(12 - 7)}{(12 - 7) + (12 - 8)} \times 8$$

$$= 16 + \frac{5}{5 + 4} \times 8$$

$$= 16 + \frac{5}{9} \times 8$$

$$= 16 + \frac{40}{9}$$

$$= 16 + 4.44 = 20.44$$

3.4.2 Graphic Location of Mode

Mode can be located graphically by the following process.

(i) Draw histogram for the given data, taking class boundaries along x-axis and frequencies along y-axis.

(ii) The highest rectangle represents the modal class.

(iii) The top right and left corners of the modal class rectangle are joined with the top right and left corner of the proceeding rectangles.

(iv) From the point of intersection of both the lines a perpendicular on the x-axis is drawn. The point where the perpendicular meets the x-axis, will be the required value of the mode.

EXAMPLE 3.23

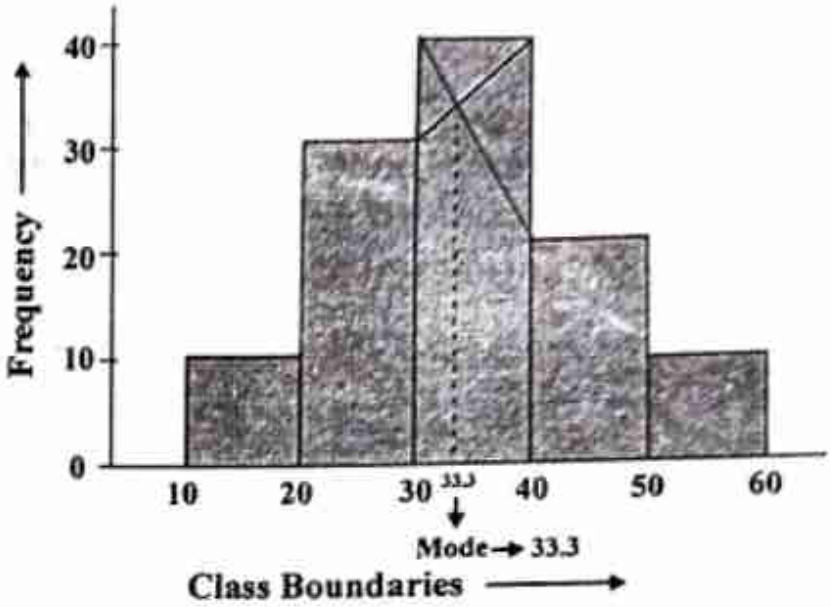
From the following frequency distribution.

Groups	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
f	10	30	40	20	10

Find approximate value of the mode from the histogram and check your answer by calculation.

SOLUTION

HISTOGRAM



Check by Calculations

Groups	f
10 – 20	10
20 – 30	30 = f_1
30 – 40	40 = f_m
40 – 50	20 = f_2
50 – 60	10

Applying the formula

$$\text{Mode} = L + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \times h$$

Modal class is 30 – 40 therefore

$L = 30,\ f_m = 40,\ f_1 = 30,\ f_2 = 20,\ h = 10$

$$\text{Mode} = 30 + \frac{(40 - 30)}{(40 - 30) + (40 - 20)} \times 10$$

$$= 30 + \frac{10}{10 + 20} \times 10$$

$$= 30 + \frac{100}{30}$$

$$= 30 + 3.33 = 33.33$$

The value of the mode from graphic method is also 33.3.

3.4.3 When to apply Mode

We apply mode when it is required to study the problems like average size of shoes, average size of ready made garments, average size of agriculture holding etc. This average is widely used in Biology and Meteorology.

3.4.4 Advantages and Disadvantages of Mode

Advantages

(i) It is easy to understand.

(ii) It is not effected by extreme values.

(iii) It can be computed even in open-end classes.

(iv) It can be useful in qualitative data.

Disadvantages

(i) It is not clearly defined.

(ii) It is not suitable for further algebraic treatment.

(iii) It is not based on all the observations.

(iv) It may not exist in some cases.

3.4.5 Empirical Relation Between Mean, Median and Mode

In a symmetrical distribution, mean, median and mode are coincide.

i.e. Mean = Median = Mode

In a moderately skewed distribution

Mode = 3 Median − 2 Mean.

EXAMPLE 3.24

In a moderately skewed distribution

(i) Mean = 15, Median = 16

Find the value of Mode.

(ii) Median = 65 and Mode = 85

Find the value of Mean.

SOLUTION

(i) Mode is given by the relation

Mode = 3 median − 2 mean

Mean = 15

Median = 16

Mode = 3(16) − 2(15)

= 48 − 30 = 18

(ii) Median = 65, Mode = 85

Mean = ?

Mode = 3 Median − 2 Mean

85 = 3(65) − 2 Mean

85 = 195 − 2 Mean

2 Mean = 195 − 85

2 Mean = 110

$$\text{Mean} = \frac{110}{2} = 55$$

EXAMPLE 3.25

Compute mean, median and mode for the data given below.

Classes	f	Classes	f
0.7312 – 0.7313	10	0.7320 – 0.7321	30
0.7314 – 0.7315	15	0.7322 – 0.7323	8
0.7316 – 0.7317	20	0.7324 – 0.7325	2
0.7318 – 0.7319	25		

SOLUTION

(i) **Arithmetic Mean**

Classes	f	x	fx
0.7312 – 0.7313	10	0.73125	7.31250
0.7314 – 0.7315	15	0.73145	10.97175
0.7316 – 0.7317	20	0.73165	14.63300
0.7318 – 0.7319	25	0.73185	18.29615
0.7320 – 0.7321	30	0.73205	21.96150
0.7322 – 0.7323	8	0.73225	5.85800
0.7324 – 0.7325	2	0.73245	1.46490
Total	110	—	80.49790

$$\bar{X} = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{80.49790}{110} = 0.73180$$

Calculation for Median and Mode

Classes	f	c.f	C.B.
0.7312 – 0.7313	10	10	0.73115 – 0.73135
0.7314 – 0.7315	15	25	0.73135 – 0.73155
0.7316 – 0.7317	20	45	0.73155 – 0.73175
0.7318 – 0.7319	25	70	0.73175 – 0.73195
0.7320 – 0.7321	30	100	0.73195 – 0.73215
0.7322 – 0.7323	8	108	0.73215 – 0.73235
0.7324 – 0.7325	2	110	0.73235 – 0.73255

Median

$$\text{Median} = \left(\frac{n}{2}\right)\text{th item}$$

$$= \frac{110}{2}\text{th item} = 55\text{th item}$$

55th item lies in the class 0.713175 – 0.73195. So this is median class.

Where L = 0.73175, h = 0.0002,

f = 25, n = 110, c = 45

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

$$= 0.73175 + \frac{0.0002}{25}\left(\frac{110}{2} - 45\right)$$

$$= 0.73175 + \frac{0.0002}{25}(55 - 45)$$

$$= 0.73175 + \frac{0.0002}{25}(10)$$

$$= 0.73175 + \frac{0.002}{25}$$

$$= 0.73175 + 0.00008 = 0.73199$$

(iii) Mode

Modal class

The class which has maximum frequency 0.73195–0.73215 is the modal class

$$\text{Mode} = L + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \times h$$

Where

$L = 0.73195,\ f_m = 30,\ f_1 = 25,\ f_2 = 8,\ h = 0.0002$

$$\text{Mode} = 0.73195 + \frac{(30 - 25)}{(30 - 25) + (30 - 8)} \times 0.0002$$

$$= 0.73195 + \frac{5}{5 + 22} \times 0.0002$$

$$= 0.73195 + \frac{0.00100}{27}$$

$$= 0.73195 + 0.00004 = 0.73199$$

EXAMPLE 3.26

Find the median and mode from the following data.

X	1.8	1.6	1.4	1.2	1.0	0.8	0.6	0.4	0.2
f	6	9	15	24	33	20	13	7	5

SOLUTION

Writing the data in an ascending order.

X	f	C.f	C.B. $\left(x - \frac{h}{2}, x + \frac{h}{2}\right)$
0.2	5	5	0.1 – 0.3
0.4	7	12	0.3 – 0.5
0.6	13	25	0.5 – 0.7
0.8	20	45	0.7 – 0.9
1.0	33	78	0.9 – 1.1
1.2	24	102	1.1 – 1.3
1.4	15	117	1.3 – 1.5
1.6	9	126	1.5 – 1.7
1.8	6	132	1.7 – 1.9
	132		

(i) Median

$$\text{Median} = \left(\frac{n}{2}\right) \text{th item}$$

$$= \frac{132}{2} \text{ th item}$$

$$= 66 \text{ th item}$$

66th item lies in the class 0.9–1.1. Therefore 0.9–1.1 is the median class.

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

Where

$L = 0.9, h = 0.2, f = 33, n = 132, c = 45$

$$\text{Median} = 0.9 + \frac{0.2}{33}\left(\frac{132}{2} - 45\right)$$

$$= 0.9 + \frac{0.2}{33}(66 - 45)$$

$$= 0.9 + \frac{0.2}{33}(21)$$

$$= 0.9 + \frac{4.2}{33}$$

$$= 0.9 + 0.127 = 1.03$$

(ii) Mode

Modal Class

The class which has maximum frequency.

i.e. 0.9 – 1.1 is the modal class.

Mode is given by the relation.

$$\text{Mode} = L + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \times h$$

$$= 0.9 + \frac{(33 - 20)}{(33 - 20) + (33 - 20)} \times 0.2$$

$$= 0.9 + \frac{13}{13 + 9} \times 0.2$$

$$= 0.9 + \frac{13 \times 0.2}{22}$$

$$= 0.9 + \frac{2.6}{22}$$

$$= 0.9 + 0.118 = 1.02$$

3.5 GEOMETRIC MEAN

Let $x_1, x_2, \ldots, x_n$ be the n positive values, then the geometric mean may be defined as nth positive root of their product.

i.e. $$G = \sqrt[n]{x_1 \cdot x_2 \ldots\ldots x_n}$$

$$G = (x_1 \cdot x_2 \ldots\ldots x_n)^{1/n}$$

When n is large, the calculation of the geometric mean becomes difficult and time consuming. For this reason we use the formula in terms of logarithims.

$$G = \text{antilog}\left[\frac{1}{n}(\Sigma \log x)\right]$$

If the data is organized into frequency distribution, then the above formula becomes

$$G = \text{antilog}\left[\frac{1}{\Sigma f}(\Sigma f \log x)\right]$$

The geometric mean is appropriate to average ratios and rate of change.

EXAMPLE 3.27

Find the geometric mean of the data given below.

10, 12, 18, 20, 25, 30, 35

SOLUTION

x	log x
10	1.0000
12	1.0792
18	1.2553
20	1.3010
25	1.3979
30	1.4771
35	1.5441
Σ	**9.0546**

$$G = \text{antilog}\,\frac{1}{n}(\Sigma \log x)$$

$$= \text{antilog}\,\frac{1}{7}(9.0546)$$

$$= \text{antilog}\,(1.2935) = 19.6569$$

EXAMPLE 3.28

Calculate geometric mean of the data given below.

Classes	1 – 3	4 – 6	7 – 9	10 – 12	13 – 15
Frequency	1	2	3	2	1

SOLUTION

Classes	F	x	log x	f log x
1 – 3	1	2	0.3010	0.3010
4 – 6	2	5	0.6990	1.3980
7 – 9	3	8	0.9031	2.7093
10 – 12	2	11	1.0414	2.0828
13 – 15	1	14	1.1461	1.1461
				7.6372

$$G = \text{antilog}\,\frac{1}{\Sigma f}(\Sigma f \log x)$$

$$= \text{antilog}\,\frac{1}{9}(7.6372)$$

$$= \text{antilog}\,(0.8486) = 7.0563$$

3.5.1 Advantages and disadvantages of Geometric Mean

Advantages

(i) It is based on all the observations.

(ii) It is rigorously defined.

(iii) It gives equal weightage to all the observations.

(iv) It is not much affected by sampling variability.

Disadvantages

(i) In case of negative value, it can not be calculated.

(ii) It vanishes if any observation is zero.

3.6 HARMONIC MEAN

Harmonic may be defined as the reciprocal of the arithmetic mean of the reciprocal of the values.

i.e. $H = \dfrac{n}{\Sigma\left(\dfrac{1}{x}\right)}$ Where $x \neq 0$

If the data is arranged in the form of frequency distribution, then Harmonic mean is defined as

$$H = \frac{\Sigma f}{\Sigma f\left(\frac{1}{x}\right)} \quad \text{Or} \quad H = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)}$$

EXAMPLE 3.29

A motorcycle is running at the rate of 15 km/hour during the first 60 km at 20 km/hour, during second 60 km, 30 km/hour, during the 3rd 60 km. What would be the average speed?

SOLUTION

$$H = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

$$= \frac{3}{\frac{1}{15} + \frac{1}{20} + \frac{1}{30}}$$

$$= \frac{3}{0.067 + 0.050 + 0.033}$$

$$= \frac{3}{0.1503} = 19.9557 \text{ km/hour}$$

EXAMPLE 3.30

Calculate Harmonic mean of the following frequency distribution.

Classes	0 - 4	4 - 8	8 - 12	12 - 16	16 - 20	20 - 24	24 - 28
Frequency	2	5	7	8	7	4	1

SOLUTION

Classes	f	x	f/x
0 - 4	2	2	1.0000
4 - 8	5	6	0.8383
8 - 12	7	10	0.7000
12 - 16	8	14	0.5714
16 - 20	7	18	0.3889
20 - 24	4	22	0.1816
24 - 28	1	26	0.0385
Σ	34	—	3.7137

$$\text{H.M.} = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)} = \frac{34}{3.7137} = 9.1553$$

3.6.1 Advantages and disadvantages of Harmonic Mean

Advantages

(i) It is rigorously defined by mathematical formula.

(ii) It is based on all the observations.

(iii) It is not much affected by sampling stability.

Disadvantages

(i) It is not easily understandable.

(ii) It can not calculate if any observation is zero.

(iii) It gives high weightage to small values.

SUMMARY

The formulae and the methods of computing mean, median, mode, geometric mean and harmonic mean are summarized below:

APPLICATION	FORMULA
Arithmetic Mean:	
Ungrouped data	
Direct method	$\overline{X} = \frac{\Sigma X}{n}$
Short cut method	$\overline{X} = A + \frac{\Sigma D}{n}$
Grouped Data	$\overline{X} = \frac{\Sigma fx}{\Sigma f}$
Short cut method	$\overline{X} = A + \frac{\Sigma fD}{\Sigma f}$, $D = X - A$
Step deviation method	$\overline{X} = A + \frac{\Sigma fu}{\Sigma f} \times h$, $u = \frac{X - A}{h}$
Weighted mean	$\overline{X}_w = \frac{\Sigma WX}{\Sigma W}$
Combined mean	$\overline{X}_c = \frac{\Sigma n_i \overline{X}_i}{\Sigma n_i}$
Median	
Ungrouped data	Median $= \frac{n+1}{2}$ th value
Grouped data	Median $= l + \frac{h}{f}\left(\frac{n}{2} - C\right)$
Graphically	From Ogive

Mode:	
Ungrouped data	Value which occurs maximum no. of times
Grouped data	$\text{Mode} = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$
Graphically	From histogram
Empirical relation	Mode = 3 Median – 2 Mean
Geometric Mean	
Ungrouped data	$G = \text{antilog}\,\frac{1}{n}(\Sigma \log x)$
Grouped data	$G = \text{antilog}\,\frac{1}{\Sigma f}(\Sigma f \log x)$
Harmonic Mean:	
Ungrouped data	$H = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$
Grouped data	$H = \frac{\Sigma f}{\Sigma(f/x)}$

EXERCISES

3.1 Define statistical average. What are the qualities of good average?

3.2 (a) Define and explain arithmetic mean.

(b) Calculate arithmetic mean for the following values.

(i) 1, 2, 3, 4, 5

(ii) 3, 5, 7, 5, 7, 9

(iii) 15, 18, 7, 12, 13, 20, 23, 18

(iv) 48, 34, 69, 25, 22, 34, 56, 68

(c) The weights of 10 students are given below. Find Mean.

40.2, 50.3, 43.1, 50.0, 52.8, 60.0, 57.2, 48.5, 49.5, 52.0

3.3 (a) From the following data. Calculate arithmetic mean.

x	0	1	2	3	4	5
f	40	20	50	30	20	10

(b) Given below the frequency distribution of number of classrooms in different colleges. Find arithmetic mean.

No. of Rooms	25	28	35	40	45
No. of Colleges	3	7	10	3	2

3.4 (a) The marks obtained by 100 students are given below. Calculate average marks.

Marks	10–25	25–40	40–55	55–70	70–85	85–100
Freq.	6	20	44	26	3	1

(b) The following frequency distribution shows the daily income of 100 house holds in a locality.

Income	35–39	40–44	45–49	50–54	55–59	60–64	65–69
Freq.	13	15	28	17	12	10	5

Calculate arithmetic mean.

(c) The mean age of a group of 100 persons was found to be 35. Later it was discovered that age 50 was misread as 25. Find the corrected mean.

3.5 The height of college students measured to nearest inches is given below.

Height (inches)	60 – 62	63 – 65	66 – 68	69 – 71	72 – 74
No. of Student	05	18	42	27	08

Calculate arithmetic mean.

3.6 (a) The marks in English of Ist year class are given below.

Marks	15–19	20–24	25–29	30–34	35–39
f	9	18	35	17	5

Calculate arithmetic mean by using.

(i) Shortcut method

(ii) Step deviation method

(b) The following data have been obtained from a frequency distribution of continuous variable X after making substitution.

$$U = \frac{X - 130}{5}$$

U	–3	–2	–1	0	1	2	3
f	5	12	26	32	13	8	4

(i) Find arithmetic mean by direct method.

(ii) Find arithmetic mean by step deviation method.

3.7 (a) Define the weighted arithmetic mean.

(b) Calculate the weighted mean for the following items.

Items	Expenditure	Weights
Food	290	7.5
Rent	54	2.0
Clothing	98	1.5
Fuel & Light	75	1.0
Misc.	75	0.5

(c) A student's final marks in Statistics, Computer science, mathematics and English are 75, 62, 85 and 57. Find the weighted arithmetic mean if weights of 3, 4, 2 and 1 respectively are attached to the subjects.

3.8 (a) A distribution consists of four components with frequencies 15, 12, 16, 21 having their means 16.5, 20.3, 21.6 and 26.2. Find the mean of the combined distribution.

(b) The average marks obtained by the students of three sections in statistics class are given below.

Section	No. of Students	Average Marks
A	50	75
B	75	62
C	60	68

Find average marks of the whole class.

(c) A variable Y is determined from a variable X by the equation $Y = 10 - 4X$.

Find Y when, X = -3, -2, -1, 0, 1, 2, 3, 4, 5 and show that $\overline{y} = 10 - 4\overline{x}$.

3.9 (a) Deviations from X = 10.5 of 10 items are given below.

−1.3, 2.0, 2.9, 7.5, −4.6, −3.4, 8.2, 9.3, −7.4, 5.6

Calculate the arithmetic mean.

(b) Find the mean from the following observations and show that

$\Sigma(X - \bar{X}) = 0$

6.5, 2.33, 7.4, 7.25, 6.50, 9.7, 8.35, 2.6, 2.43

(c) From the following frequency distribution show that sum of deviations of values from their mean is zero.

i.e. $\Sigma f(X - \bar{X}) = 0$

Classes	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
F	5	12	15	25	8	3	2

3.10 (a) Define median. Given a simple example of median.

(b) Calculate median from the following data.

(i) Heights of 9 students in inches are

60, 58, 54, 53, 55, 63, 51, 52, 57

(ii) No. of Road side accidents in 10 cities are

10, 6, 12, 8, 7, 15, 13, 5, 7, 14

(iii) Wages of 10 workers are

88, 70, 72, 125, 115, 95, 81, 90, 95, 90

3.11 (a) Calculate median for the following data.

No. of children	0	1	2	3	4	5	6	7
Frequency	2	3	4	6	5	4	3	3

(b) Given below is the frequency distribution of family members in 50 families. Find the median as the average family size.

No. of persons	2	3	4	5	6
No. of families	5	8	12	20	5

3.12 Compute the median from the following data.

Groups	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69
F	15	17	24	31	21	12

3.13 Daily wages of factory workers are given below.

Daily wages	40-60	60-80	80-100	100-120
No. of Workers	13	23	101	182
Daily wages	120-140	140-160	160-180	
No. of Workers	105	19	7	

Calculate the median.

3.14 (a) Give the advantages and Disadvantages of median.

(b) Pocket money of the student during convocation is given below. Locate the median graphically.

Rs.	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30
No. of Students	12	9	18	7	4

Also check your result by actual calculations.

3.15 (a) Define mode. Explain it with example.

(b) Find the mode for the following data.

(i) 7, 4, 9, 10, 12, 15, 7, 9, 7

(ii) 70, 75, 60, 62, 68, 70, 60

(iii) 10, 8, 15, 12, 6, 9, 7, 16

3.16 (a) Calculate the mode from the following data.

Size of shoes	3	4	5	6	7	8	9	10
Freq.	12	18	20	13	18	27	19	9

(b) Calculate the mode from the following frequency distribution showing the number of children per house.

No. of Children	0	1	2	3	4	5	6	7	8	9	10
No. of houses	4	10	12	18	25	20	16	8	4	2	1

(c) Calculate median and mode from the following frequency distribution.

No. of Children	0	1	2	3	4	5	6	7
f	1	3	3	6	5	4	2	1

3.17 (a) The age distribution of employees in a factory is given below. Find the model age of the employees.

Age of Employee (Years)	15–20	20–25	25–30	30–35	35–40
Frequency.	5	23	58	104	141
Age of Employee (Years)	40–45	45–50	50–55	55–60	
Freq.	98	43	19	6	

(b) The following table shows the distribution of maximum loads in short tons supported by certain cables produced by a company.

Maximum Load	f
9.3 – 9.7	2
9.8 – 10.2	5
10.3 – 10.7	12
10.8 – 11.2	17
11.3 – 11.7	14
11.8 – 12.2	6
12.3 – 12.7	3
12.8 – 13.2	1

Determine mean, median and mode.

3.18 (a) The weight of 40 male students at a college are given in the following frequency distribution.

Weight	118-126	127-135	136-144	145-153	154-162	163-171	172-180
Freq.	3	5	9	12	5	4	2

Calculate the mean, median and mode.

(b) Compute Mean, Median and Mode from the following data.

x	2.4	2.0	1.6	1.2	0.8	0.4
f	6	9	15	30	12	8

3.19 (a) The daily profits in rupees of 120 shops are given below.

Profit per shop	0-100	100-200	200-300	300-400	400-500	500-600
No. of Shops	15	18	25	40	15	7

Draw histogram. Calculate the mode graphically and verify the result by actual calculations.

(b) From the following data find the missing frequency when mean is 15.38.

x	10	12	14	16	18	20
f	3	7	?	20	8	5

3.20 (a) Discuss the empirical relation between mean, median and mode.

(b) For a certain frequency distribution the value of mean was 11 and the median was 12. Find the approximate value of the mode.

(c) Find the value of mode by using empirical relation between averages for the following data.

Marks	10–19	20–29	30–39	40–49	50–59
No. of Students	5	25	40	20	10

3.21 **Which average will be suitable to compare.**

(i) Height of students.

(ii) Size of shoes

(iii) Intelligence of students.

(iv) Number of Petals of a flower.

(v) Average income of different people.

(vi) Average size of ready-made garments.

(vii) Weight of students.

(viii) Marks obtained by the students.

3.22 Find the mean and Median of the following data.

Height (inches)	45 - 50	50 - 55	55 - 60	60 - 65	65 - 70	70 - 75
No. of Persons	2	7	12	18	13	3

3.23 Compute Mean and Median.

Monthly income (Rs.)	No. of Families	Monthly income (Rs.)	No. of families
110 – 119	2	160 – 169	18
120 – 129	4	170 – 179	13
130 – 139	17	180 – 189	6
140 – 149	28	190 – 199	5
150 – 159	25	200 – 209	2

3.24 (a) Calculate Q_1, Q_3, D_7, Median, P_{20} and P_{60} from the following data.

Classes	1-10	11-20	21-30	31-40	41-50	51-60	61-70
F	2	4	7	8	6	3	2

(b) Calculate Q_1, Q_3, D_5 and P_{80} from the following data.

20, 35, 18, 27, 35, 40, 48, 33, 42, 35, 28

(c) Find the median, the quartiles 8th docile and 65th percentile for the distribution of examination marks given below:

Marks	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of students	8	87	190	304	211	85	20

3.25 (a) Define geometric mean.

(b) What do you meant by Harmonic mean?

3.26 (a) Find the geometric mean of the following data.

(i) 3, 5, 6, 6, 7, 10, 12

(ii) 15, 18, 17, 20, 25, 23

(iii) 75, 62, 76, 78, 59, 67

(iv) 110, 115, 108, 112, 120, 128, 130

(v) 45, 32, 37, 46, 39, 36, 41, 48, 36

(b) Find geometric mean from the following frequency distribution.

x	2	3	4	5	6
f	5	7	8	3	2

3.27 (a) Given the following frequency distribution of weights, calculate geometric mean.

Weight (grams)	65-84	85-104	105-124	125-144	145-164	165-184	185-204
f	9	10	17	10	5	4	5

(b) Give the following frequency distribution.

Classes	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39
f	15	17	25	18	12

Find the geometric mean.

3.28 **Calculate the geometric mean from the following frequency distribution.**

Classes	100 - 200	200 - 300	300 - 400	400 - 500	500 - 600
f	15	18	30	20	17

3.29 Aslam gets a rise of 15% in salary at the end of his first year of service and further 25% and 30% at the end of the second and third year respectively. The rise in each case being calculated on his salary at the beginning of the year. To what annual percentage increase in this equivalent?

(b) A man traveling 100 kilometer has 5 stages at equal interval. The speed of the man in the various stages was observed to 10, 16, 20, 14, 15 kilometer per hour. Find the average speed at which the man travels.

3.30 (a) Define Harmonic mean

(b) Calculate Harmonic mean from the following data.

(i) 1, 2, 3, 4, 5

(ii) 12, 18, 16, 20, 25, 30

(iii) 75, 60, 65, 85, 60, 50

(c) A man traveling 100 kilometer has 5 stages at equal interval. The speed of the man in the various stages was observed to 10, 16, 20, 14, 15 kilometer per hour. Find the average speed at which the man travels.

3.31 The reciprocal of 11 values of x are given below.

0.0500, 0.0454, 0.0400, 0.0333, 0.0285, 0.0232, 0.0213, 0.0200, 0.0182, 0.0151, 0.0143

Calculate harmonic mean and arithmetic mean.

3.32 (a) Find the harmonic mean from the following data.

Weight (gm)	65 – 84	85 – 104	105 – 124	125 – 144
f	9	10	17	10
Weight (gm)	145 – 164	165 – 184	185 – 204	
f	5	4	5	

(b) Compute Harmonic mean from the following data.

Hourly Wages	40-50	50-60	60-70	70-80	80-90
Frequency	4	8	16	8	4

3.33 For the data given below calculate

(i) Arithmetic mean

(ii) Harmonic Mean

(iii) Median

Marks	30 – 40	40 – 50	50 – 60	60 – 70
frequency	1	3	12	15
Marks	70 – 80	80 – 90	90 – 100	
frequency	14	11	5	

3.34 Calculate Mean, Median and Geometric mean from the following data

Daily wages (Rs)	No. of Workers
Below 05	7
Below 10	52
Below 15	107
Below 20	170
Below 25	215
Below 30	259
Below 35	285
Below 40	300

3.35 CHOOSE THE CORRECT ANSWER.

(i) **The sum of the deviations taken from mean is**

(a) Always equal to zero

(b) Sometimes equal to zero

(c) Never equal to zero

(d) None of these

(ii) **The median of 12, 5, 6, 8 and 4 is**

(a) 7 (b) 6

(c) 8 (d) 12

(iii) **The mode of 10, 8, 6, 5, 6, 7 is**

(a) 10 (b) 7

(c) 6 (d) 5

(iv) **The mode of "STATISTICS" is**

(a) S (b) S and T

(c) C (d) none of these

(v) **Most suitable average for qualitative data is**

(a) Mean (b) Median

(c) Mode (d) None of these

(vi) **In a symmetrical distribution, mean, median and mode are**

(a) Equal (b) Unequal

(c) Sometimes equal (d) None of these

(vii) **For any two values, the mean and median are always**

(a) Equal (b) Unequal

(c) ill define (d) None of these

(viii) **The arithmetic mean of 5 values is 10, then sum of these values is**

(a) 5 (b) 10

(c) 15 (d) 50

(ix) **If $\Sigma(x - 20) = 0$, then mean will be**

(a) 10 (b) 15

(c) 20 (d) 0

(x) The most frequent value in the set of data if it exists is called

(a) Mean (b) Median

(c) Mode (d) None of these

(xi) When the observations are not of equal importance then we use

(a) Simple mean (b) Weighted mean

(c) Combined mean (d) Median

(xii) For graphic representation of median, we have to draw

(a) Histogram (b) Historigram

(c) Ogive (d) Pie chart

(xiii) For graphic representation of mode, we have to draw

(a) Histogram (b) Historigram

(c) Frequency polygon (d) Ogive

(xiv) If the data contains an extreme value, the suitable average is

(a) Mean (b) Median

(c) Mode (d) Weight mean

(xv) A distribution which has one mode is called

(a) Unimodal (b) Bimodal

(c) Multimodal (d) None of these

(xvi) A distribution which has two modes is called

(a) Uni-model (b) Bi-model

(c) Multi-model (d) None of these

(xvii) Sum of squared deviation of the observations from mean is

(a) zero (b) 1

(c) Minimum (d) None of these

(xviii) In case of skewed distribution, mean median and mode are

(a) equal (b) sometimes equal

(c) not equal (d) none of these

(xix) The mean is based on:

(a) All values (b) Small values

(c) Extreme values (d) Non of these

(xx) Arithmetic mean of the positive numbers a and b is:

(a) $\frac{ab}{2}$ (b) $\frac{2a}{b}$

(c) $\frac{a+b}{2}$ (d) $\frac{2}{a+b}$

(xxi) If $\bar{x} = 10$ and $y = 2x + 5$, then $\bar{y}$ is equal to:

(a) 20 (b) 25

(c) 30 (d) 50

(xxii) Data must be arranged before calculating:

(a) Mode (b) Median

(c) Mean (d) None of these

(xxiii) The sum of deviation from mean is:

(a) Minimum (b) Negative

(c) Zero (d) None of these

(xxiv) Change of origin and scale is used for calculation of the:

(a) Mean (b) Median

(c) Mode (d) None of these

(xxv) Mode for the word <u>PROFESSOR</u> is

(a) R (b) S

(c) O (d) R, S and O

(xxvi) The mean, median and mode for constant "a" is

(a) 0 (b) $\frac{a}{2}$

(c) a (d) a^2

Answers 3.35

(i)	(a)	**(ii)**	(b)	**(iii)**	(c)	**(iv)**	(b)	**(v)**	(c)
(vi)	(a)	**(vii)**	(a)	**(viii)**	(d)	**(ix)**	(c)	**(x)**	(c)
(xi)	(b)	**(xii)**	(c)	**(xiii)**	(a)	**(xiv)**	(b)	**(xv)**	(a)
(xvi)	(b)	**(xvii)**	(c)	**(xviii)**	(c)	**(xix)**	(a)	**(xx)**	(c)
(xxi)	(b)	**(xxii)**	(b)	**(xxiii)**	(c)	**(xxiv)**	(a)	**(xxv)**	(d)
(xxvi)	(c)								

SHORT QUESTION ANSWERS

1. Define average.

Ans. An average is a single value, which represents the set of data as a whole.

2. What are the qualities of good average.

Ans.
1. An average should be rigidly defined.
2. It should be easy to understand.
3. It should be easy to calculate.
4. It should be based on all the observations.
5. It should be unaffected by extreme observation.

3. What are the commonly used average?

Ans. The commonly used averages are

1. Arithmetic mean 2. Median 3. Mode
4. Geometric mean 5. Harmonic mean

4. Define arithmetic mean.

Ans. It is defined as sum of all the observations divided by the number of observation. It is denote by $\bar{X}$.

i.e. $\bar{X} = \frac{\Sigma x}{n}$

5. Write down at least two properties of mean.

Ans. 1. The sum of deviations of the observations from their mean is zero.

$$\Sigma(X - \bar{X}) = 0$$

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

2. The sum of squared deviations of observations from mean is minimum.
3. Combined mean is

$$\bar{X}_c = \frac{\Sigma n_i \bar{X}_i}{\Sigma n_i}$$

6. Define median.

Ans. Median is the middle most value of a set of data, when arranged in order of magnitude.

7. Give the merits of the median.

Ans.
(a) It is easy to calculate and understand.
(b) It is not affected by extreme values.
(c) It can be computed even in open end classes.
(d) It is possible to locate graphically.

8. **Define Mode.**

Ans. Mode is the value which occurs maximum number of times in the set of data.

9. **Define weighted mean.**

Ans. When the observations are not of equal importance we assign weights to their relative importance.

i.e. $\bar{X}_w = \frac{\Sigma WX}{\Sigma W}$

10. **What is the empirical relation between mean, median and mode?**

Ans. In symmetrical distribution mean, median and mode are equal.

In a moderately skewed distribution.

Mode = 3 Median − 2 Mean

11. **For a certain frequency distribution, the value of mean is 15 and median is 20. What will be value of mode?**

Ans. We know that

Mode = 3 median − 2 mean

= 3 (20) − 2(15)

= 60 − 30 = 30

12. **Define Geometric mean**

Ans. Let $x_1, x_2, \ldots\ldots x_n$ be n positive values, then geometric mean may be defined as nth positive root of their product

i.e. $G = \sqrt[n]{x_1 \cdot x_2 \ldots\ldots x_n}$

$G = (x_1 \cdot x_2 \ldots\ldots x_n)^{1/n}$

13. **Define Harmonic mean**

Ans. Harmonic mean may be defined as the reciprocal of the arithmetic mean of the reciprocal of the values.

i.e. $H = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$

CHAPTER # 4

MEASURE OF DISPERSION

4.1 INTRODUCTION

The measure of central tendency does not tell us any thing about the spread of the data, because any two sets of data may have the same central tendency with vast difference magnitude of their variability. Consider two types of data;

(a) 10, 12, 11, 14, 13

(b) 2, 10, 18, 27, 3

These two data have same mean 12, but differ in their variations. There is more variation in data (b) as compared to data (a). This illustrates the fact that measure of central tendency is not sufficient. We there for need some additional information concerning with how the data are dispersed about the average. This is done by measuring the dispersion. By dispersion we means the degree to which numerical data tend to spread about an average value.

There are two types of measures of dispersion.

(i) Absolute dispersion

(ii) Relative dispersion

(i) Absolute Measure of Dispersion

An absolute dispersion is one that measure the dispersion in term of some units. For example, if the units of data are rupees, kilograms, centimeter etc. the units of the measure of dispersion will also be rupees, kilograms, centimeter etc.

(ii) Relative Measure of Dispersion

It is expressed in the form of ratio, coefficient and it is independent of the units of measurements. It is useful for comparison of data of different nature.

MEASURES OF DISPERSION

The main measure of dispersion are the followings.

(i) The Range

(ii) The semi Interquartile Range or the Quartile Deviation

(iii) The Mean Deviation

(iv) The variance and the standard deviation

4.2 THE RANGE

It is defined as the difference between the largest and the smallest observations in a set of data.

Range = R = $X_m - X_o$

Where X_m = The largest observation

X_o = The smallest observation

The range is very simple measure of variability and only takes into account two most extreme observations.

Its relative measure known as the co-efficient of dispersion.

$$\text{Co-efficient of Range} = \frac{X_m - X_o}{X_m + X_o}$$

EXAMPLE 4.1

Calculate Range and Co-efficient of Range from the following data.

15, 20, 18, 16, 30, 42, 12, 25

SOLUTION

$X_m = 42,\ X_o = 12$

$$R = X_m - X_o$$
$$= 42 - 12$$
$$= 30$$

$$\text{Coefficient of Range} = \frac{X_m - X_o}{X_m + X_o}$$

$$= \frac{42 - 12}{42 + 12} = \frac{30}{54} = 0.56$$

EXAMPLE 4.2

Calculate Range and Coefficient of Range from the following data.

Classes	1 – 3	4 – 6	7 – 9	10 – 12	13 – 15
Freq.	1	2	3	2	1

SOLUTION

R = Mid value of the highest class − Mid value of the largest class

Classes	f	x
1 − 3	1	2
4 − 6	2	5
7 − 9	3	8
10 − 12	2	11
13 − 15	1	14

$X_m = 14$, $X_o = 2$

$$R = X_m - X_o$$
$$= 14 - 2$$
$$= 12$$

$$\text{Co-efficient of Range} = \frac{X_m - X_o}{X_m + X_o}$$
$$= \frac{14 - 2}{14 + 2} = \frac{12}{16} = 0.75$$

4.3 THE SEMI INTER QUARTILE RANGE OR QUARTILE DEVIATION

The semi-inter quartile range is defined as half of the difference between the third and the first quartiles.

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

Where Q_1 and Q_3 are the first and the third quartiles of the data.

Its relative measure called the coefficient of quartile deviation is defined by the relation.

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

EXAMPLE 4.3

Given below are the marks obtained by 9 students.

45, 32, 36, 37, 46, 39, 36, 48, 41

Find the quartile deviation and coefficient of quartile deviation.

SOLUTION

Arranging the data in an order

32, 36, 36, 37, 39, 41, 45, 46, 48

n = 9

$$Q_1 = \left(\frac{n+1}{4}\right) \text{the value}$$

$$= \left(\frac{9+1}{4}\right) \text{th value}$$

$$= 2.5\text{th value}$$

$$= 2\text{nd value} + 0.5\,(3\text{rd value} - 2\text{nd value})$$

$$= 36 + 0.5(36 - 36)$$

$$= 36 + 0.5(0) = 36 \text{ marks}$$

$$Q_3 = 3\left(\frac{n+1}{4}\right) \text{th value}$$

$$= 3\left(\frac{9+1}{4}\right) \text{th value}$$

$$= \frac{30}{4} \text{th value}$$

$$= 7.5\text{th value}$$

$$= 7\text{th value} + 0.5(8\text{th value} - 7\text{the value})$$

$$= 45 + 0.5(46 - 45)$$

$$= 45 + 0.5(1) = 45.5 \text{ marks}$$

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

$$= \frac{45.5 - 36}{2} = \frac{9.5}{2} = 4.75$$

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

$$= \frac{45.5 - 36}{45.5 + 36} = \frac{9.5}{81.5} = 0.117$$

EXAMPLE 4.4

Compute the quartile deviation and coefficient of quartile deviation from the following data

Groups	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69
f	15	17	24	31	21	12

SOLUTION

First of all we calculate Q_1 and Q_3

Groups	f	C.B.	C.f
10 – 19	15	9.5 – 19.5	15
20 – 29	17	19.5 – 29.5	32
30 – 39	24	29.5 – 39.5	56
40 – 49	31	39.5 – 49.5	87
50 – 59	21	49.5 – 59.5	108
60 – 69	12	59.5 – 69.5	120

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right) \quad \text{..... (i)}$$

$$Q_1 = \frac{n}{4} \text{ th value}$$

$$= \frac{120}{4} \text{ th value}$$

$$= 30\text{th value}$$

30th value lies in the class 19.5 – 29.5

Therefore

$l = 19.5,\ h = 10,\ f = 17,\ n = 20,\ C = 15$

Putting these values in (i), we have

$$Q_1 = 19.5 + \frac{10}{17}\left(\frac{120}{4} - 15\right)$$

$$= 19.5 + \frac{10}{17}(30 - 15)$$

$$= 19.5 + \frac{10}{17}(15)$$

$$= 19.5 + \frac{150}{17}$$

$$= 19.5 + 8.82 = 28.32$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - C\right) \quad \text{..... (ii)}$$

$$Q_3 = \frac{3n}{4} \text{ th value}$$

$$= \frac{3(120)}{4} \text{ th value}$$

$$= 90\text{th value}$$

So 90th value lies in the group 49.5 – 59.5

Therefore

$l = 49.5,\ h = 10,\ f = 21,\ c = 87$

Putting the values in (ii), we have

$$Q_3 = 49.5 + \frac{10}{21}\left(\frac{3 \times 120}{4} - 87\right)$$

$$= 49.5 + \frac{10}{21}(90 - 87)$$

$$= 49.5 + \frac{10}{21}(3)$$

$$= 49.5 + 1.43 \quad = 50.93$$

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

$$= \frac{50.93 - 28.32}{2} = \frac{22.61}{2} = 11.31$$

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

$$= \frac{50.93 - 28.32}{50.93 + 28.32} = \frac{22.61}{79.25} = 0.285$$

4.4 THE MEAN DEVIATION OR AVERAGE DEVIATION

Mean deviation is defined as the mean of the absolute deviations measured either from mean, median or mode. By absolute deviations we mean that all the deviations are positive.

$$\text{M.D.} = \frac{\Sigma\,|X - \bar{X}|}{n} \text{(M.D. from Mean)}$$

$$\text{M.D.} = \frac{\Sigma\,|X - \text{Median}|}{n} \text{(M.D. from Median)}$$

$$\text{M.D.} = \frac{\Sigma\,|X - \text{Mode}|}{n} \text{(M.D. from Mode)}$$

(In case of Group Data)

(i) Mean deviation from Mean $= \dfrac{\Sigma f\,|X - \bar{X}|}{\Sigma f}$

(ii) Mean deviation from Median $= \dfrac{\Sigma f\,|X - \text{Median}|}{\Sigma f}$

(iii) Mean deviation from Mode $= \dfrac{\Sigma f\,|X - \text{Mode}|}{\Sigma f}$

$$\text{Co-efficient of M.D. (Mean)} = \frac{\text{Mean deviation from Mean}}{\text{Mean}}$$

$$\text{Co-efficient of M.D. (Median)} = \frac{\text{M.D. from Median}}{\text{Median}}$$

$$\text{Co-efficient of M.D. (Mode)} = \frac{\text{M.D. from Mode}}{\text{Mode}}$$

EXAMPLE 4.5

Calculate mean deviation from the mean and the median for the values 30, 36, 32, 33, 35, 39, 36.5, 35 and 34

SOLUTION

$$\text{Mean} = \frac{\Sigma x}{n} = \frac{310.5}{9} = 34.5$$

To find the median, we first arrange the values
30, 32, 33, 34, 35, 35, 36, 36.5 and 39

$\tilde{X} = 35$

X	$\lvert X - \bar{X} \rvert$	$\lvert X - \tilde{X} \rvert$
30	$\lvert 30 - 34.5 \rvert = 4.5$	5.0
36	1.5	1.0
32	2.5	3.0
33	1.5	2.0
35	0.5	0
39	4.5	4.0
36.5	2.0	1.5
35	0.5	0
34	0.5	1.0
310.5	$\Sigma \lvert X - \bar{X} \rvert = 18.0$	**17.50**

$$\text{M.D. (From Mean)} = \frac{\Sigma \lvert X - \bar{X} \rvert}{n}$$

$$= \frac{18}{9} = 2.0$$

$$\text{M.D. (From Median)} = \frac{\Sigma |X - \tilde{X}|}{n}$$

$$= \frac{17.5}{9} = 1.94$$

EXAMPLE 4.6

Compute mean deviation from Mean. Also calculate the coefficient of mean deviation

Weights	f	Weight	f
50 – 53	23	65 – 68	66
53 – 56	24	68 – 71	49
56 – 59	39	71 – 74	38
59 – 62	46	74 – 77	21
62 – 65	54	77 – 80	12

SOLUTION

Weight	F	x	$u = \frac{x-a}{h}$	fu	$\lvert x - \bar{x} \rvert$	$f \lvert x - \bar{x} \rvert$
50 – 53	23	51.5	–5	–115	13.08	300.84
53 – 56	24	54.5	–4	–96	10.08	241.92
56 – 59	39	57.5	–3	–117	7.08	276.12
59 – 62	46	60.5	–2	–92	4.08	187.68
62 – 65	54	63.5	–1	–54	1.08	58.32
65 – 68	66	66.5	0	0	1.92	126.72
68 – 71	49	69.5	1	49	4.92	241.08
71 – 74	38	72.5	2	76	7.92	300.96
74 – 77	21	75.5	3	63	10.92	229.32
77 – 80	12	78.5	4	48	13.92	167.04
—	**372**	—	—	–238	—	**2130.00**

$$\bar{x} = a + \frac{\Sigma fu}{\Sigma f} \times h$$

Let $u = \frac{x - a}{n}$, $a = 66.5$ and $h = 3$

$$\bar{x} = 66.5 + \frac{(-238)}{372} \times 3$$

$$= 66.5 - 1.92 = 64.58$$

$$\text{M.D.} = \frac{\Sigma f\,|x - \bar{x}|}{\Sigma f}$$

$$= \frac{2130}{372} = 5.726$$

$$\text{Coefficient of M.D.} = \frac{\text{M.D.}}{\text{Mean}}$$

$$= \frac{5.726}{64.58} = 0.089$$

4.5 THE VARIANCE AND STANDARD DEVIATION

The Variance

The variance is defined as the mean of the squares of deviation of all the observations from their mean.

The concept of variance was introduced in 1918 by R.A. Fisher. Due to its importance variance is commonly used measure of dispersion. The symbolic definition for variance is

$$S^2 = \frac{\Sigma(x - \bar{x})^2}{n}$$

Or $$S^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

In case of frequency distribution variance may be defined as

$$S^2 = \frac{\Sigma f(x - \bar{x})^2}{\Sigma f}$$

Or $$S^2 = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2$$

Standard Deviation

The positive square root of the variance is called standard deviation. Symbolically

$$S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

Or $$S = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

In case of frequency distribution

$$S = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{\Sigma f}}$$

Or $$S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

The standard deviation is expressed in the same units as the observations themselves.

EXAMPLE 4.7

Calculate the variance and standard deviation from the following marks obtained by 9 students.

45, 32, 37, 46, 39, 36, 41, 48, 36

SOLUTION

X	X^2
45	2025
32	1024
37	1369
46	2116
39	1521
36	1296
41	1681
48	2304
36	1296
360	**14632**

$$S^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

$$= \frac{14632}{9} - \left(\frac{360}{9}\right)^2$$

$= 1625.78 - 1600$

$= 25.78 \text{ (Mark)}^2$

and $S = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$

$= 5.08$ Marks

EXAMPLE 4.8

Determine the variance and standard deviation from the following data.

Groups	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54
f	1	4	8	11	15	9	2

SOLUTION

Classes	F	x	fx	fx^2
20 – 24	1	22	22	484
25 – 29	4	27	108	2916
30 – 34	8	32	256	8192
35 – 39	11	37	407	15059
40 – 44	15	42	630	26460
45 – 49	9	47	423	19881
50 – 54	2	52	104	5408
Total	**50**	**—**	**1950**	**78400**

$S^2 = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2$

$= \frac{78400}{50} - \left(\frac{1950}{50}\right)^2$

$= 1568 - 1521$

$S^2 = 47$

Standard Deviation

$S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$

$= \sqrt{47} = 6.856$

composed by ARFAN ALI MATHEMATICIAN 03148100197

Change of Origin and Scale (Short cut method for calculating variance and standard deviation)

Let $u = \frac{x - a}{h}$

Then $S^2 = h^2\left[\frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2\right]$

Where h = class interval

Standard Deviation = $S = h\sqrt{\frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2}$

EXAMPLE 4.9

Calculate variance and standard deviation by using short cut method to the following data.

Weight	28 – 31	32 – 35	36 – 39	40 – 43	44 – 47
f	1	14	56	172	245
Weight	48 – 51	52 – 55	56 – 59	60 – 63	64 – 67
f	263	156	67	23	3

SOLUTION

The necessary calculations are given below.

Weight (Kg)	f	x	$u = \frac{x - 49.5}{4}$	fu	fu^2
28 – 31	1	29.5	–5	–5	25
32 – 35	14	33.5	–4	–56	224
36 – 39	56	37.5	–3	–168	504
40 – 43	172	41.5	–2	–344	688
44 – 47	245	45.5	–1	–245	245
48 – 51	263	49.5	0	0	0
52 – 55	156	53.5	1	156	156
56 – 59	67	57.6	2	134	268
60 – 63	23	61.5	3	69	207
64 – 67	3	65.5	4	12	48
Σ	**1000**	—	—	**–447**	**2365**

$$S = h \times \sqrt{\frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2}$$

$$= 4 \times \sqrt{\frac{2365}{1000} - \left(\frac{-447}{1000}\right)^2}$$

$$= 4\sqrt{2.365 - 0.1998}$$

$$= 4\sqrt{2.1652}$$

$$= 4 \times 1.47 = 5.88 \text{ kg}$$

Variance $(S)^2 = (5.88)^2$

$= 34.5744 \text{ (kg)}^2$

4.5.1 Co-efficient of Variation

The most important of all the relative measures of dispersion is co-efficient of variation. It is used to compare the variations in two or more set of data or distributions that are measured in different units. For example, one may be measured in hours and the other in rupees. The group which has lower value of co-efficient of variation is comparatively more consistent. The coefficient of variation is defined as

$$C.V. = \frac{S}{\bar{x}} \times 100$$

EXAMPLE 4.10

Goals scored by two teams A and B in a hockey season were as follows.

No. of goals x	Number of matches	
	A	B
0	24	17
1	9	9
2	8	6
3	5	5
4	4	3

By calculating the co-efficient of variation in each case, find which team may be considered more consistent.

SOLUTION

No. of Goals	Team A			Team B		
x	f	fx	fx^2	f	fx	fx^2
0	24	0	0	17	0	0
1	9	9	9	9	9	9
2	8	16	32	6	12	24
3	5	15	45	5	15	45
4	4	16	64	3	12	48
Total	50	56	150	40	48	126

Team A

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f} = \frac{56}{50} = 1.12$$

$$S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

$$= \sqrt{\frac{150}{50} - \left(\frac{56}{50}\right)^2}$$

$$= \sqrt{3 - 1.2544}$$

$$= \sqrt{1.7456} = 1.321$$

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100$$

$$= \frac{1.321}{1.12} \times 100 = 117.95\ \%$$

Team B

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f} = \frac{48}{40} = 1.20$$

$$S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

$$= \sqrt{\frac{126}{40} - \left(\frac{48}{40}\right)^2}$$

$$= \sqrt{1.71} = 1.308$$

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100$$

$$= \frac{1.308}{1.20} \times 100 = 109.0\%$$

We see that the coefficient of variation for the team B is smaller than that for the team A.

Hence team B is more consistent than team A.

4.5.2 Properties of Variance and Standard Deviation

The variance and standard deviation have the following properties.

(i) The variance and standard deviation of a constant is zero. If "a" is a constant, then

$$var(a) = 0$$

$$S.D.(a) = 0$$

(ii) The variance and standard deviation are independent of origin.

$$var(x + a) = var(x)$$

$$var(x - a) = var(x)$$

and $S.D.(x \pm a) = S.D.(x)$

(iii) When all the values are multiplied with a constant, the variance of the values is multiplied by square of the constant, and their standard deviation is multiplied by the constant.

i.e. $var(ax) = a^2\, var(x)$

$$var\left(\frac{x}{a}\right) = \frac{1}{a^2} var(x)$$

and $S.D.(ax) = a.S.D.(x)$

$$S.D.\left(\frac{x}{a}\right) = \frac{1}{a} \cdot S.D.(x)$$

(iv) The variance of the sum or difference of two independent variables is equal to the sum of their respective variances.

If x and y are two independent variables, then

$$var(x + y) = var(x) + var(y)$$

$$var(x - y) = var(x) + var(y)$$

and $S.D.(x + y) = S.D.(x) + S.D.(y)$

$$S.D.(x - y) = S.D.(x) + S.D.(y)$$

(v) If sets of data constants of $n_1, n_2, \ldots\ldots, n_k$ values having corresponding means $\bar{x}_1, \bar{x}_2, \ldots\ldots, \bar{x}_k$ and variances $S_1^2, S_2^2, \ldots\ldots S_k^2$, then the combined variance of a set data is given by

$$S_c^2 = \frac{\Sigma n_i [S_i^2 + (\bar{x}_i - \bar{x}_c)^2]}{\Sigma n_i}$$

Where $\bar{x}_c = \dfrac{\Sigma n_i \bar{x}_i}{\Sigma n_i}$

EXAMPLE 4.11

A distribution consists of 3 components with frequency 100, 120 and 150 having means 5.5, 15.8 and 10.5 and standard deviations 2.4, 4.2 and 3.7 respectively. Find the coefficient of variation for the combined distribution.

SOLUTION

$n_1 = 100,\ n_2 = 120,\ n_3 = 150$

$\bar{x}_1 = 5.5,\ \bar{x}_2 = 15.8,\ \bar{x}_3 = 10.5$

$S_1 = 2.4,\ S_2 = 4.2,\ S_3 = 3.7$

$$\text{Combined mean} = \bar{X}_C = \frac{\Sigma n_i \bar{x}_i}{\Sigma n_i}$$

$$\bar{X}_C = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + x_3\bar{x}_3}{n_1 + n_2 + n_3}$$

$$= \frac{100(5.5) + 120(15.8) + 150(10.5)}{100 + 120 + 150}$$

$$= \frac{550 + 1896 + 1575}{370}$$

$$= \frac{4021}{370} = 10.87$$

$$S_c^2 = \frac{\Sigma n_i[S_i^2 + (\bar{x}_1 - \bar{x}_c)^2]}{\Sigma n_i}$$

$$S_c^2 = \frac{n_1[S_1^2 + (\bar{x}_1 - \bar{x}_c)^2] + n_2[S_2^2 + (\bar{x}_2 - \bar{x}_c)^2] + n_3[S_3^2 + (\bar{x}_3 - \bar{x}_c)^2]}{n_1 + n_2 + n_3}$$

$$= \frac{100[5.76 + (5.5 - 10.87)^2] + 120[17.64 + (15.8 - 10.87)^2] + 150[13.69 + (10.5 - 10.87)^2]}{100 + 120 + 150}$$

$$= \frac{100\,[5.76 + 28.84] + 120[17.64 + 24.30] + 150[13.69 + 0.137]}{370}$$

$$= \frac{100(34.6) + 120(41.94) + 150(13.827)}{370}$$

$$= \frac{3460 + 5032.8 + 2074.05}{370}$$

$$= \frac{10566.85}{370} = 28.56$$

$$S_c = \sqrt{28.56} = 5.34$$

$$\text{Combined C.V.} = \frac{S_c}{\bar{x}_c} \times 100$$

$$= \frac{5.34}{10.87} \times 100 = 49.13\%$$

4.6 MOMENTS

The measure of location alongwith measure of variability are useful to describe a data set but fail to tell anything about the shape of the distribution. For this purpose we need to define certain other measures. Important measures about the shape of the distribution is called moments. These measures are discussed as skewness and Kurtosis.

Moments are defined as the arithmetic means of the powers of the deviations of the observations from any value. These are denoted by u_1, u_2, u_3, u_4 for population data and m_1, m_2, m_3 and m_4 etc for sample data.

4.6.1 Moments about Mean

The rth sample moments about mean is denoted by m_r is given by

$$m_r = \frac{\Sigma(x-\bar{x})^r}{n}, \text{ (for ungroup data)}$$

$$m_r = \frac{\Sigma f(x-\bar{x})^r}{\Sigma f}, \text{ (for grouped data)}$$

where r = 1, 2, 3, 4,.....

m_1 is always zero while m_2 is called variance

$$m_1 = \frac{\Sigma(x-\bar{x})^1}{n} = 0$$

$$m_2 = \frac{\Sigma(x-\bar{x})^2}{n} = \text{variance}$$

$$m_3 = \frac{\Sigma(x-\bar{x})^4}{n} \text{ and } m_4 = \frac{\Sigma(x-\bar{x})^4}{n}$$

EXAMPLE 4.12

Calculate first four moments about mean for the following data given below.

4, 7, 5, 9, 8, 3, 6

SOLUTION

x	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
4	−2	4	−8	16
7	1	1	1	1
5	−1	1	−1	1
9	3	9	27	81
8	2	4	8	16
3	−3	9	−27	81
6	0	0	0	0
42	0	28	0	196

First of all we calculate mean as

$$\bar{x} = \frac{\Sigma x}{n} = \frac{42}{7} = 6$$

$$m_r = \frac{\Sigma(x-\bar{x})^r}{n}\ ;\ r = 1, 2, 3, 4$$

$$m_1 = \frac{\Sigma(x-\bar{x})}{n} = \frac{0}{7} = 0$$

$$m_2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{28}{7} = 4$$

$$m_3 = \frac{\Sigma(x-\bar{x})^3}{n} = \frac{0}{7} = 0$$

$$m_4 = \frac{\Sigma(x-\bar{x})^4}{n} = \frac{196}{7} = 28$$

EXAMPLE 4.13

Calculate first four moments about mean from the following data

SOLUTION

Classes	f	x	fx	$x-\bar{x}$	$f(x-\bar{x})$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
20 - 24	1	22	22	–17	–17	289	–4913	83521
25 - 29	4	27	108	–12	–48	576	–6912	82944
30 - 34	8	32	256	–7	–56	392	–2744	19208
35 - 39	11	37	407	–2	–22	44	–88	176
40 - 44	15	42	630	–3	45	135	405	1215
45 - 49	9	47	423	8	72	576	4608	36864
50 - 54	2	52	104	13	26	338	4394	57122
Total	50	–	1950	–	0	2350	–5250	281050

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{1950}{50} = 39$$

$$m_r = \frac{\Sigma f(x-\bar{x})^r}{\Sigma f}\ ;\ r = 1, 2, 3, \text{ and } 4$$

$$m_1 = \frac{\Sigma f(x-\bar{x})^1}{\Sigma f} = 0$$

$$m_2 = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f} = \frac{2350}{50} = 47$$

$$m_3 = \frac{\Sigma f(x-\bar{x})^3}{\Sigma f} = \frac{-5250}{50} = -105$$

$$m_4 = \frac{\Sigma f(x-\bar{x})^4}{\Sigma f} = \frac{281050}{50} = 5621$$

4.6.2 Moments about an Arbitrary Value

The rth sample moment about any arbitrary origin a is denoted by m'_r is defined as

$$m'_r = \frac{\Sigma (x-a)^r}{n} = \frac{\Sigma D^r}{n}$$

where $D = x - a$

$r = 1, 2, 3, 4, \ldots\ldots$

So

$$m'_1 = \frac{\Sigma (x-a)^1}{n} = \frac{\Sigma D}{n}$$

$$m'_2 = \frac{\Sigma (x-a)^2}{n} = \frac{\Sigma D^2}{n}$$

$$m'_3 = \frac{\Sigma (x-a)^3}{n} = \frac{\Sigma D^3}{n}$$

$$m'_r = \frac{\Sigma (x-a)^4}{n} = \frac{\Sigma D^4}{n}$$

The rth sample moments for grouped data about any arbitrary value a is denoted by m'_r, is defined as

$$m'_r = \frac{\Sigma f (x-a)^r}{\Sigma f} = \frac{\Sigma f D^r}{\Sigma f}, \; r = 1, 2, ,3 ,4$$

$$m'_1 = \frac{\Sigma f (x-a)^1}{\Sigma f} = \frac{\Sigma f D}{\Sigma f}$$

$$m'_2 = \frac{\Sigma f (x-a)^2}{\Sigma f} = \frac{\Sigma f D^2}{\Sigma f}$$

Similar

$$m'_3 = \frac{\Sigma f (x-a)^3}{\Sigma f} = \frac{\Sigma f D^3}{\Sigma f} \text{ and}$$

$$m'_4 = \frac{\Sigma f (x-a)^4}{\Sigma f} = \frac{\Sigma f D^4}{\Sigma f}$$

4.6.3 Relationship of Moments between mean and about Arbitrary Origin

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2$$

$$m_3 = m'_3 - 3m'_2 m'_1 + 2(m'_1)^3$$

$$m_4 = m'_4 - 4m'_3 m'_1 + 6m'_2 (m'_1)^2 - 3(m'_1)^4$$

EXAMPLE 4.14

(a) Calculate first four moments about origin and convert them into moment about mean.

(b) Also calculate fist four moments about x = 5 and convert them into moments about mean from the following data: 4, 7, 5, 9, 8, 3, 6.

SOLUTION

(a) Moments about origin

x	x^2	x^3	x^4
4	16	64	256
7	49	343	2401
5	25	125	625
9	81	729	6561
8	64	512	4096
3	9	27	81
6	36	216	1296
42	280	2016	15,316

$$m'_r = \frac{\sum x^r}{n}; r = 1, 2, 3, 4$$

$$m'_1 = \frac{\sum x}{n} = \frac{42}{7} = 6$$

$$m'_2 = \frac{\sum x^2}{n} = \frac{280}{7} = 40$$

$$m'_3 = \frac{\sum x^3}{n} = \frac{2016}{7} = 288$$

$$m'_4 = \frac{\sum x^4}{n} = \frac{15316}{7} = 2188$$

Converting into moments about mean, we have

$m_1 = 0$

$m_2 = m'_2 - (m'_1)^2 = 40 - (6)^2 = 40 - 36 = 4$

$m_3 = m'_3 - 3m'_2\, m'_1 + 2(m'_1)^3$

$= 288 - 3(40\ (6) + 2(6)^3$

$= 288 - 720 + 432$

$m_3 = 0$ and

$m_4 = m'_4 - 4m'_3\, m'_1 + 6m'_2\, (m'_1)^2 - 3(m'_1)^4$

$= 2188 - 4(288)\ (6) + 6(40)\ (6)^2 - 3(6)^4$

$= 2188 - 6912 + 8640 - 3888$

$m_4 = 28$

(b) Moments about x = 5

x	x – 5	$(x-5)^2$	$(x-5)^3$	$(x-5)^4$
4	–1	1	–1	1
7	2	4	8	16
5	0	0	0	0
9	4	16	64	256
8	3	9	27	81
3	–2	4	–8	16
6	1	1	1	1
42	7	35	91	371

$$m'_r = \frac{\sum (x-a)^r}{n} = \frac{\sum (x-5)^r}{n} \; ; r = 1, 2, 3, 4$$

$$m'_1 = \frac{\sum (x-5)}{n} = \frac{7}{7} = 1$$

$$m'_2 = \frac{\sum (x-5)^2}{n} = \frac{35}{7} = 5$$

$$m'_3 = \frac{\sum (x-5)^3}{n} = \frac{91}{7} = 13$$

$$m'_4 = \frac{\sum (x-5)^4}{n} = \frac{371}{7} = 53$$

Converting into moments about mean, we have

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1) = 5 - (1)^2 = 4$$

$$m_3 = m'_3 - 3m'_2\, m'_1 + 2(m'_1)^3$$

$$13 - 15 + 2 = 0$$

$$m_4 = m'_4 - 4m'_3\, m'_1 + 6m'_2\,(m'_1)^2 - 3(m'_1)^4$$

$$= 53 - 4\,(13)\,(1) + 6\,(5)\,(1)^2 - 3(1)4$$

$$= 53 - 52 + 30 - 3$$

$$m_4 = 83 - 55 = 28$$

4.7 SKEWNESS

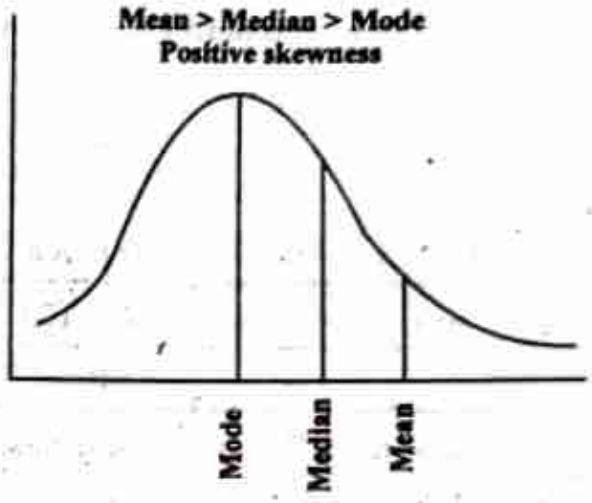
A distribution in which the values equidistant from the mean have equal frequencies is called symmetrical. A distribution is called skewed if it is not symmetrical. A skewed distribution has a curve with a longer tail on any direction. If the right tail is longer than the left tail, the distribution is said to have positive

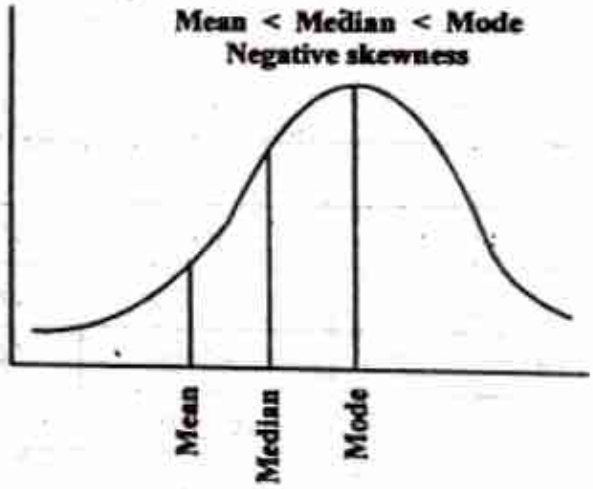
skewness. If the left tail is longer than the right tail, it is said to have negatively skewed. In positive skewness, the mean is greater than the median and the median is greater than the mode.

i.e. Mean > Median > Mode

And in the negatively skewed distribution

Mode > Median > Mean





4.7.1 Measure of Skewness

Karl Pearson introduced a coefficient of skewness denoted by S_k and defined by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

We know that mode is sometimes ill-defined and it is difficult to locate by simple methods. In such cases the median is calculated and the formula becomes,

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

This coefficient usually varies between –3(negative skewness) and +3(positive skewness)

Another measure of skewness was suggested by Bowley. The Bowley's coefficient of skewness is

$$S_k = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

It values lies between –1 and +1.

Skewness may also be defined by moment ratio given as

$$b_1 = \frac{m_3}{\sqrt{m_2^3}}$$

EXAMPLE 4.15

The weight of 38 students at a college are given below.

Weight	118 – 126	127 – 135	136 – 144	145 – 153	154 – 162	163–171
f	3	5	9	12	5	4

(i) Calculate Karl Pearson coefficient of skewness.

(ii) Calculate Bowley's coefficient of skewness.

SOLUTION

(i) Karl Pearson coefficient

Weight	F	x	fx	fx^2	C.B	C.f
118 – 126	3	122	366	44652	117.5 – 126.5	3
127 – 135	5	131	655	85805	126.5 – 135.5	8
136 – 144	9	14	1260	176400	135.5 – 144.5	17
145 – 153	12	149	1788	266412	144.5 – 153.5	29
154 – 162	5	158	790	124820	153.5 – 162.5	34
163 – 171	4	167	668	111556	162.5 – 171.5	38
Σ	38	—	5527	809645		

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{5527}{38} = 145.45$$

$$S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

$$= \sqrt{\frac{809645}{38} - \left(\frac{5527}{38}\right)^2}$$

$$= \sqrt{21306.45 - 21154.94}$$

$$= \sqrt{151.51} = 12.31$$

$$\text{Mode} = l + \frac{(f_m - f_1)}{(fm - f_1) + (f_m - f_2)} \times h$$

$$= 144.5 + \frac{(12 - 9)}{(12 - 9) + (12 - 9)} \times 9$$

$$= 144.5 + \frac{3}{3 + 7} \times 9$$

$$= 144.5 + \frac{27}{10}$$

$$= 144.5 + 2.7 = 147.2$$

$$S_k = \frac{\text{Mean} - \text{Mode}}{S}$$

$$= \frac{145.45 - 147.2}{12.31}$$

$$= \frac{-1.75}{12.31} = -0.142$$

(ii) Bowlay's Co-efficient

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right)$$

$$Q_1 = \frac{n}{4} \text{ th value}$$

$$= \frac{38}{4} \text{ th value} = 9.5 \text{ th value}$$

9.5th value lie in the group 135.5 – 144.5

$$Q_1 = 135.5 + \frac{9}{9}\left(\frac{38}{4} - 8\right)$$

$$= 135.5 + (1.5)$$

$$= 137.0$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - C\right)$$

$$Q_3 = \frac{3n}{4} \text{th value}$$

$$= \frac{3 \times 38}{4} \text{ th value}$$

$$= 28.5\text{th value}$$

Therefore

$$Q_3 = 144.5 + \frac{9}{12}\left(\frac{3 \times 38}{4} - 17\right)$$

$$= 144.5 + \frac{9}{12}(11.5)$$

$$= 144.5 + 8.625$$

$$= 153.13$$

$$\text{Median} = 1 + \frac{h}{f}\left(\frac{n}{2} - C\right)$$

$$\text{Median} = \frac{n}{2}\text{th value}$$

$$= \frac{38}{2}\text{ th value} = 19\text{th value}$$

19th value lie in the group 144.5 – 153.5, therefore

$$\text{Median} = 144.5 + \frac{9}{12}\left(\frac{38}{2} - 17\right)$$

$$= 146.0$$

Bowley's coefficient of skewness

$$= \frac{Q_3 + Q_1 - 2\text{ Median}}{Q_3 - Q_1}$$

$$= \frac{153.13 + 137.0 - 2(146.0)}{153.13 - 137.0}$$

$$= \frac{290.13 - 292}{16.13}$$

$$= -0.116$$

Empirical Relation between Measure of Dispersion

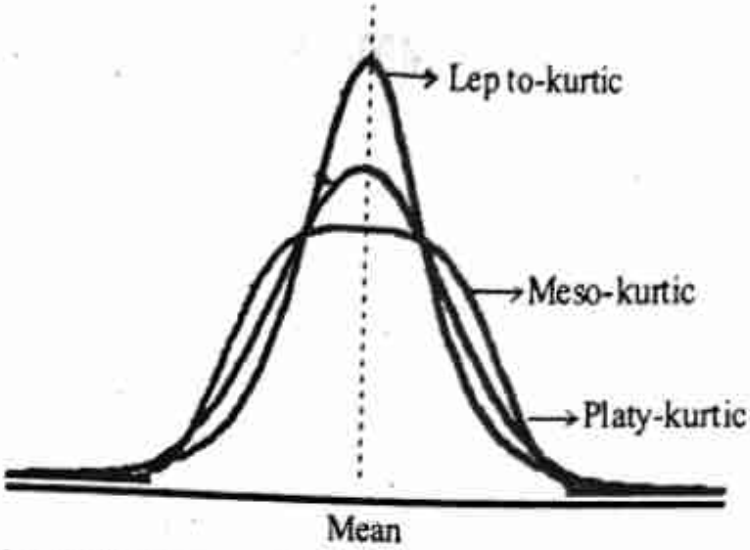
$$\text{Mean deviation} = \frac{4}{5}\text{ (Standard Deviation)}$$

$$\text{Semi-Inter Quartile Range} = \frac{2}{3}\text{ (Standard Deviation)}$$

$$\text{Semi-Inter Quartile Range} = \frac{5}{6}\text{ (Mean Deviation)}$$

4.8 KURTOSIS

Kurtosis is the degree of peakness or flatness of unimodal frequency curve. A distribution having relatively high peak is caled Leptokurtic. If on the otherhand a distribution is flat-topped is called playtkurtic. A normal distribution which is neither very peaked nor very flat-toped is called mesokurtic. Following figure shows the shapes of the distribution.



Kurtosis is usually measured by the moment ratio defined as

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2}$$

If $\beta_2 = 3$, then the distribution is mesokurtic (normal). If $\beta_2 < 3$, the distribution is platykurtic and if $\beta_2 > 3$, the distribution is lepokurtic.

The corresponding measure of kurtosis for the sample data is $b_2 = \frac{m_4}{(m_2)^2}$.

EXAMPLE 4.16

Calculate Skewness and Kurtosis for data of example 4.13. Also interpret the result.

SOLUTION

From example 4.13. We know that $m_2 = 47$, $m_3 = -105$ and $m_4 = 5621$

$$\text{Skewness} = b_1 = \frac{m_3}{\sqrt{m_2^3}}$$

$$b_1 = \frac{-105}{\sqrt{(47)^3}} = \frac{-105}{322.22}$$

$$b_1 = -0.326$$

The distribution is negatively skewed.

$$\text{Kurtosis} = b_2 = \frac{m_2}{(m_2)^2}$$

$$= \frac{5621}{(47)^2} = \frac{5621}{2209}$$

$$b_2 = 2.545$$

as $b_2 < 3$, so the distribution is platykurtic.

SUMMARY

The formula and methods of computing the different types of dispersion and skewness are summarized below:

APPLICATION	FORMULA
Range	R = Maximum value – Min value i.e. $R = X_m - X_o$
Quartile Deviation	$Q.D. = \frac{Q_3 - Q_1}{2}$ Where $Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C\right)$ $Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - C\right)$
Coefficient of Q.D.	$\frac{Q_3 - Q_1}{Q_3 + Q_1}$
Mean deviation	
Ungroupd data	(i) $M.D = \frac{\Sigma\lvert X - Mean\rvert}{n}$ (ii) $M.D = \frac{\Sigma\lvert X - Median\rvert}{n}$ (iii) $M.D = \frac{\Sigma\lvert X - Mode\rvert}{n}$
Grouped data	(i) $M.D. = \frac{\Sigma f\lvert X - Mean\rvert}{\Sigma f}$ (ii) $M.D. = \frac{\Sigma f\lvert X - Median\rvert}{\Sigma f}$ (iii) $M.D. = \frac{\Sigma f\lvert X - Mode\rvert}{\Sigma f}$
Standard deviation	
Ungrouped data	$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}}$ Or $S = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}$
Short cut method	$S = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$ Where $d = X - a$

Grouped Data	
Direct method	$S = \sqrt{\frac{\Sigma f (X - \bar{X})^2}{\Sigma f}}$ Or $S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$
Step deviation method	$S = \sqrt{\frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2} \times h$
Coefficient of variation	$C.V = \frac{S}{\bar{X}} \times 100$
Variance	Square of standard deviation
Measure of Skewness	
Person's formula	(i) $Sk = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$ (ii) $= \frac{3(\text{mean} - \text{median})}{\text{S.D.}}$
Bowley's formula	$SK = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$
Moment ratio (skewness)	$b_1 = \frac{m_3}{\sqrt{m_2^3}}$
Kurtosis	$b_2 = \frac{m_4}{\sqrt{m_2^2}}$

EXERCISES

4.1 What is meant by dispersion? Discuss the various measures of dispersion.

4.2 What is range and how is calculated?

4.3 Find the range and coefficient of range from the following data.

(i) 15, 12, 18, 16, 11, 19, 25, 13, 17, 21

(ii) 105, 103, 110, 108, 106, 115, 110, 109, 102

4.4 (a) Find range from the following frequency distribution.

Classes	70 – 74	75 – 79	80 – 84	85 – 89	90 – 94
f	2	5	12	8	7

(b) Find range and its coefficient from the following data.

Daily wages	120	150	170	200	250	300
f	15	20	27	23	15	10

4.5 (a) Define semi-interquartile range or quartile deviation.

(b) Find the quartile deviation from the following data. Also calculate the coefficient of quartile deviation.

(i) 15, 12, 18, 16, 11, 19, 25, 13, 17, 21

(ii) 105, 103, 110, 108, 106, 115, 110, 109, 102

4.6 Calculate the quartile deviation from the following data.

Mark	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79
f	8	87	190	86	20

4.7 Find the quartile deviation and its coefficient from the following data.

Classes	41 – 50	51 – 60	61 – 70	71 – 80	81 – 90	91– 100
f	30	36	43	104	73	14

4.8 (a) Define mean deviation and its coefficient.

(b) Calculate mean deviation from the following data.

9, 2, 6, 12, 8, 13, 23, 16, 6, 5

(c) Calculate mean deviation from median from the following data.

9, 2, 6, 12, 18, 13, 23, 16, 6, 5

4.9 Calculate mean deviation from mean and its coefficient from the data given below.

Income	35 – 39	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69
F	13	15	28	17	12	10	5

4.10 Calculate mean deviation from median.

Classes	86 – 90	91 – 95	96 – 100	101– 105	106– 110	111– 115
f	6	4	10	6	3	1

4.11 Calculate mean deviation from mode from the data given below. Also calculate its coefficient.

Groups	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13	13 – 15	15 – 17
F	6	53	83	56	21	26	4	4

4.12 **(a)** Define variance and standard deviation. Describe their properties.

(b) Calculate variance and standard deviation from the following data.

(i) 1, 2, 3, 4, 5

(ii) 3, 5, 7, 13, 15, 17, 23, 27

(iii) 10, 8, 7, 9, 5, 12, 8, 6, 8, 2

4.13 Calculate the variance and standard deviation of the data given below. Also calculate the coefficient of variation.

x	63	64	65	66	67	68	69	70	71
f	4	6	10	20	30	13	12	3	2

(b) Calculate variance and standard deviation from the following frequency distribution.

x, fx, fx^2

Classes	1-3	4-6	7-9	10-12	13-15
f	1	2	3	2	1

4.14 Calculate the variance and standard deviation for weight distribution of 120 students at the Govt. College, Lahore.

Weight	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74
f	1	4	17	28	25	18
Weight	75 – 79	80 – 84	85 – 89	90 – 94	95 – 99	
f	13	6	5	2	1	

4.15 Determine the variance, standard deviation and coefficient of variation of the following data.

Classes	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54
f	1	4	8	11	15	9	2

4.16 The following frequency distribution shows the marks of students.

Marks	30 – 39	40 – 40	50 – 59	60 – 69	70 – 79	80 – 89
f	10	14	26	20	18	12

Using the transformation $u = \frac{x - 54.5}{10}$.

Calculate the variance and coefficient of variation.

4.17 It is often stated that in frequency distribution there exists the approximate relation $\frac{\text{Mean deviation}}{\text{Standard deviation}} = 0.8$. Test this statement in the following distribution.

Weight (grams)	65 – 84	85 – 104	105 – 124	125 – 144
f	9	10	17	10
Weight (grams)	145 – 164	165 – 184	185 – 204	
f	5	4	5	

4.18 The breaking strength of 20 test pieces of a certain alloy is given as under.

95, 103, 97, 130, 76, 73, 78, 95, 89, 68, 82, 79, 69, 67, 83, 108, 94, 87, 93, 117

Calculate the average breaking strength of the alloy and standard deviation. Calculate the percentage of observations lying within the limit : mean ± S, mean ± 2S, mean ± 3S, where S stands for standard deviation.

4.19 What do you understand by variance? The wages of 1000 employees ranges from Rs. 4.50 to Rs. 19.50. They are grouped in 15 classes with a common class interval of Re. 1, and the class frequencies from the lowest class to the highest class are 6, 17, 35, 48, 65, 90, 131, 173, 155, 117, 75, 52, 21, 9 and 6. Find the mean wage and its standard deviation.

4.20 (a) Find the arithmetic mean and the standard deviation of the observations, 40, 40, 50, 60, 70, 70, 80, 80, 90. Also find the mean and standard deviation after increasing the observation by

(i) 10 units

(ii) 10 percent

(b) What will be the standard deviation and the variance in each of the following cases.

(i) $2x$ (ii) $x + 2$

(iii) $2x + 4$

If var(x) = 25.

4.21 Compute the mean wages and co-efficient of variation for the employees working in two factors are given below.

Wages	Factories	
	Factory A	Factory B
30 – 35	12	4
35 – 40	18	10
40 – 45	29	31
45 – 50	32	67
50 – 55	16	35
55 – 60	8	15

4.22 Compare the variability of expenditure in two towns as given below.

Expenditure (Rupees)	No. of Families	
	Town A	Town B
21 – 30	3	2
31 – 40	61	14
41 – 50	132	20
51 – 60	153	27
61 – 70	140	28
71 – 80	51	7
81 – 90	2	2

4.23 The following are the marks of two students in different subjects.

A	75	60	72	75	56	82	50
B	90	35	95	88	19	60	76

Who is better student? Who is more consistent student?

standard deviation is 8 on the same test. Find the mean and standard deviation of the combined group of 100 children.

(b) Coefficient of variation of two series are 75% and 90% and their standard deviations are 15 and 18 respectively. Find their mean.

4.25 Given the following data. Calculate the variance and standard deviation by step deviation method.

Classes	20 – 24	25 – 29	30 – 34	35 – 39
f	1	4	8	11
Classes	40 – 44	45 – 49	50 – 59	
f	15	9	2	

4.26 (a) What is meant by skewness? Distinguish between positive and negative skewness.

(b) What can you say of skewness in each of the following distributions?

(i) Mean = 19, Mode = 52

(ii) Q_1 = 136, Median = 160, Q_3 = 184

(iii) Mean = 78, Median = 61

4.27 The heights of 100 college students measured to nearest inch are given below.

Height	60-62	63-65	66-68	69-71	72-74
f	5	18	42	27	8

Calculate Coefficient of skewness by Karl Person method.

4.28 Find the coefficient of skewness by Bowley's formula from the following frequency distribution and interpret the result.

Age Group	0 – 10	10 – 20	20 – 30	30 – 40
f	18	16	15	13
Age Group	40 – 50	50 – 60	60 – 70	70 – 80
f	10	5	2	1

4.29 Calculate

(i) Bowley's coefficient of skewness

(ii) Karl Pearson's coefficient of skewness from the following data.

Classes	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54
Cumulative frequency	22	50	268	495	730	946	1000

4.30 The daily income of employees range from Rs. 0 to Rs. 18. They are grouped in intervals of Rs. 2 and class frequencies from lowest to the highest class are 5, 39, 69, 41, 29, 22, 16, 7, 5. Find coefficient of skewness.

4.31

Classes	2.5 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4	4.5 – 4.9
f	18	22	36	48	42
Classes	5.0 – 5.4	5.5 – 5.9	6.0 – 6.4	6.5 – 6.9	7.0 – 7.4
f	26	14	10	03	01

Required

(i) Standard Deviation

(ii) Variance

(iii) Pearsonian measure of skewness

4.32 (a) What do you meant by moments?

(b) In a certain distribution, the first and second moments about value 2 are –1 and 16 respectively. find the mean and variance of the distribution.

(c) Find b_1 and b_2, given that first four raw moments are:
$m'_1 = -2$, $m'_2 = 15$, $m'_3 = -25$, and $m'_4 = 80$

4.33 (a) Given the following information. $\Sigma f = 290$ $\Sigma fx = 2610$, $\Sigma fx^2 = 23780$, $\Sigma fx^3 = 219530$ and $\Sigma fx^4 = 2056100$. Calculate first four moments about mean.

(b) First three moments of distribution about x = 4 are 1,4 and 10 respectively. Is the distribution symmetrical or skewed.

(c) Lower and upper quartiles of a distribution are 142.36 and 167.73 respectively. While median is 153.50. Find the coefficient of Skewness.

4.34 (a) Calculate first four moments about mean. Also calculate Skewness and Kurtosis for the following data: 1, 2, 3, 4, 5.

(b) Calculate first four moments about mean.

Class	f
1 – 3	1
4 – 6	2
7 – 9	3
10 – 12	2
13 – 15	1

4.35 Calculate the first four moments about the mean for the following data. Also calculate Skewness and Kurtosis.

x	1	2	3	4	5	6	7	8	9
f	1	6	13	25	30	22	9	5	2

4.36 SELECT THE CORRECT ANSWER:

(i) Which one of the following is not a measure of dispersion?

(a) Range (b) Standard deviation

(c) Second quartile (d) Variance

(ii) The standard deviation is

(a) Square of the variances (b) Half of the Variance

(c) Two times standard deviation (d) Square root of the Variance

(iii) The Coefficient of variation is measure in

(a) The same unit as the means and Standard deviation

(b) Percent

(c) Square unit (d) None of these

(iv) If the tail of a frequency distribution is to the right, the Coefficient of Skewness is

(a) Zero (b) Positive

(c) Negative (d) None

(v) Symmetrical distribution will always have skewness equal to.

(a) Zero (b) Negative

(c) Positive (d) None

(vi) If a distribution has zero variance, then which of the following is true?

(a) All of the observations are negative

(b) All the observations are positive

(c) All the observations are equal

(d) None of these

(vii) For a normal distribution, the measure of kurtosis equal to

(a) Zero (b) 3

(c) Negative (d) Positive

(viii) If the original units are measure in kg, the variance is

(a) Also measured in Kg (b) Measure in Kg squared

(c) Measure in half Kg (d) None of the these

(ix) If standard deviation of frequency distribution is 10, means is 40 and mode is also 40, then coefficient of Skewness is:

(a) Zero (b) Positive

(c) Negative (d) None of these

ANSWERS 4.36

(i)	(c)	**(ii)**	(d)	**(iii)**	(b)	**(iv)**	(b)	**(v)**	(a)
(vi)	(c)	**(vii)**	(b)	**(viii)**	(b)	**(ix)**	(a)		

SHORT QUESTION AND ANSWERS

1. Define dispersion.

Ans: By dispersion we mean the degree to which numerical data tends to spread about an average value.

2. What is the difference between absolute dispersion and relative dispersion?

Ans: Absolute Dispersion:

An absolute dispersion is one that measure the dispersion in term of same units. e.g. if the units of data are rupees, kilograms etc, the units of the measure of dispersion will also be rupees, kilograms etc.

Relative Dispersion:

It is expressed in the form of ratio and coefficients. It is independent of the units of measurements.

3. What are the main measures of dispersion? Types

Ans: The main measures of dispersion are the followings:

(i) The Range

(ii) The semi inter quartile range or quartile deviation.

(iii) The mean deviation

(iv) The variance and standard deviation

4. Define Range

Ans: Range is defined as the difference between the largest and the smallest observation in a set of data

$$\text{Range} = R = X_m - X_o$$

$$\text{Coefficient of Range} = \frac{X_m - X_o}{X_m + X_o}$$

5. What do you meant by Quartile deviation?

Ans: Quartile deviation is defined as half of the difference between the third and the first quartile, *i.e.*

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

6. Define mean deviation.

Ans: Mean deviation is defined as the mean of the absolute deviations measured either from mean, median or mode. By absolute deviation we mean that all the deviations are positive.

$$M.D. = \frac{\Sigma|X - Mean|}{n}$$

or

$$M.D. = \frac{\Sigma|X - Median|}{n}$$

or

$$M.D. = \frac{\Sigma|X - Mode|}{n}$$

7. Define variance.

Ans: The variance is defined as the mean of the squares of deviations of all the observations from their mean.

or

$$S^2 = \frac{\Sigma(X - \overline{X})^2}{n}$$

8. Define standard deviation.

Ans: The positive square root of the variance is called standard deviation symbolically.

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}}$$

9. What do you know about coefficient of variation?

Ans: It is used to compare the variations in two or more than two sets of data. The group which has lower value of coefficient of variation is comparatively more consistent. The coefficient of variation is defined as:

$$C.V = \frac{S}{\overline{X}} \times 100$$

10. What are the properties of variance.

Ans: The variance has the following properties:

(i) The variance of a constant is zero, var (a) = 0.

(ii) var (x ± a) = var (X)

(iii) var (ax) = a^2 var (x)

(iv) var (x ± y) = var (x) + var (y)

11. Define symmetrical distribution.

Ans: A distribution in which the values equidistant from the mean have equal frequency is called symmetrical distribution.

12. What do you meant by skewness?

Ans: A distribution is called skewed if it is not symmetrical.

13. Define positive skewness.

Ans: If the right tail is longer than the left tail, the distribution is said to have positive skewness. In case of positively skewness:

Mean > Median > Mode

14. Define negative skewness.

Ans: If the left tail is longer than right tail, it is called negative skewness. In case of negative skewness.

Mode > Median > Mean

15: What is the measure of Skewness?

Ans: (i) Karl Person Coefficient of Skewness:

$$S_k = \frac{\text{Mean} - \text{Mode}}{S}$$

(ii) Bowley coefficient of Skewness:

$$S_k = \frac{Q_3 + Q_1 - 2\text{ median}}{Q_3 - Q_1}$$

16. Define moments.

Ans: Moments are defined s the arithmetic means of the powers of the deviations of observations from any value.

17. What do you mean by kurtosis?

Ans: It is the degree of peakness or flatness of unimodel frequency curve.

18. Define moment ratio.

Ans: The ratio between the moment about mean are called moment ratio. i.e.

$$b_1 = \frac{m_3^2}{m_2^3} \text{ and } b_2 = \frac{m_4}{m_2^2}$$

CHAPTER # 5

SIMPLE REGRESSION AND CORRELATION

5.1 REGRESSION

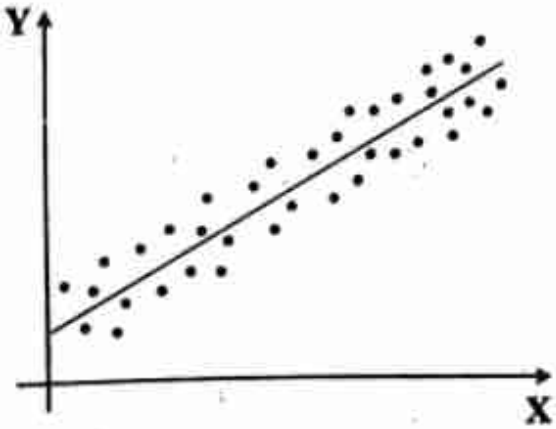
The term regression was introduced by the English biometrician, Sir Francis Galton in 1877. He studied the heights of parents and their children and observed an interesting relationship between them. He found that, though all tall parents have tall children and short parents have short children, the average height of children tends to step back or to regress towards the average height of all men. This tendency towards the average height of all men was called a regression by Galton.

Today, the world regression is used in quite different sense. It investigates the dependence of one variable on other variables, called independent variables and provides an equation to be used for estimating the average value of the dependent variable from the known values of independent variable. For example, if we want to estimate the heights of children on the basis of their ages, the height would be the dependent variable and the ages would be the independent variable.

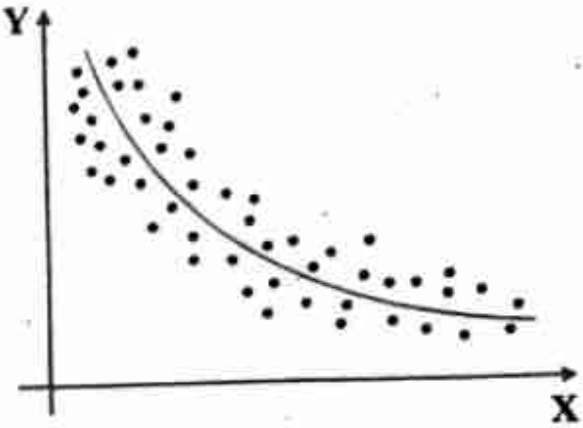
5.2 SCATTER DIAGRAM

A scatter diagram can be used to find whether or not a relationship between two variables exists. The pair of independent - dependent observations as a point on Graph paper using the x-axis for the independent variable and y-axis for the dependent variable. Such a diagram is called a scatter diagram. By looking to the scatter of the various points on the graph we can form an idea as whether variables are related or not. The greater is the scatter of the plotted points, the lesser is the relationship between the variables. If a relationship between the variables exists, then the points in the scatter diagram will show a tendency to cluster around a straight line or some curve. Such a line or curve around which the points cluster, is called the regression line.

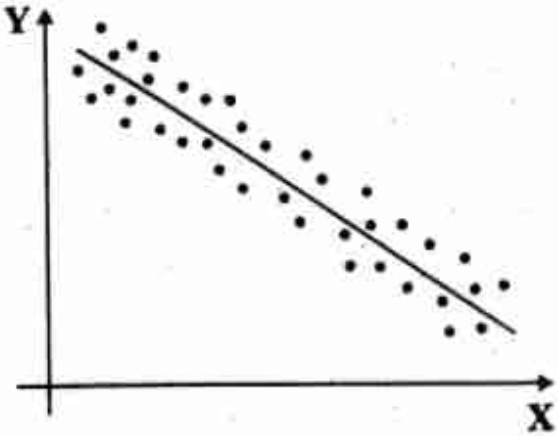
The following scatter diagrams show different types of relation between the variables.



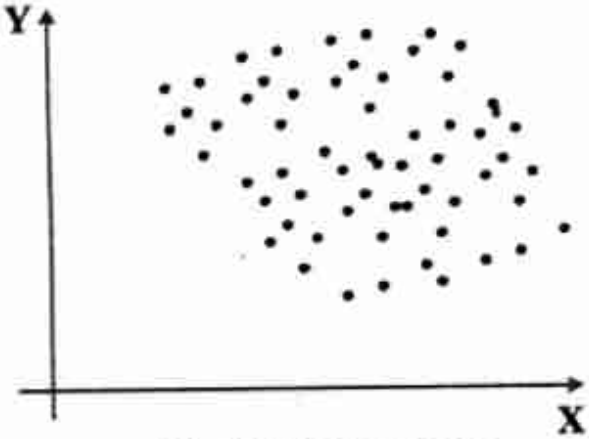
(a) Positive Linear Relationship



(c) Curvilinear Relationship



(b) Negative Linear Relationship



(d) No Relationship

5.3 SIMPLE LINER REGRESSION MODEL

We assume that the linear relationship between the dependent variable Y_i and the value X_i of the regressor X is

$$Y_i = \alpha + \beta X_i + \varepsilon i,$$

where the X_i's are fixed or predetermined values,

the Y_i's are observations randomly drawn from a population,

the ε_i's are error components or random deviations,

α and β are population parameters, α is the intercept and the slope β is called *regression coefficient*.

Furthermore, we assume that

(i) $E(\varepsilon_i) = 0$, *i.e.* the expected value of error term is zero, it implies that the expected value of Y is related to X in the population by a straight line;

(ii) $\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ for all i, *i.e.* the variance of error term is constant. It means that the distribution of error has the same variance for all values of X. *(Homoscedasticity assumption)*;

(iii) $E(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, *i.e.* error terms are independent of each other (*assumption of no serial or auto correlation between* ε's);

(iv) $E(X, \varepsilon i) = 0$, *i.e.* X and ε are also independent of each other;

(v) ε_i's are normally distributed with a mean of zero and a constant variance σ^2. This implies that Y values are also normally distributed.

But in practice, we have a sample from a population, therefore we desire to estimate the population regression line from the sample data. The estimated regression line may be written as

$$y = a + bx$$

Where y is the dependent variable and x is the independent variable, while "b" is the regression coefficient and a is an intercept.

Types of Regression Lines

(i) Regression line y on x

$$y = a + b_{yx}$$

Where $b_{yx} = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$

Or $b_{yx} = \dfrac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$

and $a = \bar{y} - b\bar{x}$

Regression equation interms of correlation is

$$(y - \bar{y}) = r\frac{S_y}{S_x}(x - \bar{x})$$

Where r = correlation coefficient

S_y = standard deviation of y

S_x = standard deviation of x

(ii) Regression Equation x on y

$$x = a + b_{xy}y$$

Where $b_{xy} = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$

Or $b_{xy} = \dfrac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$

and $a = \bar{x} - b\bar{y}$

Regression equation x on y interms of correlation coefficient is

$$(x - \bar{x}) = r\frac{S_x}{S_y}(y - \bar{y}) \qquad \therefore\ b_{xy} = r\frac{S_x}{S_y}$$

EXAMPLE 5.1

Calculation regression line y on x from the data given below. Also predict y when x = 5.

X	4	3	1	2	6	7	2	3
y	39	38	16	18	41	45	25	38

SOLUTION

x	y	xy	x^2
4	39	156	16
3	38	114	9
1	16	16	1
2	18	36	4
6	41	246	36
7	45	315	49
2	25	50	4
3	38	114	9
28	**260**	**1047**	**128**

We know that regression line y on x is

$$y = a + bx \quad \text{..... (i)}$$

$$b = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{8(1047) - 28(260)}{8(128) - (28)^2}$$

$$= \frac{8376 - 7280}{1024 - 784} = \frac{1096}{240} = 4.57$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{260}{8} = 32.5$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{8} = 3.5$$

$$a = \bar{y} - b\bar{x}$$

$$= 32.5 - 4.57(3.5)$$

$$= 32.5 - 15.995 = 16.505$$

Putting the values of a and b in (i), we have

$$\hat{y} = 16.505 + 4.57x$$

Predict y, when x = 5

$$\hat{y} = 16.505 + 4.57\,(5)$$
$$= 16.505 + 22.85$$
$$= 39.36$$

EXAMPLE 5.2

The relationship between money spent on research and development and chemical firms of annual profit are given below.

Year	Million spent on R & D X	Annual profit (millions) Y
1995	5	31
1994	11	40
1993	4	30
1992	5	34
1991	3	25
1990	2	20

Find regression equation y on x

SOLUTION

The estimated regression line y on x is

$$\hat{y} = a + bx$$

where $$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$a = \bar{y} - b\bar{x}$$

The necessary calculations are given below.

x	y	xy	x^2
5	31	155	25
11	40	440	121
4	30	120	16
5	34	170	25
3	25	75	9
2	20	40	4
30	**180**	**1000**	**200**

Substituting the values we get

$$b = \frac{6(1000) - 30(180)}{6(200) - (30)^2}$$

$$b = \frac{6000 - 5400}{1200 - 900}$$

$$b = \frac{600}{300} = 2$$

$$\bar{x} = 30/6 = 5, \bar{y} = 180/6 = 30$$

$$a = 30 - 2(5) = 20$$

The estimated regression line is

$$\hat{y} = 20 + 2x$$

5.4 PROPERTIES OF REGRESSION LINE

(i) Least square regression line always passes through mean $(\bar{x}, \bar{y})$ of the data

(ii) The sum of the deviations of the values y from the regression line is always equal to zero.

i.e. $\Sigma(y - \hat{y}) = 0$

(iii) The sum of the square deviations of the observed values from the least square regression line is minimum.

i.e. $\Sigma(y - \hat{y})^2 = \text{minimum}$

(iv) Regression line is the line of best fit, because a and b are unbiased estimates of parameters α and β.

5.5 STANDARD DEVIATION OF REGRESSION OR STANDARD ERROR OF ESTIMATE

The degree of scatter of the observed values about the regression line is called standard error of estimate of y on x. For sample data the standard error of estimate is denoted by S_{yx} and defined as

$$S_{yx} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n-2}}$$

Or $$= \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$

5.6 COEFFICIENT OF DETERMINATION

The coefficient of determination measures the proportion of variability in the values of dependent variable (y) explained by its linear relationship with the independent variable (x). It is defined as the ratio of explained variation to the total variation. Sample coefficient of determination is denoted by r^2 and defined by

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}, \text{ where}$$

$$\Sigma(y - \bar{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$$

and

$$\Sigma(y - \bar{y})^2 = \Sigma y^2 (\Sigma y)^2/n$$

The alternative formula is

$$r^2 = \frac{a\Sigma y + b\Sigma xy - (\Sigma y)^2/n}{\Sigma y^2 - (\Sigma y)^2/n}$$

EXAMPLE 5.3

Compute the standard error of estimate, coefficient of determination and coefficient of correlation for the data in Example, 5.1

SOLUTION

For the data in example 5.1, we found from the regression calculation, that

$$\Sigma y = 260, \Sigma y^2 = 9320 \; n = 8$$

$$a = 16.505, b = 4.57, \Sigma xy = 1047$$

We know that

$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$

$$= \sqrt{\frac{9320 - 16.505(260) - 4.57(1047)}{8 - 2}}$$

$$= \sqrt{\frac{9320 - 4291.3 - 4784.79}{6}}$$

$$= \sqrt{40.65} \qquad \Rightarrow S_{y.x} = 6.376$$

Which is the standard error of estimate. The coefficient of determination is

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

$$= \frac{a\sum y + b\sum xy - (\sum y)^2/n}{\sum y^2 - (\sum y)^2/n}$$

$$r^2 = \frac{16.505\,(260) + 4.57\,(1407) - (260)^2/8}{9320 - (260)^2/8}$$

$$= \frac{4291.3 + 4784.79 - 8450}{9320 - 8450} = \frac{626.09}{870} = 0.719$$

The coefficient of correlation is

$$r = \sqrt{0.719} = 0.848$$

5.7 CORRELATION

Correlation measures the strength or closeness of relationship between two variables. The purpose of correlation is to determine whether or not two variables are related. In other words the degree of interdependence between two variables is called correlation. For example, heights and weights of persons ages of husbands and their wives at the time of their marriages, demand and supply of a commodity etc.

5.8 POSITIVE AND NEGATIVE CORRELATION

Positive Correlation

If both the variables tend to increase or decrease together, the correlation is said to be direct or positive e.g. length of an iron bar will increase as the temperature increases.

Negative Correlation

If one variable tends to increase as the other variable decreases the correlation is said to be negative or inverse, e.g. the volume of gas will decrease as pressure increases. Coldness increases as the temperature decreases.

5.9 PERFECT CORRELATION

The correlation coefficient r is a pure number (*i.e.* independent of the units) and it assumes values that can range from +1 for perfect positive linear relationship −1 for perfect negative linear relationship. In case of no correlation, the correlation is said to be zero correlation.

5.10 PEARSON PRODUCT MOMENT CORRELATION CO-EFFICIENT

A numerical measure of strength in the linear relationship between any two variables is known as correlation co-efficient. It is denoted by r is defined as;

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \,.\, \Sigma(y - \bar{y})^2}}$$

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{nS_x S_y}$$

The correlation coefficient r always lies between −1 & +1.

If $r = +1$ (Perfect positive correlation)

If $r = -1$ (Perfect negative correlation)

If $r = 0$ (The two variables are independent)

EXAMPLE 5.4

Calculate correlation coefficient between x and y from the following data.

x	85	50	93	35	62	65	87	58	50	45
y	78	38	95	25	75	80	95	63	66	40

SOLUTION

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x-\bar{x})^2$	$(y - \bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
85	78	22	12.5	484	156.25	275.0
50	38	−13	−27.5	169	756.25	357.5
93	95	30	29.5	900	870.25	885.0
35	25	−28	−40.5	784	1640.25	1134.0
62	75	−1	9.5	1	90.25	9.50
65	80	2	14.5	4	210.25	29.0
87	95	24	29.5	576	810.25	708.0
58	63	−5	−2.5	25	6.25	12.5
50	66	−13	0.5	169	0.25	6.5
45	40	−18	−25.5	324	650.25	459.0
630	**655**	**0**	**0**	**34369**	**5250.50**	**3844.0**

Method 1

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 . \Sigma(y - \bar{y})^2}}$$

$$= \frac{3844.00}{\sqrt{(34369)(5250.50)}}$$

$$= 0.91$$

Therefore

$$\bar{x} = \frac{\Sigma x}{n} = \frac{630}{10} = 63.0$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{655}{10} = 65.5$$

Method 2

x	y	xy	x^2	y^2
85	78	6630	7225	6084
50	38	1900	2500	1444
93	95	8835	8649	9025
35	25	875	1225	625
62	75	4650	3844	5625
65	80	5200	4225	6400
87	95	8265	7569	9025
58	63	3654	3364	3969
50	66	3300	2500	4356
45	40	1800	2025	1600
630	**655**	**45109**	**43126**	**48153**

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][x\Sigma y^2 - (\Sigma y)^2]}}$$

$$= \frac{10(45109) - 630(655)}{\sqrt{[10(43125) - (630)^2][10(48153) - (655)^2]}}$$

$$= \frac{451090 - 412650}{\sqrt{(431260 - 396900)(481530 - 429025)}}$$

$$= \frac{38440}{\sqrt{(34360)(52505)}}$$

$$= \frac{38440}{42474.3664} = 0.91$$

5.11 PROPERTIES OF COEFFICIENT OF CORRELATION

The sample correlation coefficient has the following properties.

(i) The correlation coefficient is symmetrical with respect to variables x and y.

i.e. $r_{xy} = r_{yx}$

(ii) The correlation coefficient lies between −1 and +1.

i.e. $-1 \le r \le 1$

(iii) The correlation coefficient is independent of origin and scale. *i.e.* $r_{xy} = r_{yx}$

(iv) The correlation coefficient is the geometric mean of two regression coefficient.

i.e. $r = \sqrt{(b_{yx})(b_{xy})}$

(v) The correlation coefficient is a pure number. *i.e.* It has no unit.

(vi) For two independent random variables correlation coefficient is zero.

5.12 SHORT CUT METHOD FOR THE CALCULATION OF CORRELATION COEFFICIENT

Let $u = x - a,\ v = y - b$

Then $r_{uv} = \dfrac{n\Sigma uv - \Sigma u\,\Sigma v}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2]\,[n\Sigma v^2 - (\Sigma v)^2]}}$

EXAMPLE 5.5

Calculate the co-efficient of correlation between the variables x and y given below by using short cut method.

x	78	89	97	69	59	79	68	61
y	125	137	156	112	107	136	123	108

SOLUTION

Let $u = x - 69$ and $v = y - 112$

x	y	u=x−69	v = y − 112	u^2	v^2	uv
78	125	9	13	81	169	117
89	137	20	25	400	625	500
97	156	28	44	784	1936	1232
69	112	0	0	0	0	0
59	107	−10	−5	100	25	50
79	136	10	24	100	576	540
68	123	−1	11	1	121	−11
61	108	−8	−4	64	16	32
600	**1004**	**48**	**108**	**1530**	**3468**	**2160**

$$r_{uv} = \frac{n\Sigma uv - \Sigma u\,\Sigma v}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2]\,[n\Sigma v^2 - (\Sigma v)^2]}}$$

$$= \frac{8(2160) - 48(188)}{\sqrt{[8(1530) - (48)^2]\,[8(3468) - (188)^2]}}$$

$$= \frac{17280 - 5184}{\sqrt{(12240 - 2304)\,(27744 - 11664)}}$$

$$= \frac{12096}{\sqrt{(9936)\,(16080)}}$$

$$= \frac{12096}{12640.05} = 0.96$$

5.13 RANK CORRELATION

Sometimes the actual measurements of objects are not available. They are then arranged in order according to some characteristic of interest. Such an ordered arrangements is called a ranking and the order given to an object is called its rank. The correlation between two sets of rankings is called rank correlation.

The spearman rank correlation coefficient is given by

$$r_s = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)}, \text{ where } d = x\text{-}y$$

EXAMPLE 5.6

Two Judges in a contest, who were asked to rank 8 candidates A, B, C, D, E, F, G, H is order of their preference. Find the coefficient of rank correlation.

	A	B	C	D	E	F	G	H
First Judge	5	2	8	1	4	5	3	7
Second Judge	4	5	7	3	2	8	1	6

We calculate the coefficient of rank correlation as follows.

x_i	y_i	$d = x_i - y_i$	d^2
5	4	1	1
2	5	–3	9
8	7	1	1
1	3	–2	4
4	2	2	4
6	8	–2	4
3	1	2	4
7	6	1	1
–	–	–	28

Then $$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(28)}{8(64-1)} = 0.67$$

This indicates that Judges agreed well in their choices.

5.14 RANK CORRELATION FOR TIED RANKS

The Spearman's formula applies only when no ties are present. When there are ties in ranks the ranks are adjusted by assigning the means of the ranks which they jointly occupy. For example if fifth, sixth and seventh largest values of variables are the same, we assign each the rank $\frac{5+6+7}{3} = 6$. In tied ranks, one of the following two methods is to be used:

First, for each tie, add a quantity $\frac{1}{12}(t^3 - t)$ to $\sum d^2$ before substituting the values in the Spearman's formula of rank correlation.

Second, use the product moment coefficient of correlation to find the correlation between the two sets of adjusted ranks.

EXAMPLE 5.7

Compute the coefficient of rank correlation for the followings ranks:

X	8	3	6.5	3	6.5	9	3	1	5
Y	8	9	6.5	2.5	4	5	6.5	1	2.5

SOLUTION

We observe that both the sets of ranks have 2,2 ties. In the first set, 3 is repeated 3 times and 6.5 two times. Similarly in the second set 2.5 and 6.5 repeated two times. We add the following $\frac{1}{12}(3^3-3)+\frac{1}{12}(2^3-2)+\frac{1}{12}(2^3-2)+\frac{1}{12}(2^3-2)=2+0.5+0.5+0.5=3.5$ quantities to $\sum d^2$.

x	y	d = x-y	d^2
8	8	0	0
3	9	–6	36
6.5	6.5	0	0
3	2.5	0.5	0.25
6.5	4	2.5	6.25
9	5	4	16
3	6.5	–3.5	12.25
1	1	0	0
5	2.5	2.5	6.25
Total	–	**0**	**77**

hence $r_s = 1-\frac{6(\sum d^2+3.5)}{n(n^2-1)}$

$$= 1-\frac{6(77+3.5)}{9(81-1)}$$

$$= 1-\frac{483}{720} = 0.3291$$

Alternative Method

Let us denote the ranks given by the first member by x and second member by y. The necessary calculations are given below.

x	y	x^2	y^2	xy
8	8	64	64	64
3	9	9	81	27
6.5	6.5	42.25	42.25	42.25
3	2.5	9	6.25	7.5
6.5	4	42.25	16	26
9	5	81	25	45
3	6.5	9	42.25	19.5
1	1	1	1	1
5	2.5	25	6.25	12.5
45	**45**	**282.5**	**284**	**244.75**

Hence the coefficient of rank correlation is

$$r = \frac{n\sum xy - \sum xy \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{9(244.75) - (45)(45)}{\sqrt{[9(282.5) - (45)^2]\,[9(284) - (45)^2]}}$$

$$r = 0.3391$$

Both the methods gives almost the same answer.

SUMMARY

The formula and methods of computing the correlation and regression are given below:

APPLICATION	FORMULA
Range of Correlation	$-1 \le r \le 1$
CORRELATION	(i) $r = \dfrac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2\,\Sigma(Y-\overline{Y})^2}}$
	(ii) $r = \dfrac{n\,\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{[n\,\Sigma X^2 - (\Sigma X)^2]\,[n\,\Sigma Y^2 - (\Sigma Y)^2}}$
	(iii) $r = \dfrac{S_{XY}}{S_x\,S_Y}$
	Where $S_{XY} = \dfrac{\Sigma XY}{n} - (\overline{X}\,\overline{Y})$
	$S_X = \sqrt{\dfrac{\Sigma X^2}{n} - \left(\dfrac{\Sigma X}{n}\right)^2}$
	$S_Y = \sqrt{\dfrac{\Sigma Y^2}{n} - \left(\dfrac{\Sigma Y}{n}\right)^2}$
Regression line Y on X	$y = a + bx$
	$a = \overline{y} - b\,\overline{X}$
	Where
	$b_{YX} = \dfrac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(X-\overline{X})^2}$
	Or $b_{YX} = \dfrac{n\,\Sigma XY - \Sigma X\,\Sigma Y}{n\,\Sigma X^2 - (\Sigma X)^2}$
	Or $b_{YX} = r\dfrac{S_y}{S_x}$

Regression line X on Y	$X = a + by$ Where $b_{XY} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2}$ $b_{XY} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma y^2 - (\Sigma y)^2}$ Or $b_{XY} = r\frac{S_x}{S_y}$ $a = \bar{X} - b\bar{y}$
Relation between regression & correlation	$r = \sqrt{(b_{YX})(b_{XY})}$
Standard error of estimate	$S_{yx} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$ Where $\Sigma(y - \hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$
Coefficient of determination	$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$ Or $1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - y)^2}$ Or $\frac{a\Sigma y + b\Sigma xy - (\Sigma y)^2/n}{\Sigma y^2 - (\Sigma y)^2/n}$
Rank correlation	$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$

EXERCISES

5.1 (a) What do you meant by Regression?

(b) Write a short note on scatter diagram.

5.2 The following sample of 8 grade point averages and marks in matriculation was observed for students from a College.

Score	480	490	510	510	530	550	610	640
G.P.A.	2.7	2.9	3.3	2.9	3.1	3.0	3.2	3.7

Find the least square line. Estimate the mean GPA of students scoring 600 marks.

5.3 The following table shows the marks obtained by students in Math and Physics.

Math (x)	48	40	32	34	30	50	26	50	22	43
Physics (y)	76	56	40	50	34	70	56	68	40	57

(i) Construct scatter diagram for the above data.

(ii) Find the least square regression line of y on x and of x on y.

5.4 Find out coefficient of correlation and regression equations for the following data, x on y and y on x. Calculate the value of y when the value of x is 60.

x	10	15	20	25	30
y	15	17	21	23	26

5.5

x	80	82	86	91	83	85	89	96	93
y	145	140	130	124	133	127	120	110	116

Required

Calculate coefficient of correlation and also the line of regression y on x.

5.6 Fit a regression line of y on x from the following data.

x	5	7	6	12	17	19	20	29
y	22	14	11	9	9	8	6	2

5.7 (a) The correlation coefficient between two variables x and y is $r = 0.63$.

If $S_x = 1.50, S_y = 2.00, \bar{x} = 10, \bar{y} = 20$

Find the regression lines y on x and of x on y.

(b) If the equations of the least squares regression lines are:

$y = 20.8 - 0.219\,x$ (y on x)

$x = 16.2 - 0.785\,y$ (x on y)

Find the product moment coefficient of correlation.

5.8 The following data show the son's height and father's height.

Father's height (x)	59	61	63	65	67	69	71	73	75
Son's height (y)	64	66	67	67	68	69	70	72	72

Estimate the regression line $y = a + bx$ using $u = \frac{x - 67}{2}$ and $v = y - 68$.

Predict the mean height of sons whose fathers are 70 inches in height.

5.9 Fit the line of Regression of marks in Maths on marks in Economics from the following data:

Math	5	7	16	12	17	13	10	9
Eco	2	4	17	19	9	8	6	5

5.10 The following table gives the aptitude test score and productivity indices of 10 workers selected at random estimate.

Aptitude score (x)	60	62	65	70	72	48	53	73	65	82
Productivity index (y)	68	60	62	80	85	40	52	62	60	81

Calculate correlation coefficient between aptitude scores and productivity index.

5.11 A financial analyst has gathered the following data bout the relationship between income and investment in securities in respect of 8 randomly selected families.

Income	8	12	9	24	43	37	19	16
Percent Invested	36	25	33	15	28	19	20	20

Find the correlation coefficient and interpret it.

5.12 On the basis of the figured recorded below for supply and price find:

(a) correlation coefficient

(b) Both regression coefficients

Supply(x)	80	82	86	91	83	85	89	96	93
Price (y)	145	140	130	124	133	127	120	110	116

5.13 Find Regression Coefficients of the followings:

$\Sigma x = 17.6$, $\Sigma y = 32.8$, $\Sigma xy = 94.7$

$\Sigma x^2 = 49.64$, $\Sigma y^2 = 182$, $n = 8$

5.14

x	16	72	73	63	83	80	66	66	74	62
y	40	52	43	49	61	58	44	58	50	45

Required: Calculate Coefficient of correlation and comment on the answer.

5.15 Calculate coefficient of correlation between x and y. Find regression equation of y on x. Estimate value of y when x = 75.

x	79	89	97	69	59	79	68	61
y	125	137	156	112	107	136	123	104

5.16 Find out the regression equation for the following data. Calculate the value of y when the value of x is 22.

x	10	15	20	25	30
y	15	17	21	23	26

5.17 Find out:

(i) Regression equation x on y

(ii) Compute trend values and prove that $\Sigma (X - \hat{X}) = 0$

x	5	6	8	10	12	13	15	16	17
y	16	19	23	28	36	40	42	45	50

5.18 (a) Define correlation, write down the properties of correlation.

(b) Calculate coefficient of correlation.

X	1	2	3	4	5	6
Y	2	3	4	3	6	5

(c) The following data show the marks in economics and marks in statistics obtained by ten students.

Student	1	2	3	4	5	6	7	8	9	10
Eco (x)	78	36	96	25	75	82	90	62	65	39
Stat (y)	84	51	91	60	68	62	86	58	53	47

Compute the co-efficient of correlation.

5.19 (a) Calculate coefficient of correlation from the following data.

x	3	5	6	9	10	12	15	20	22	28
y	10	12	15	18	20	22	27	30	32	34

(b) Find the coefficient of correlation between x and y.

x	30	35	40	45	50	60	70	80	90
y	2	4	5	5	8	15	24	30	32

5.20 Given the following data.

x	20	25	30	35	40	45	50	55	60	65
y	90	85	75	65	95	35	20	15	15	5

(i) Calculate the coefficient of correlation between x and y.

(ii) Calculate correlation coefficient between u and v where

$$u = \frac{x - 45}{5} \text{ and } v = \frac{y - 35}{5}$$

5.21 Find the correlation coefficient between x and y, given.

x	5	12	14	16	18	21	22	23	25
y	11	16	15	20	17	19	25	24	21

5.22 (a) For a sample of 20 pairs of observations, we have

$\bar{x} = 2$, $\bar{y} = 8$, $\Sigma x^2 = 180$, $\Sigma y^2 = 3424$, $\Sigma xy = 604$

Calculate the coefficient of correlation.

(b) For a given set of data, we have

$n = 10$, $\Sigma(x - \bar{x})(y - \bar{y}) = 120$, $S_y = 8$

$\Sigma(x - \bar{x})^2 = 90$

Find coefficient of correlation.

5.23 For a set of 50 pairs of observations, the standard deviations of x and y are 4.5 and 3.5 respectively. If the sum of products of deviations of x and y values from their respective means be 420. Find the coefficient of correlation.

5.24 A computer while calculating the coefficient of correlation between two variables x and y from 25 pairs of observations obtained the following sums.

$\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$

$\Sigma xy = 508$

It was however, later discovered at the time of checking that he had copied down two pairs

X	Y
6	14
8	6

while the correct values were

x	Y
8	12
6	8

Obtain the correct value of the coefficient of correlation.

5.25 Calculate the coefficient of correlation between supply and demand from the following data.

Supply	400	200	700	100	500	300	600
Demand	60	30	70	10	40	20	50

5.26 The following table gives the aptitude test scores and productivity indices of 10 workers selected at random estimate.

Aptitude. Score (x)	60	62	65	70	72	48	53	73	65	82
Productivity. index (y)	68	60	62	80	85	40	52	62	60	81

Calculate correlation coefficient between aptitude test scores and productivity indices.

5.27 **(a)** What is meant by standard error of estimate?

(b) Explain coefficient of determination.

(c) Compute coefficient of determination and standard error of estimate for the following data.

Income (x) (0000)	10	20	30	40	50	60
Expenditure (y) (0000)	7	21	23	34	36	53

5.28 **(a)** What is rank correlation?

(b) Ten students got the following positions on the basis of their performance in two subjects.

Subject	1	2	3	4	5	6	7	8	9	10
Subject-I	1	5	2	6	4	8	3	7	10	9
Subject-II	3	6	2	7	5	8.	1	4	9	10

Calculate Spearman's rank correlation coefficient

5.29 (a) Find rank correlation coefficient for the following data:

a	4.7	2.9	6.4	2.5	4.9
b	8.6	5.4	6.2	4.8	8.3

(b) The following table shows the marks of six candidates in the subjects.

Candidate	A	B	C	D	E	F
Math (x)	38	62	56	42	59	48
Stat (y)	64	89	84	60	73	69

Calculate the coefficient of rank correlation.

5.30 SELECT THE CORRECT ANSWER:

(i) **The word regression was introduced by:**

(a) Karl Pearson (b) Galton

(c) R.A. Fisher (d) Marshall

(ii) **The diagram of a set of bivariate data is:**

(a) Histogram (b) Historigram

(c) Scatter diagram (d) Bar diagram

(iii) **If the value of regression Coefficient is zero, then two variables are**

(a) Dependent (b) Independent

(c) Both (a) and (b) (d) None of these

(iv) **The signs of regression coefficients (b_{yx} and b_{xy}) are always:**

(a) Different (b) Same

(c) Positive (d) Negative

(v) **The sum of deviations of observed values from regression line is:**

(a) Zero (b) Minimum

(c) Maximum (d) None

(vi) **The sum of square deviations of values from regression line is**

(a) Zero (b) Minimum

(c) Maximum (d) None of these

(vii) **If $y = 10.5 + 6.3\,x$, then the value of regression coefficient is:**

(a) 0 (b) 10.5

(c) 6.3 (d) x

(viii) If $y = 20 + 2.5x$, then the value of intercept is:

(a) 0 (b) 20
(c) 25 (d) None of these

(ix) The correlation coefficient lies between

(a) -1 and 0 (b) 0 and 1
(c) -1 and 1 (d) None of these

(x) The Correlation Coefficient is independent of

(a) Origin (b) Scale
(c) Both (a) and (b) (d) None of these

(xi) If one variable increase and the other variable also increase, then Correlation is

(a) Positive (b) Negative
(c) Perfect (d) None of these

(xii) If $u = x$ and $v = -x$, then correlation γ_{uv} is:

(a) 1 (b) -1
(c) 0 (d) 0.5

(xiii) If both the variables are independent, then correlation will be

(a) No (b) Positive
(c) Negative (d) None of these

(xiv) The Correlation Coefficient is the geometric mean of:

(a) Variables (b) Regression Coefficients
(c) Coefficient of determination
(d) None of these

ANSWERS 5.30

(i)	(b)	**(ii)**	(c)	**(iii)**	(b)	**(iv)**	(b)	**(v)**	(a)
(vi)	(b)	**(vii)**	(c)	**(viii)**	(b)	**(ix)**	(c)	**(x)**	(c)
(xi)	(a)	**(xii)**	(b)	**(xiii)**	(a)	**(xiv)**	(b)		

SHORT QUESTION AND ANSWERS

1. Define Regression.

Ans: The dependence of one variable on other is called regression e.g. If we want to estimate the heights of children on the basis of their ages, the height would be the dependent variable and the ages would be the independent variable.

2. What is the standard form of least square regression line?

Ans: $y = a + bx$

where y is dependent variable, x is independent variable, while a and b a constants.

3. Define Scatter diagram.

Ans: The pair of independent-dependent observation as a point on the graph pape using the x-axis for independent variable and the y-axis for the dependen variable. Such a diagram is called scatter diagram.

4. Define Correlation.

Ans: The degree of interdependence between two variables is called correlation e.g. ages of husbands and ages of wives at the time of their marriage, demand and supply of a commodity etc.

5. What do you meant by Positive and Negative Correlation?

Ans: **Positive Correlation:**

If both the variables tend to increase or decrease together, the correlation is said to be direct or positive.

Negative Correlation:

If one variable tends to increase as the other variable decreases the correlation is said to be negative or inverse.

6. Write down the properties of correlation.

Ans: (i) Correlation coefficient is symmetrical with respect to be variables x and y.

i.e. $r_{xy} = r_{yx}$

(ii) The correlation coefficient lies between −1 and +1 *i.e.*

$-1 \le r \le 1$

(iii) The correlation coefficient is independent of origin and scale. *i.e.*

$r_{xy} = r_{uv}$

(iv) The correlation is the geometric mean of two regression coefficients. *i.e.*

$$\sqrt{(b_{yx})(b_{xy})}$$

7. What are the properties of Regression?

Ans: (i) Least square regression line always passes through mean $(\bar{x}, \bar{y})$.

(ii) $\Sigma(y - \hat{y}) = 0$

(iii) $\Sigma(y - \hat{y})^2 = \text{minimum}$

CHAPTER # 6

MULTIPLE REGRESSION AND CORRELATION

6.1 MULTIPLE REGRESSION

The technique of simple regression involves on independent variable and one dependent variable. If the number of independent variables in a regression model is more than one, then the model is called as multiple regression. Infact, many of the real-world applications demand the use of multiple regression model. For example, the yield of wheat depends upon the fertility of land, rainfall, fertilizer applied, etc.

The general multiple regression model is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i, \ (i = 1, 2, \dots, n)$$

where ε_i's are the random errors,

α and β_i's are the unknown population parameters, α is the intercept and β_1, β_2, ..., β_k are the regression coefficients for variables $X_1, X_2, \dots, X_k$ respectively.

We assume that

(i) $E(\varepsilon_i) = 0$ for all i. This implies that for given values of X_i's,

$$E(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}.$$

(ii) $\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ for all i, i.e.

(iii) $E(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, i.e. error terms are independent of each other.

(iv) $E(X, \varepsilon_i) = 0$ for all regressors, i.e. ε and each X variable are independent.

(v) ε_i's are normally distributed with a mean of zero and a constant variance σ^2.

(vi) We assume further in a multiple regression model that there exists no exact linear relationship between any two of the regressors.

The corresponding regression equation estimated from sample data then takes the following form

$$\hat{y}_i = a + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki},$$

6.2 MULTIPLE LINEAR REGRESSION WITH TWO REGRESSORS

For two independent variables x_1 and x_2, the estimated multiple linear regression based on sample data is

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

The error or residual in each case is given as

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + b_1X_{1i} + b_2X_{2i})$$

Using the least-squares criterion, we determine those values of a, b_1 and b_2 which will minimize the sum of squared residuals, Σe_i^2. To minimize Σe_i^2, we find $\frac{\partial \Sigma e_i^2}{\partial a}$, $\frac{\partial \Sigma e_i^2}{\partial b_1}$ and $\frac{\partial \Sigma e_i^2}{\partial b_2}$ and set equal to zero. Thus

$$\frac{\partial \Sigma e_i^2}{\partial a} = -2\,\Sigma[Y_i - (a + b_1X_{1i} + b_2X_{2i})] = 0,$$

$$\frac{\partial \Sigma e_i^2}{\partial b_1} = -2\,\Sigma X_{1i}\,[Y_i - (a + b_1X_{1i} + b_2X_{2i})] = 0,$$

$$\frac{\partial \Sigma e_i^2}{\partial b_2} = -2\,\Sigma X_{2i}\,[Y_i - (a + b_1X_{1i} + b_2X_{2i})] = 0.$$

Simplifying, we get the following three normal equations

$$\Sigma Y = na + b_1\Sigma X_1 + b_2\Sigma X_2,$$

$$\Sigma X_1Y = a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2,$$

$$\Sigma X_2Y = a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_1X_2^2.$$

The values of a, b_1 and b_2 are determined by solving these three normal equations simultaneously and are substituted into

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

to obtain the estimated multiple linear regression equation.

EXAMPLE 6.1

Given the following set of data

(a) Calculate the multiple regression equation

(b) Predict y when $x_1 = 3.0$ and $x_2 = 2.7$

y	25	30	11	22	27	19
x_1	3.5	6.7	1.5	0.3	4.6	2.0
x_2	5.0	4.2	8.5	1.4	3.6	1.3

SOLUTION

The estimated multiple regression equation is

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

The normal equations are

$$\Sigma y = na + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

$$\Sigma x_1y = a\Sigma x_1 + b_1\Sigma_1^2 + b_2 \Sigma x_1x_2$$

$$\sum x_2 y = a\sum x_2 + b_1\sum x_1 x_2 + b_2\sum x_2^2$$

The necessary calculations are given below

y	x_1	x_2	x_1y	x_2y	x_1x_2	x_1^2	x_2^2
25	3.5	5.0	87.5	125.0	17.50	12.25	25.00
30	6.7	4.2	201.0	126.0	28.14	44.89	17.64
11	1.5	8.5	16.5	93.5	12.75	2.25	72.25
22	0.3	1.4	6.6	30.8	0.42	0.09	1.96
27	4.6	3.6	124.2	97.2	16.56	21.16	12.96
19	2.0	1.3	38.2	24.7	2.60	4.00	1.69
134	**18.6**	**24.0**	**473.8**	**497.2**	**77.97**	**84.64**	**131.50**

Substituting the sums in the normal equations.

We get

$$134 = 6a + 18.6\, b_1 + 24.0\, b_2$$
$$473.8 = 18.6a + 84.64\, b_1 + 77.97\, b_2$$
$$497.2 = 24.0a + 77.97\, b_1 + 131.50\, b_2$$

Solving these equations simultaneously, we get

$$a = 20.3916, \quad b_1 = 2.3403, \quad b_2 = -1.3293$$

Hence the desired regression equation is

$$\hat{y} = 20.3916 + 2.3403\, x_1 - 1.3283\, x_2$$

(b) When $x_1 = 3$ and $x_2 = 2.7$

$$\hat{y} = 20.3916 + 2.3403(3.0) - 1.3293(2.7)$$

$$\hat{y} = 23.83$$

6.3 STANDARD ERROR OF ESTIMATE

The standard error of estimate is the standard deviation of multiple regression. It measures the dispersion of y values about the population multiple regression equation. The sample standard error of estimate for two independent variables x_1 and x_2, denoted by $S_{y.12}$ is given by

$$S_{y.12} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 3}}$$

Where

$$\sum(y - \hat{y})^2 = \sum[y - (a + b_1 x_1 + b_2 x_2)]^2$$
$$= \sum y^2 - a\sum y - b_1 \sum x_1 y - b_2 \sum x_2 y$$

6.4 CO-EFFICIENT OF MULTIPLE DETERMINATION

The co-efficient of multiple determination measures the proportion of variability in the values of dependent variable y explained by its linear relation with the independent variables. It is defined by the ratio of the explained variation to the total variation. For two independent variable x_1 and x_2, the co-efficient of multiple determination is denoted by $R^2_{y.12}$ and computed as

$$R^2_{y.12} = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

where $\Sigma(\hat{y} - \bar{y})^2 = a\Sigma y + b_1 \Sigma x_1 y + b_2 \Sigma x_2 y - (\Sigma y)^2/n$

and $\Sigma(y - \bar{y})^2 = \Sigma y^2 - (\Sigma y)^2/n$

The coefficient of multiple determination lies between 0 and 1.

The positive square root of the coefficient of multiple determination is called the co-efficient of multiple correlation ($R_{y.12}$). It measures the degree of association between y and both the regressors x_1 and x_2 combined.

6.5 MULTIPLE REGRESSION IN TERMS OF LINEAR CORRELATION COEFFICIENTS

The multiple regression equation of a variable x_1, on other variable x_2 and x_3 can be expressed in terms of linear correlation coefficients and standard deviations, as

$$\frac{x_1}{S_1} = \left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r^2_{23}}\right)\left(\frac{x_2}{S_2}\right) + \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r^2_{23}}\right)\left(\frac{x_3}{S_3}\right)$$

Similarly, the other two multiple regression equations of x_2 on x_1 and x_3, and of x_3 on x_1 and x_2 are obtained as

$$\frac{x_2}{S_2} = \left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r^2_{23}}\right)\left(\frac{x_1}{S_1}\right) + \left(\frac{r_{23} - r_{12}\, r_{13}}{1 - r^2_{13}}\right)\left(\frac{x_3}{S_3}\right), \text{ and}$$

$$\frac{x_3}{S_3} = \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r^2_{12}}\right)\left(\frac{x_1}{S_1}\right) + \left(\frac{r_{23} - r_{12}\, r_{13}}{1 - r^2_{12}}\right)\left(\frac{x_2}{S_2}\right).$$

To obtained the regression equations in terms of original values, we replace x1 by $X1 - \bar{X}_1$, x_2 by $X_2 - \bar{X}_2$ and x_3 by $X_3 - \bar{X}_3$ respectively.

EXAMPLE 6.2

Given

$\bar{X}_1 = 6.8$, $\bar{X}_2 = 7.0$, $\bar{X}_3 = 74$

$S_1 = 1.0$, $S_2 = 0.80$, $S_3 = 9.0$

$r_{12} = 0.60, \quad r_{13} = 0.70, \quad r_{23} = 0.65$

Find the regression equation of x_1 on x_2 and x_3.

SOLUTION

Since the standard deviations and linear correlation coefficients are given, therefore estimated regression equation of x_1 on x_2 and x_3 is

$$\frac{X_1 - \bar{X}_1}{S_1} = \left(\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{X_2 - \bar{X}_2}{S_2}\right) + \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{X_3 - \bar{X}_3}{S_3}\right)$$

Now

$$\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2} = \frac{0.60 - (0.70)\,(0.65)}{1 - (0.65)^2} = \frac{0.145}{0.5775} = 0.2511$$

$$\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2} = \frac{0.70 - (0.60)\,(0.65)}{1 - (0.65)^2} = \frac{0.31}{0.5775} = 0.5368$$

Putting the values, we get

$$\frac{X_1 - 6.8}{1.0} = 0.2511\left(\frac{X_2 - 7.0}{0.80}\right) + 0.5368\left(\frac{X_3 - 74}{9.0}\right)$$

$$X_1 - 6.8 = 0.3139\,(X_2 - 7.0) + 0.0596\,(X_3 - 74)$$

$$X_1 - 6.8 = 0.3139\,X_2 - 2.1973 + 0.0596\,X_3 - 4.4104$$

$$X_1 - 6.8 = 0.6077 + 0.3139\,X_2 + 0.0596\,X_3$$

$$\hat{X} = 0.1923 + 0.3139\,X_2 + 0.0596\,X_3$$

EXAMPLE 6.3

Compute the standard error of estimate, coefficient of multiple determination and coefficient of multiple correlation for the data in Example 6.1.

SOLUTION

For the data in example 6.1, we found from the regression calculations, that

$\Sigma y = 134,\ \Sigma y^2 = 3220\ n = 6,\ \Sigma x_1 y = 473.8,\ \Sigma x_2 y = 497.2$

$a = 20.3916,\ b_1 = 2.3404,\ b_2 = -1.3283$

Therefore,

$$S_{y.12} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b_1\Sigma x_1 y - b_2\Sigma x_2 y}{n - 3}}$$

$$= \sqrt{\frac{3220 - 20.3916\,(134) - 2.3403\,(473.8) - (-1.3283)\,(497.2)}{6 - 3}}$$

$$= \sqrt{\frac{3220 - 2732.4744 - 1108.8341 + 660.4308}{3}}$$

$$= \sqrt{\frac{39.1223}{3}} = \sqrt{13.0408}$$

$S_{y.12} = 3.6112$

The coefficient of multiple determination is

$$R^2_{y.12} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

$$= \frac{a\sum y + b_1 \sum x_1 y + b_2 \sum x_2 y - (\sum y)^2/n}{\sum y^2 - (\sum y)^2/n}$$

$$= \frac{20.3916\,(134) + 2.3403(473.8) + (-1.3283)(497.2) - (134)^2/6}{3220 - (134)^2/6}$$

$$= \frac{2732.4744 + 1108.8341 - 660.4308 - 2992.6667}{3320 - 2992.6667}$$

$$= \frac{188.211}{227.3333}$$

$R^2_{y.12} = 0.8279$

This means that 83% of the variability in dependent variable is explained by its linear relationship with independent variables.

The coefficient of multiple correlation, $R_{y.12}$ is

$$R_{y.12} = \sqrt{0.8279} = 0.9099$$

6.6 MULTIPLE CORRELATION COEFFICIENT

The coefficient of multiple correlation measures the degree of relationship between a variable and its estimate from the regression equation. The coefficient of multiple correlation between x_1 and the variables x_2 and x_3 combined is 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 other two multiple correlation coefficients are

$$R_{2.13} = \sqrt{\frac{r_{23}^2 + r_{21}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{12}^2}}$$

6.7 PARTIAL CORRELATION

The coefficient of partial correlation measures the strength of relationship between x_1 and x_2, when the influence of x_3 has been removed, denoted by $r_{12.3}$ and is given by

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

In an similar way, by the change of variables, we get

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \text{ and}$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

EXAMPLE 6.4

Calculate multiple correlation coefficient $R_{1.23}$ and partial correlation coefficient $r_{23.1}$ from the following correlation coefficients.

$r_{12} = 0.80, \quad r_{13} = -0.40, \quad r_{23} = -0.56$

SOLUTION

We know that

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.80)^2 + (-0.40)^2 - 2(0.80)(-0.40)(-0.56)}{1 - (-0.56)^2}}$$

$$= \sqrt{\frac{0.4416}{0.6864}}$$

$$= \sqrt{0.6434}$$

$$R_{1.23} = 0.8021$$

$$r_{23.1} = \frac{r_{23} - r_{13}\, r_{12}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

$$= \frac{-0.56 - (0.80)(-0.40)}{\sqrt{[1 - (0.80)^2][1 - (-0.40)^2]}}$$

$$= \frac{-0.24}{0.5499}$$

$$r_{23.1} = -0.4364$$

SUMMARY

The formula and methods of computing, multiple regressions and correlation are given below:

APPLICATION	FORMULA
Multiple regression equation y on x_1 and x_2	$\hat{y} = a + b_1 x_1 + b_2 x_2$ The normal equations are: $\sum y = na + b_1 \sum x_1 + \sum b_2 \sum x_2$ $\sum x_1 y = a\sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$ $\sum x_2 y = a\sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$
Standard error of estimate	$S_{y.12} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-3}}$ Where $(y-\hat{y})^2 = \sum y^2 - a\sum y - b_1 \sum x_1 y - b_2 \sum x_2 y$
Coefficient of multiple determination	$R_{y.12} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$ Where $\sum(\hat{y}-\bar{y})^2 = a\sum y + b_1 \sum x_1 y + b_2 \sum x_2 y - \frac{(\sum y)^2}{n}$ and $\sum(y-\bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$
Multiple correlation coefficient	$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$
Partial correlation coefficient	$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - {}^2_{13})(1 - r_{23}^2)}}$

EXERCISES

6.1 (a) What is multiple regression? Explain the basic differences between simple and multiple regression.

(b) What do you meant by coefficient by determination and multiple correlation?

6.2 The following information has been gathered from a random sample of apartment rents in a city. We are trying to predict rent (in pounds per month) based on the size of apartment (number of rooms) and the distance from main city (in kilometers).

Rent	No. of Rooms	Distance from City
y	x_1	x_2
360	2	1
1000	6	1
450	3	2
525	4	3
350	2	10
300	1	4

(a) Calculate the least square equation that best relates these three variables.

(b) If some one is looking for 2 bed room apartment 2 k.m from city. What rent should be expect to pay?

6.3 Given that data

y	12	10	9	13	20
x_1	2	2	3	4	4
x_2	1	1	0	0	3

(a) Calculate the estimated regression equation, (i.e. $\hat{y} = a + b_1 x_1 + b_2 x_2$) for the above data.

(b) Calculate the standard error of estimate, $S_{y.12}$.

(c) Calculate the coefficient of multiple determination and multiple correlation.

6.4 Given the data

y	12	15	14	18	17	15	20
x_1	12	16	18	20	18	25	30
x_2	30	50	65	75	80	95	105

Calculate the estimated regression equation:

$$\hat{y} = a + b_1 x_1 + b_2 x_2.$$

6.5 Given the following information

$n = 6$, $\Sigma x_1 = 48$, $\Sigma x_2 = 42$, $\Sigma x_3 = 300$, $\Sigma x_1^2 = 474$, $\Sigma x_2^2 = 434$

$\Sigma x_3^2 = 19008$, $\Sigma x_1 x_2 = 236$, $\Sigma x_2 x_3 = 2820$ and $\Sigma x_1 x_3 = 1818$

(a) Find the regression equation of x_3 on x_1 and x_2.

(b) Estimate x_3 when $x_1 = 10$ and $x_2 = 6$

6.6 Given

$\bar{x}_1 = 6.8$	$S_1 = 1.0$	$r_{12} = 0.60$
$\bar{x}_2 = 7.0$	$S_2 = 0.8$	$r_{13} = 0.70$
$\bar{x}_3 = 74$	$S_3 = 9.0$	$r_{23} = 0.65$

(a) Find the least square regression equation of x_3 on x_1 and x_2.

(b) Compute $R_{3.12}$

6.7 The following means, standard deviations are correlations are found for

x_1 = Seed hay crop in cwt per acre,

x_2 = Spring rain fall in inches

x_3 = Accumulated temperature above 42°F

$\bar{x}_1 = 28.02$	$S_1 = 4.42$	$r_{12} = 0.80$
$\bar{x}_2 = 4.91$	$S_2 = 1.10$	$r_{13} = -0.40$
$\bar{x}_3 = 594$	$S_3 = 85$	$r_{23} = -0.56$

Find the partial correlations and the regression equation for hay-crop on spring rain fall and accumulated temperature.

6.8 (a) Distinguish between multiple correlation and partial correlation.

(b) Given

$r_{13} = 0.492$, $r_{13} = 0.927$ and $r_{23} = 0.758$

Calculate all the partial correlation coefficient.

(c) Three variables have in pairs, simple correlation coefficients given by

$r_{12} = 0.8$, $r_{23} = -0.9$, $r_{13} = -0.7$

Find all the partial correlation coefficients and multiple correlation coefficient $R_{1.23}$ of x_1 on x_2 and x_3.

6.9 **(a)** Calculate multiple correlation coefficient $R_{2.13}$ of x_2 on x_1 and x_3.

$b_{12} = 0.75$, $b_{13} = 0.58$, $b_{21} = 0.88$

$b_{23} = 0.53$, $b_{31} = 1.68$ and $b_{32} = 1.30$

(b) Calculate the partial correlation coefficient $r_{23.1}$ and the multiple correlation coefficient $R_{2.13}$ from the values given below

$b_{21} = -0.1$, $b_{12} = -0.4$, $b_{13} = 0.27$

$b_{31} = 0.6$, $b_{23} = 0.67$, $b_{32} = 0.38$

6.10 **(a)** Given that following information

$n = 5$, $\Sigma y = 89$, $\Sigma x_1 = 30$, $\Sigma x_2 = 52$

$\Sigma x_1^2 = 238$, $\Sigma_2^2 = 582$, $\Sigma x_1 x_2 = 357$

$\Sigma x_1 y = 619$, $\Sigma x_2 y = 1007$ $\Sigma y^2 = 1885$

Calculate multiple regression equation y on x_1 andx_2.

(b) Compute the standard error of estimate, coefficient of multiple determination and coefficient of multiple correlation for the data in question 6.10(a).

6.11 SELECT THE CORRECT ANSWER:

(i) The range of multiple correlation coefficient is

(a) -1 to $+1$ (b) 0 to ∞

(c) 0 to 1 (d) None of these

(ii) The range of partial correlation coefficient is

(a) -1 to 1 (b) 0 to 1

(c) 0 to ∞ (d) -1 to 0

(iii) If the multiple correlation coefficient $R_{1.23} = 1$, then it means that

(a) High relationship (b) Weak relationship

(c) Perfect relationship (d) Perfect linear relationship

(iv) Multiple correlation coefficient is

(a) Square of the coefficient of determination

(b) Square root of the coefficient of determination

(c) Half of the coefficient of determination

(d) None of all above

(v) In the regression analysis, the explained variation is given by

(a) $\Sigma(y - \hat{y})^2$ (b) $\Sigma(y - \bar{y})^2$

(c) $\Sigma(\hat{y} - \bar{y})^2$ (d) None of these

(vi) **The slope b_2 in multiple regression equation $\hat{y} = a + b_1x_1 + b_2x_2$ measures**

(a) The change in $\hat{y}$ per unit change in x_1

(b) The change in $\hat{y}$ per unit change in x_2

(c) The change in $\hat{y}$ per unit change in x_1, holding x_2 constant

(d) The change in $\hat{y}$ per unit change in x_2, holding x_1 constant

(vii) **The predicated value of y for $x_1 = 2$ and $x_2 = 3$ by using regression line; $\hat{y} = 20 + 7x_1 + 2x_2$ is**

(a) 20 (b) 30

(c) 40 (d) 50

(viii) **The range of coefficient of determination is**

(a) 0 to 1 (b) 0 to ∞

(c) $-\infty$ to ∞ (d) -1 to 1

ANSWERS 6.11

(i)	(c)	(ii)	(a)	(iii)	(c)	(iv)	(b)	(v)	(c)
(vi)	(d)	(vii)	(c)	(viii)	(a)				

SHORT QUESTION AND ANSWERS

1. **What is multiple regression?**

Ans: A regression which involves two or more independent variables is called multiple regression. e.g student CGPA depends upon intelligence, no. of hours of study etc.

2. **Define standard error of estimate.**

Ans: The standard error of estimate is the standard deviation of regression. It measures the dispersion of y values about the population multiple regression equation.

3. **What do you mean by coefficient of determination?**

Ans: It is defined by the ratio of explained variation to the total variation.

$$R^2_{y.12} = \frac{\Sigma(\hat{y} - \bar{y})^2}{(y - \bar{y})^2}$$

4. **Define multiple correlation.**

Ans: The coefficient of multiple correlation measures the degree of relationship between a variable and its estimate from the regression equation.

5. **Define partial correlation.**

Ans: The coefficient of partial correlation measures the strength of relationship between x_1 and x_2, when the influence of x_3 has been removed, denoted by $r_{12.3}$.

CHAPTER # 7

PROBABILITY THEORY

7.1 INTRODUCTION

The concept of probability plays an important role in all problems of science, business and every day life, which in any way involve an element of uncertainty.

It is mainly concerned with drawing inferences from uncertain experiments or situations. For example, when we make the statements that it is likely to rain today, Pakistan will win the Hockey World Cup, A train will be late or will be in time etc. We face such type of problems in our daily life. All these problems are uncertain in nature. A similar type of uncertainty occurs when we toss a coin, throw a dice or draw a card from well-shuffled pack of cards etc. The uncertainty in all these cases is measured in terms of probability.

Two French mathematicians Blaise Pascal and De Fermat in connection with gambling problems found the theory of probability. Later on Jacob Bernoulli, Abraham De Moivire and Laplace developed it. The modern probability theory was developed during the twenties and thirties of this century.

Probability is widely used in Economics, Psychology, Meteorology, Astronomy, Sociology, Physics and Insurance where risk and uncertainty are involved.

Now a day, probability theory is best understood through the application of set theory, therefore basic concepts, notations, and operations of set theory that are relevant to probability are given below.

7.2 SET

A set is a well-defined collection of distinct objects. The objects of the set are called elements. A set is denoted by capital letters A, B, C etc. Sets are always written between pair of brackets.

e.g. A = {1, 2, 3, 4, 5}

B = {Karachi, Lahore, Peshawar, Quetta}

Finite Set

A set is finite if it contains limited number of elements. *e.g.* The number of students in a college, the books in a library etc.

Infinite Set

A set is called infinite if it contains infinite number of elements *e.g.* the number of stars on the sky, the drops of water in a lake. etc.

Null or Empty Set

A set, which has no element, is called Null set or empty set. It is denoted by ϕ or { }.

e.g. A = {A male student admitted in girl's college}

Singleton Set

A set, which has only one element, is called singleton set.

e.g. A = {2}, B = {Lahore}

Subset

If every element of set A is also an element of set B, then set A is called subset of set B. It is denoted by $A \subset B$.

e.g. A = {1, 2, 3}

B = {1, 2, 3, 4, 5}

Then $A \subset B$

Universal Set

A set, which contains all elements of all the sets under consideration, is called universal set. It is denoted by U or S.

e.g. A = {1, 3, 5}

B = {2, 4}

U = {1, 2, 3, 4, 5}

7.3 VENN DIAGRAM

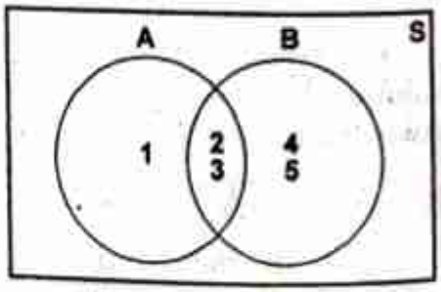
Venn diagram is a pictorial (Diagrammatic) way of representing sets and their relationship. The universal set "U" or "S" is shown by a rectangle. Inside the rectangle, circles show other sets.

e.g.

Let, U = {1, 2, 3, 4, 5}

A = {1, 2, 3}

B = {2, 3, 4, 5}



7.4 OPERATIONS ON SETS

The basic operations are union, intersection, difference and complement.

(i) Union of Sets

The union of two sets A and B is the set of all elements which belongs to either A or B or both A and B. It is denoted by $A \cup B$.

e.g. $A = \{1, 2, 3\}$

$B = \{3, 4, 5\}$

Then $A \cup B = \{1, 2, 3, 4, 5\}$

(ii) Intersection of Sets

The intersection of two sets A and B is the set of those elements which are common in A and B.

It is denoted by $A \cap B$.

e.g. $A = \{1, 2, 3, 4\}$

$B = \{3, 4, 5, 6\}$

$A \cap B = \{3, 4\}$

(iii) Difference of Sets

The difference of two sets A and B is denoted by "A–B" is the set of all elements which belong to A, but not to B.

e.g. $A = \{1, 2, 3, 4, 5\}$

$B = \{2, 5\}$

$A - B = \{1, 3, 4\}$

(iv) Complement of a Set

The complement of a set A denoted by $\bar{A}$ is a particular difference "S – A" *i.e.* the set of all those elements which belongs to S, but not to A.

e.g. $S = \{1, 2, 3, 4, 5, 6\}$

$A = \{1, 2, 3\}$

$$\bar{A} = S - A = \{1, 2, 3, 4, 5, 6\} - \{1, 2, 3\}$$
$$= \{4, 5, 6\}$$

7.5 PRODUCT SET

The product sets A and B denoted by $A \times B$ is a set of all ordered pairs (x, y), where x belongs to A and y belongs to B.

e.g. $A = \{1, 2, 3\}$, $B = \{5, 6\}$

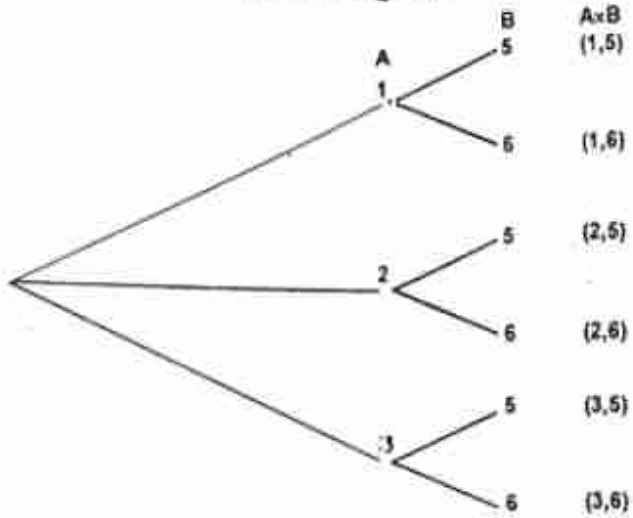
Then $A \times B = \{(1, 5), (1, 6), (2, 5), 2, 6), 3, 5), 3, 6)\}$

Similarly

$B \times A = \{(5, 1), (5, 2), (5, 3), (6, 1), (6, 2), (6, 3)\}$

The product set $A \times B$ may be easily formed by means of "Tree Diagram"

Tree Diagram



Similarly Tree diagram for $B \times A$ my be formed.

Ordered Pair

An ordered pair consists of two elements say x and y, in which one of them (say x) is designated as first element and the other as a second element.

An ordered pair is denoted by (x, y)

7.6 TECHNIQUES OF COUNTING

Factorial

The numbers 1, 2, 3, ..., n are called natural numbers. The product of first "n" natural numbers is called "n" factorial. It is denoted by n!.

Mathematically,

$n! = n(n - 1)(n - 2), \ldots\ldots \times 3 \times 2 \times 1$

e.g. $4! = 4 \times 3 \times 2 \times 1 = 24$

Permutation

An orderly arrangement of distinct objects is called permutation. The number of permutations of "r" objects selected from "n" distinct objects is given by

$$^{n}P_r = \frac{n!}{(n - r)!}$$

e.g. $$^{4}P_2 = \frac{4!}{(4 - 2)!}$$

$$= \frac{4!}{2!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = 12$$

Combination

A non-orderly arrangement of objects is called combination.

The umber of combinations of "r" objects selected from "n" objects is given by

$$^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

e.g. $\binom{6}{2} = \frac{6!}{2!(6-2)!}$

$$= \frac{6!}{2!(4!)} = \frac{6 \times 5 \times 4!}{2 \times 1 \times 4!} = 15$$

7.7 EXPERIMENT

An experiment is a well-defined action of obtaining an observation.

Trial

Single performance of an experiment is called trial.

Out Come

The result obtained from an experiment is called outcome.

Random Experiment

An experiment which gives different results even though it is repeated a large number of times under similar conditions is called random experiment. The tossing of a fair coin and throwing a die etc. are the examples of random experiments.

7.8 SAMPLE SPACE

The set of all possible outcomes of a random experiment is called a sample space. *e.g.* when we toss a coin the possible outcomes are head and tail. It is denoted by "S".

$S = \{H, T\}$

If we throw a balance die, then the sample space will be

$S = \{1, 2, 3, 4, 5, 6\}$

Sample Point

Each element of the sample space is called sample point. For example, when we toss a fair coin head and tail are two sample points.

Event

Any subset of a sample space is called an event.

Simple Event

An event that contains only one sample point is called a simple event *e.g.* the occurrence of 3, when a die is thrown.

i.e. Let $A = \{3\}$

Compound Event

An event that contains more than one sample point is called compound event. *e.g.* occurrence of even numbers when a die is thrown.

i.e. {2, 4, 6}

Sure Event

An event that is certain to occur is called certain event or sure event. The sample space S is a sure event.

Impossible Event

An event, which can never occur, is called an impossible event. The empty set *i.e.* ϕ is an impossible event.

Equally Likely Events

Two events A and B are said to be equally likely, if they have same chances of occurrence. *e.g.* when a coin is tossed, it has to possible out comes *i.e.* head and tail. Both these faces are equally likely.

Mutually Exclusive Events

Two events A and B are mutually exclusive if and only if they can not both occur at the same time. For example, when a coin is tossed, we get either a head or a tail, but not both. Head and tail are mutually exclusive events.

Collectively Exhaustive Events

Two mutually exclusive events are called collectively exhaustive if their union constitutes the entire sample space. For example, when we toss a coin, the two events *i.e.* head and tail are collectively exhaustive.

7.9 BASIC INFORMATIONS ABOUT PLAYING CARDS

Total number of cards = 52

They are arranged in four suits. Each suit contains 13 cards. The four suits are diamond, heart, club and spade.

Diamond and Heart are red cards while club and space are black.

The face values of 13 cards in each suit is Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen and King.

Jack, Queen and King are called picture cards.

Mathematically,

Total cards = 52

Diamond cards = 13

Heart cards = 13

Spade cards = 13

Club cards = 13

Red cards = Diamonds + Hearts = 13 + 13 = 26

Black cards = Spades + Clubs = 13 + 13 = 26

13+13=26

Total number of kings = 4

Total number of queens = 4

Total number of jacks = 4

Picture cards = 12 (three from each suit)

EXAMPLE 7.1

Construct the sample space for the following random experiments.

(i) When a fair coin is tossed.

(ii) When two coins are tossed.

(iii) When a die is rolled.

(iv) When pair of dice is thrown

SOLUTION

Coin N=2 , Die N=6

SOLUTION

(i) Total number of outcomes = 2^1 = 2

The two possible outcomes are head and tail therefore,

S = {H, T}

(ii) Total number of outcomes when two coins are tossed

= 2^2 = 4

Possible outcomes are as under

	Coin I	
Coin II	**H**	**T**
H	HH	HT
T	TH	TT

Therefore

S = {(HH), (HT), (TH), (TT)}

(iii) When a dire is rolled.

Total no. of samples / outcomes = 6^1 = 6

The sample space will be

S = {1, 2, 3, 4, 5, 6}

(iv) When a pair of dice is rolled / thrown.

Total number of outcomes $= 6^2 = 36$

The possible outcomes are shown below.

Ist die	IInd die					
	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Sample space of the experiment is

$$S = \left\{ \begin{array}{l} (1,1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ (2,1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ (3,1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ (4,1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ (5,1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ (6,1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \end{array} \right\}$$

$P(A) = \frac{n(A)}{n(S)}$

$= 6/36$

$= 1/6$

7.10 DEFINITIONS OF PROBABILITY

There are two ways to define probability namely objective approach and subjective approach.

(a) Objective Approach

In term of objective approach there are three different definitions.

1. Classical or A Priori Definition of Probability

This definition may be stated as follows if a random experiment can produce 'n' mutually exclusive and equally likely outcomes and 'm' of which are favorable to the occurrence of event A, then the probability of event A denoted by P(A) is defined as the ratio $\frac{m}{n}$.

$$P(A) = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

i.e. $P(A) = \frac{m}{n}$

This definition was given by P.S. Laplace. The classical definition has following shortcomings.

(a) This definition is not applicable, when the outcomes are not equally likely.

(b) This definition becomes violate, when the number of possible outcomes may be infinite.

2. Relative Frequency or Posteriori Definition of Probability

If a random experiment is repeated a large number of times say n under uniform conditions and if an event A occurs m times, then the probability of A is defined as the limit of relative frequency $\frac{m}{n}$ as n tends to infinity.

i.e. $$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

This definition has two defects.

(i) It is not applicable if the total numbers of ways are limited.

(ii) This definition becomes violate if the conditions are not same.

3. The Axiomatic Definition of Probability

Let S be the sample space with the sample points $E_1, E_2, \ldots, E_i, \ldots, E_n$. Each sample point is assigned a real number like $P(E_1), P(E_2), P(E_i), \ldots P(E_n)$, which are called the probabilities of E_i, that must satisfy the following axioms.

(i) $0 \le P(Ei) \le 1$

(ii) $P(S) = 1$, for the sure event.

(iii) If A and B are mutually exclusive events.

Then $P(A \cup B) = P(A) + P(B)$

(a) Probability of an Event

Let an event belongs to a sample space which contains a finite and equally likely sample points, then probability of an event A is given by

$$P(A) = \frac{\text{Number of sample points in A}}{\text{Number of smaple point in S}}$$

i.e. $$P(A) = \frac{n(A)}{n(S)}$$

(b) Subjective or Personalistic Probability

In this approach probability is based on the degree of confidence or belief that a reasonable person has in the occurrence of an event A. This probability is subjective in nature, because the probability vary from person to person for the same situation.

EXAMPLE 7.2

A fair coin is tossed. What is the probability of getting a

(i) Head

(ii) Tail

SOLUTION

Total number of outcomes $= 2^1 = 2$

The possible outcomes are head and tail.

Therefore the sample space is

$S = \{H, T\}, \ n(s) = 2$

(i) Let A denote the event that head appears,

$A = \{H\}, \ n(A) = 1$

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{2}$$

(ii) Let B denote the event that tails appears.

$B = \{T\}, \ n(B) = 1$

$$P(B) = \frac{n(B)}{n(S)} = \frac{1}{2}$$

EXAMPLE 7.3

A die is rolled. What is the probability that it shows.

(i) a six

(ii) even number

(iii) odd number

SOLUTION

Total number of possible outcomes, when a die is rolled $= 6^1 = 6$

Sample space will be

$S = \{1, 2, 3, 4, 5, 6\}, \ n(S) = 6$

(i) Let A denotes the event that 6 appears.

$A = \{6\}, \ n(A) = 1$

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{6}$$

composed by ARFAN ALI MATHEMATICIAN 03148100197

(ii) Let B denotes the event that a die shows an even numbers.

$B = \{2, 4, 6\}, \ n(B) = 3$

$$P(B) = \frac{n(B)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

(iii) Let C denotes the event that die shows an odd numbers.

$C = \{1, 3, 5\}, \ n(C) = 3$

$$P(C) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

EXAMPLE 7.4

What is the probability of getting exactly two heads when 3 coins are tossed?

SOLUTION

Total number of possible outcomes $= 2^3 = 8$

The sample space of the experiment is

S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

$n(S) = 8$

Let A denote the event that exactly two heads appears,

A = {HHT, HTH, THH}, $n(A) = 3$

Therefore

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{8}$$

EXAMPLE 7.5

From a well-shuffled pack of 52 cards, a card is drawn at random. What is the probability that it is

(i) Black card
(ii) An ace
(iii) A card of spade
(iv) A jack of club
(v) A pictured card

SOLUTION

The total number of possible outcomes $= \binom{52}{1} = 52$

$n(S) = 52$

(i) Let A represents the event that the card is black.

Then $n(A) = 26$

Therefore

$$P(A) = \frac{n(A)}{n(S)} = \frac{26}{52} = \frac{1}{2}$$

(ii) Let B represents the event that the card drawn is an ace $n(B) = 4$.

(There are four aces in a pack of cards)

Therefore

$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{52} = \frac{1}{13}$$

(iii) Let C = {A card is spade} , $n(C) = 13$

$$P(C) = \frac{n(C)}{n(S)} = \frac{13}{52} = \frac{1}{4}$$

(iv) Let D = {A jack of club}

then $n(D) = 1$ (There is only one jack of club)

Therefore

$$P(D) = \frac{n(D)}{n(S)} = \frac{1}{52}$$

(v) Let E = {A pictured card} , $n(E) = 12$

$$P(E) = \frac{n(E)}{n(S)} = \frac{12}{52} = \frac{3}{13}$$

EXAMPLE 7.6

(a) If two six faced balance dice are thrown. Find the probability that

(i) The same number appears on both the dice.

(ii) The sum of two numbers is greater than 9.

(b) Two fair dice are thrown. Find the chance that sum of numbers on two dice is at least seven.

(c) What is the probability of throwing a sum 7 with two dice.

SOLUTION

Total number of outcomes when two dice are thrown = $6^2 = 36$

The sample space for the experiment is

$$S = \left\{ \begin{array}{l} (1,1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ (2,1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ (3,1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ (4,1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ (5,1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ (6,1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \end{array} \right\}$$

$n(S) = 36$

(a) **(i)**

Let A = {The same number appears on both the dice}

$A = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

Then $n(A) = 6$

$$P(A) = \frac{n(A)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

(ii) Let B = {The sum of two number is greater than 9}

$B = \{(4, 6), (5, 5), (6, 4), (5, 6), (6, 5), (6, 6)\}$

Then $n(B) = 6$

$$P(B) = \frac{n(B)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

(b) Let C = {Sum of numbers on two dice is at least seven}

$$C = \left\{ \begin{array}{l} (1, 6), (2, 5), (2, 6), (3, 4), (3, 5), (3, 6), (4, 3), (4, 4) \\ (4, 5), (4, 6), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1) \\ (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \end{array} \right\}$$

$n(C) = 21$

Therefore

$$P(C) = \frac{n(C)}{n(S)} = \frac{21}{36} = \frac{7}{12}$$

(c) Let D = {Sum is 7 with two dice}

$D = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$

Then $n(D) = 6$

$$P(D) = \frac{n(D)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

EXAMPLE 7.7

A digit is selected at random from the first ten natural numbers. Find the probability that the selected digit is

(i) Greater than 6

(ii) Multiple of 3

(iii) An even number

SOLUTION

$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, $n(S) = 10$

(i) Let A be the event that the selected digit is greater than 6,

then $A = \{7, 8, 9, 10\}$

Then $n(A) = 4$

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{10} = \frac{2}{5}$$

(ii) Let B be the event that the selected digit is multiple of 3.

$B = \{3, 6, 9\}$, $n(B) = 3$

$$P(B) = \frac{n(B)}{n(S)} = \frac{3}{10}$$

(iii) Let C be the event that the selected digit is an even number.

Then $C = \{2, 4, 6, 8, 10\}$, $n(C) = 5$

$$P(C) = \frac{n(C)}{n(S)} = \frac{5}{10} = \frac{1}{2}$$

EXAMPLE 7.8

Two cards are drawn from a well-shuffled pack of 52 playing cards. Find the probability that

(i) Both are King

(ii) One is king and other is queen

(iii) One is red and other is black

(iv) Both cards are of the same colour.

* One card is queen of diamond

* Both are of different colour

SOLUTION

Total no. of possible outcomes $= \binom{N}{n} = \binom{52}{2} = 1326$

$n(S) = 1326$

(i) Let A denotes the event that both cards are king

$n(A) = \binom{4}{2}\binom{48}{0}$, $n(A) = 6$

Let E = Both are of different colour

$E = \binom{26}{1}\binom{26}{1} = 676$

$P(A) = \frac{n(A)}{n(S)} = \frac{6}{1326} = \frac{1}{221}$

(ii) Let B denotes the event that one card is king and other is queen.

$n(B) = \binom{4}{1}\binom{4}{1}\binom{44}{0}$

$= 4 \times 4 \times 1 = 16$

$P(B) = \frac{n(B)}{n(S)} = \frac{16}{1326} = \frac{8}{663}$

(iii) Let C be the event that one card is red and other is black.

$n(C) = \binom{26}{1}\binom{26}{1}$

$= 26 \times 26 = 676$

$P(C) = \frac{n(C)}{n(S)} = \frac{676}{1326}$

(iv) Let D denotes the event that both cards are of the same colour (*i.e.* either both are black or both are red)

$n(D) = \binom{26}{2} + \binom{26}{2}$

$= 325 + 325 = 650$

$P(D) = \frac{n(D)}{n(S)} = \frac{650}{1326} = \frac{25}{51}$

EXAMPLE 7.9

A bag contains 5 red and 7 black balls. What is the probability of drawing at random.

(i) One red ball

(ii) One black ball

SOLUTION

Total balls in the bag = 5 + 7 = 12

Total no. of outcomes when one ball is selected from 12 balls $= \binom{12}{1} = 12$

(i) Let A = {Drawing red ball}

$n(A) = 5$

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{12}$$

(ii) Let B = {Drawing black ball}

$n(B) = 7$

$$P(B) = \frac{n(B)}{n(S)} = \frac{7}{12}$$

EXAMPLE 7.10

A bag contains 5 white and 7 black balls. If 3 balls are drawn at random from the bag. What is the probability that

(i) All are black

(ii) Two are white

(iii) All balls are of the same colour

SOLUTION

Whit balls	Black balls	=	Total
5	7	=	5 + 7 = 12

Total number of possible outcomes $= \binom{12}{3} = 220$

$n(S) = 220$

(i) Let A = {All balls are black}

$$n(A) = \binom{7}{3}\binom{5}{0}$$

$= 35 \times 1 = 35$

$$P(A) = \frac{n(A)}{n(S)} = \frac{35}{220} = \frac{7}{44}$$

(ii) Let B = {Two balls are white}

$$n(B) = \binom{5}{2} \times \binom{7}{1}$$

$= 10 \times 7 = 70$

$$P(B) = \frac{n(B)}{n(S)} = \frac{70}{220} = \frac{7}{22}$$

(iii) Let C = {All balls are of the same colour}

$$n(C) = \binom{5}{3} + \binom{7}{3}$$

$$= 10 + 35 = 45$$

$$P(C) = \frac{n(C)}{n(S)} = \frac{45}{220} = \frac{9}{44}$$

7.11 IMPORTANT POINTS THAT ARE COMMONLY USED IN THE PROBABILITY QUESTIONS

Either, Or, At Least

While solving the different questions on probability we face the word Either, Or, At least used in the question. For example, A or B, either A or B, At least one of them. In these situations we use the addition law of probability for mutually and not mutually exclusive events.

A and B, Both

If A and B or both appear in the question, we use multiplication law of probabilities.

Law of Complementation

If $\bar{A}$ is the complement of an event A relative to the sample space S, then

$$P(\bar{A}) = 1 - P(A)$$

7.12 ADDITION LAW FOR NOT MUTUALLY EXCLUSIVE EVENTS

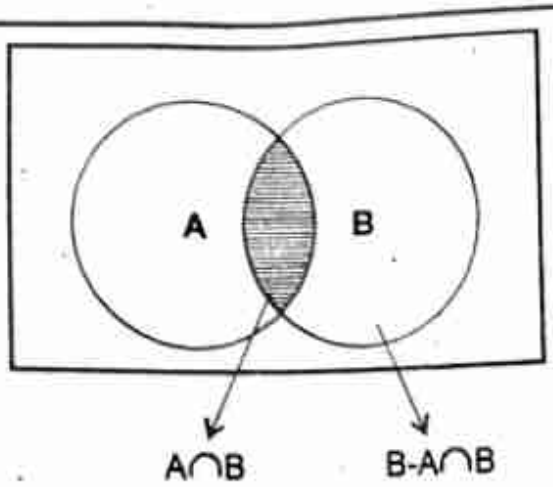
If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This law states that If A and B are two not mutually exclusive events, then the probability that at least one of them occurs is given by sum of separate probabilities of events A and B minus the probability of the joint event $A \cap B$.

Proof

From the Venn diagram we know that, the set $A \cup B$ may be written as the union of the two mutually exclusive event A and $(B - A \cup B)$.



That is

$A \cup B = A \cup [B - A \cap B]$

$P(A \cup B) = P[A \cup (B - A \cap B)]$

$P(A \cup B) = P(A) + P(B - A \cap B)$ ..... (i)

Again Event B may be written as the union of two mutually exclusive event $A \cap B$ and $(B - A \cap B)$.

i.e. $B = (A \cap B) \cup (B - A \cap B)$

$P(B) = P(A \cap B) + P(B - A \cap B)$

$P(B) - P(A \cap B) - P(B - A \cap B) = 0$

$P(B - A \cap B) = P(B) - P(A \cap B)$

Putting the value of $P(B - A \cap B)$ in equation (i) we have

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

EXAMPLE 7.11

Two coins are tossed. Find the probability that both faces are head or at least one head.

SOLUTION

Total no. of possible outcomes $= 2^2 = 4$

The sample space for the experiment is

$S = \{HH, HT, TH, TT\}$, $n(S) = 4$

Let $A = \{$Both faces are head$\}$

$B = \{$At least one head$\}$

We have to calculate $P(A \cup B)$

$A = \{HH\}$, $n(A) = 1$

$$P(A) = \frac{n(A)}{n(S)} = \frac{1}{4}$$

$B = \{TH, HT, HH\}$, $n(B) = 3$

$$P(B) = \frac{n(B)}{n(S)} = \frac{3}{4}$$

Events A and B are not mutually exclusive, therefore

$A \cap B = \{HH\}$, $n(A \cap B) = 1$

$$P(A \cap B = \frac{n(A \cap B)}{n(S)} = \frac{1}{4}$$

We know that from addition law

$P(A \cap B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{1}{4} + \frac{3}{4} - \frac{1}{4} = \frac{3}{4}$$

EXAMPLE 7.12

One card is selected at random from a deck of 52 playing cards. What is the probability that the card is Heart or picture card or both?

SOLUTION

Total number of possible outcomes $= \binom{52}{1} = 52$

$n(S) = 52$

Let $A = \{\text{Card is heart}\}$

$B = \{\text{Pictured card}\}$

$A \cap B = \{\text{Card selected is both heart and pictured}\}$

Then we need $P(A \cup B)$

$n(A) = 13$

$$P(A) = \frac{n(A)}{n(S)} = \frac{13}{52}$$

$n(B) = 12$ (3 cards in each suit)

i.e. $3 \times 4 = 12$

$$P(B) = \frac{n(B)}{n(S)} = \frac{12}{52}$$

$n(A \cap B) = 3$, (Because 3 of hearts are also picture cards)

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{3}{52}$$

Therefore

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{13}{52} + \frac{12}{52} - \frac{3}{52}$$

$$= \frac{13 + 12 - 3}{52} = \frac{22}{52} = \frac{11}{26}$$

EXAMPLE 7.13

An integer is chosen at random from the first 50 positive integers. What is the probability that the integer chosen is divisible by 4 or by 6?

SOLUTION

The sample space for this experiment is

$S = \{1, 2, 3, \ldots\ldots, 50\}$, $n(S) = 50$

Let $A = \{\text{Integer is divisible by } 4\}$

$B = \{\text{Integer chosen is divisible by } 6\}$

$A \cap B = \{\text{Int. chosen is divisible by both } 4 \text{ \& } 6 \textit{ i.e. } 12\}$

Then we need $P(A \cup B)$

$A = \{4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48\}$

$n(A) = 12$, $P(A) = \frac{12}{50}$

$B = \{6, 12, 18, 24, 30, 36, 42, 48\}$

$n(B) = 8$, $P(B) = \frac{8}{50}$

$A \cap B = \{12, 24, 36, 48\}$

$n(A \cap B) = 4$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{4}{50}$$

Therefore

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{12}{50} + \frac{8}{50} - \frac{4}{50} = \frac{16}{50} = \frac{8}{28}$$

EXAMPLE 7.14

Two students A and B can solve 50% and 80% problems respectively from the exercise. What is the probability that either A or B can solve a problem chosen at random from that exercise.

SOLUTION

P(A can solve the problem) = 50%

$$P(A) = \frac{50}{100} = 0.5$$

P(B can solve the problem) = 80%

$$P(A) = \frac{80}{100} = 0.8$$

P(Both A and B can solve the problem)

$P(A \cap B) = P(A) \cdot P(B)$

$= 0.5 \times 0.8 = 0.4$

Therefore

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$= 0.5 + 0.8 - 0.4$

$= 13 - 0.4 = 0.90$

7.12.1 Addition Law for Mutually Exclusive Events

If two events A and B are mutually exclusive, then

$P(A \cup B) = P(A) + P(B)$

Proof

As we know that from addition law for not mutually exclusive events

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ..... (i)

In this case events A and B are mutually exclusive, therefore

$A \cap B = \phi$, $P(A \cap B) = 0$

Putting the value of $P(A \cap B)$ in equation (i), we have

$P(A \cup B) = P(A) + P(B)$

EXAMPLE 7.15

Two six faced balance dice are thrown. Find the probability that the sum of two numbers is 8 or 10.

SOLUTION

Total number of outcomes $= 6^2 = 36$

The sample space for the experiment is

$$S = \left\{ \begin{array}{l} (1,1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ (2,1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ (3,1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ (4,1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ (5,1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ (6,1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \end{array} \right\}$$

$n(S) = 36$

Let A be the event that the sum is 8 and B be the event that the sum is 10.

We need to find $P(A \cup B)$

Then the two events are

$A = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}, \; n(A) = 5$

$B = \{(4, 6), (5, 5), (6, 4)\}, \; n(B) = 3$

The events A and B are mutually exclusive, *i.e.* sum of 8 and 10 cannot both occur at the same time.

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{3}{36}$$

Therefore

$$P(A \cup B) = P(A) + P(B)$$

$$= \frac{5}{36} + \frac{3}{36} = \frac{8}{36} = \frac{2}{9}$$

EXAMPLE 7.16

A card is selected at random from a well-shuffled pack of 52 playing cards. Find the probability that the card is king or queen.

SOLUTION

There are 52 possible outcomes

$n(S) = 52$

Let $A = \{\text{A card is king}\}$

$B = \{\text{A card is queen}\}$

Then we need $P(A \cup B)$

$n(A) = 4$, (as there are 4 kings)

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{52} = \frac{1}{13}$$

$n(B) = 4$, (as there are 4 queens)

$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{52} = \frac{1}{13}$$

The two events A and B are mutually exclusive, because king and queen can not both occur together.

Therefore

$$P(A \cup B) = P(A) + P(B)$$

$$= \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

EXAMPLE 7.17

For any two events A and B, it is known that $P(A) = \frac{2}{3}$

$P(A \cup B) = \frac{7}{12}$ and $P(A \cap B) = \frac{5}{12}$. Find P(B).

SOLUTION

We know that form Addition law of probabilities

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Putting the values, we have

$$\frac{7}{12} = \frac{2}{3} + P(B) - \frac{5}{12}$$

$$P(B) = \frac{7}{12} - \frac{2}{3} + \frac{5}{12}$$

$$= \frac{7 - 8 + 5}{12} = \frac{4}{12} = \frac{1}{3}$$

7.13 CONDITIONAL PROBABILITY

If A and B are two events in a sample space S, then conditional probability of event A, given that event B has occurred, denoted by P(A/B) is defined as

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

Similarly

$$P(BA) = \frac{P(A \cap B)}{P(A)}, \quad P(A) \neq 0$$

EXAMPLE 7.18

A pair of dice is rolled. What is the conditional probability that the sum of two dice will be 8, given that the two dice had the same outcomes.

SOLUTION

The sample space consists of 36 sample points.

i.e. $n(S) = 36$ (See example 5.15)

Let $A = \{\text{Sum is } 8\}$

$B = \{\text{the two dice had the same outcome}\}$

Then we require P(A/B)

Therefore

$A = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$

$B = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

$n(A) = 5$

$(A \cap B) = \{(4, 4)\}$

$n(A \cap B) = 1$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{36}$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{1}{36}$$

$$p(B) = \frac{n(B)}{n(S)} = \frac{6}{36}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{1}{36} \times \frac{36}{6} = \frac{1}{6}$$

7.14 INDEPENDENT AND DEPENDENT EVENTS

Two events A and B are said to be independent if the occurrence of one does not effect the other.

Mathematically,

Two events A and B are independent iff

(i) $P(A/B) = P(A)$

$P(B/A) = P(B)$

(ii) $P(A \cap B) = P(A) \cdot P(B)$

Otherwise they are dependent.

7.15 MULTIPLICATION LAW

If A and B are any two events

Then $P(A \cap B) = P(A) \cdot P(B / A)$

$= P(B) \cdot P(A / B)$

Proof

As we know that from conditional probability

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Multiplying both sides by P(A), we have

$$P(A) \cdot P(B/A) = \frac{P(A \cap B)}{P(A)} \cdot P(A)$$

$$P(A) \cdot P(B/A) = P(A \cap B)$$

Or $P(A \cap B) = P(A) \cdot P(B/A)$

Similarly

$P(A \cap B) = P(B) \cdot P(A/B)$ can also be proved.

EXAMPLE 7.19

A box contains 8 items, 3 of which are defective and 5 are good. Two items are selected at random. What is the probability that the first is good and second is defective.

SOLUTION

Let A be the event that the first item selected is good and B be the event that second item is defective.

Then we need to calculate the probability of $A \cap B$ which is given by the formula

$$P(A \cap B) = P(A) \cdot P(B/A)$$

$$n(A) = 5$$

$$P(A) = \frac{5}{8}$$

One good item is selected. Now there remains 7 items of which 4 are good and 3 are defective. Therefore the probability of selecting a defective after a good has been selected.

$$P(B/A) = \frac{3}{7}$$

$$P(A \cap B) = P(A) \cdot P(B/A)$$

$$= \frac{5}{8} \times \frac{3}{7} = \frac{15}{56}$$

7.15.1 Multiplication Law for Independents Events

If A and B are two independent events

Then $P(A \cap B) = P(A) \cdot P(B)$

EXAMPLE 7.20

Two dice are thrown twice. Find the probability of getting a total of 6 on the first throw and a total of 10 on the second.

SOLUTION

The sample space of this experiment consists of 36 sample points. *i.e.* $n(S) = 36$. (See example 5.15)

Let A be the event that getting a total of 6 on first throw and B be the event that getting a total of 10 on the second throw.

Then $A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$

$n(A) = 5$

$P(A) = \frac{n(A)}{n(S)} = \frac{5}{36}$

$B = \{(4, 6), (5, 5), (6, 4)\}$

$n(B) = 3$

$P(B) = \frac{n(B)}{n(S)} = \frac{3}{36}$

The event A and B are independent, therefore

$P(A \cap B) = P(A) \cdot P(B)$

$= \frac{5}{36} \times \frac{3}{36} = \frac{5}{432}$

EXAMPLE 7.21

The probability that A will be alive after 10 years is $\frac{5}{7}$ and for B it is $\frac{7}{9}$. Find the probability

(i) Both of them will die

(ii) A will be alive and B dead

(iii) B will be alive and A dead

(iv) Both of them will be alive

(v) At least one will be alive in 10 years to come.

SOLUTION

$P(A \text{ alive}) = \frac{5}{7} \quad \Rightarrow P(A) = \frac{5}{7}$

$P(A \text{ dead}) = 1 - P(A) \quad \Rightarrow P(\bar{A}) = 1 - \frac{5}{7} = \frac{2}{7}$

$P(B \text{ alive}) = P(B) = \frac{7}{9}$

$P(B \text{ dead}) = 1 - P(B) \Rightarrow P(B) = 1 - \frac{7}{9} = \frac{2}{9}$

(i) We need the probability that both of them will die.

i.e. $P(\bar{A} \cap \bar{B})$ since events A and B are independent.

Therefore

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \cdot (\bar{B})$$

$$= \frac{2}{7} \times \frac{2}{9} = \frac{4}{63}$$

(ii) We need the probability that A will be alive and B dead.

i.e. $P(A \cap \bar{B})$

Therefore

$$P(A \cap \bar{B}) = P(A) \cdot (\bar{B})$$

$$= \frac{5}{7} \times \frac{2}{9} = \frac{10}{63}$$

(iii) We need the probability that A will be dead and B alive.

i.e. $P(\bar{A} \cap B)$

$$P(\bar{A} \cap B) = P(\bar{A}) \cdot P(B)$$

$$= \frac{2}{7} \times \frac{7}{9} = \frac{2}{9}$$

(iv) We require the probability that both of them will be alive.

i.e. $P(A \cap B)$

Therefore

$P(A \cap B) = P(A) \cdot P(B)$ (Events A and B are independent)

$$= \frac{5}{7} \times \frac{7}{9} = \frac{5}{9}$$

(v) We require the probability that at least one will be alive.

i.e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{5}{7} + \frac{7}{9} - \frac{5}{9}$$

$$= \frac{45 + 49 - 35}{63} = \frac{59}{63}$$

EXAMPLE 7.22

Two cards are drawn at random from a well shuffled pack of ordinary cards. Find the probability that the first card is a king and the second cards is queen, if the first cards is

(i) replaced (ii) not replaced

SOLUTION

Let A be the event that the first card is king and B be the event that second card is queen.

(i) In case of replacement, event A and B are independent,

Therefore

$P(A \cap B) = P(A) \cdot P(B)$

$= \frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$

(ii) If the first card is not replaced, then A and B are dependent events.

Therefore

$P(A \cap B) = P(A) \cdot P(B/A)$

= P (First card is King) P (Second card is queen, given that first card is king)

$P(A \cap B) = P(A) \cdot P(B / A)$

$= \frac{4}{52} \times \frac{4}{51} = \frac{4}{663}$

SUMMARY

The summary of probability is given below:

NAME	FORMULA
Probability	$= \frac{\text{No. of favourable outcomes}}{\text{Total no. of outcomes}}$ $P(A) = \frac{n(A)}{n(S)}$ $P(B) = \frac{n(B)}{n(S)}$
Addition law for mutually exclusive events	$P(A \cup B) = P(A) + P(B)$
Addition law for not mutually exclusive events	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Law of complementation	$P(\overline{A}) = 1 - P(A)$
Conditional Probability	$P(A/B) = \frac{P(A \cap B)}{P(B)}$
Independent law	$P(A/B) = P(A)$ $P(B/A) = P(B)$
Multiplication law for dependent events	$P(A \cap B) = P(A) . P(B/A)$
Multiplication law for Independent events	$P(A \cap B) = P(A) . P(B)$

EXERCISES

7.1 Define the following with examples.

(i) Set **(ii)** Null set

(iii) Sub set **(iv)** Intersection of two sets

(v) Universal set **(vi)** Product of sets

7.2 Given

$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$A = \{1, 2, 3\}$, $B = \{2, 4, 6, 8\}$

$C = \{4, 5, 6, 7, 8\}$, $D\ \{3, 4\}$

Find

(i) $A \cap B$ **(ii)** $B \cup D$

(iii) $A \cap C$ **(iv)** $A \times D$

7.3 (a) Define complement of the set

(b) Let $U = \{1, 2, 3, \ldots, 10\}$

$A = \{1, 3, 5, 7, 9\}$

$B = \{2, 4, 6, 8, 10\}$

Find **(i)** $\bar{A}$ **(ii)** $\bar{B}$

(iii) $\bar{A} \cup \bar{B}$ **(iv)** $\bar{A} \cap \bar{B}$

7.4 (a) Explain the followings.

(i) Permutation

(ii) Combination

(iii) Factorial

(b) Solve the following

(i) ${}^{9}P_2$

(ii) ${}^{20}P_3$

(iii) $\binom{10}{2}$

(iv) $\binom{52}{5}$

7.5 Define the following terms with examples.

(i) Random experiment

(ii) Sample space

(iii) Simple and compound event

(iv) Mutually exclusive events

(v) Equally likely events

7.6 (a) Define classical or priori definition of probability.

(b) Find the probability of each of the following outcomes.

(i) A tail appears, when a fair coin is tossed.

(ii) Two head appears when two coins are tossed.

(iii) An ace comes up, when a fair die is rolled.

(iv) An odd number comes up, when a true die is rolled.

(c) Two coins are tossed. What is the probability of getting?

(i) One head (ii) Two heads

(iii) At least one head (iv) At most two heads

7.7 (a) If three coins are tossed? Construct the sample space and find the probability of 3 heads.

(b) Three coins are tossed. What is the probability of getting?

(i) Two heads

(ii) At least one head

(iii) At most one head

7.8 (a) Two fair dice are rolled. Find the probability that

(i) Sum of two dice will be even

(ii) The two dice had the same outcomes.

(b) Two fair dice are rolled. Find the probability.

(i) The sum is odd

(ii) The sum is at least 9

(iii) Sum is exactly seven

(iv) A double six

(c) Two dice are rolled. Find the probability of getting sum is less than 7 is equal to sum is greater than 7.

(d) Two dice are rolled once. Find the probability that

(i) Sum is greater than 9. (ii) Sum is at most 5.

(iii) Their product is even

7.9 (a) a bag contains 3 white, 5 yellow and 2 red balls. What is the probability of drawing at random.

(i) One red ball

(ii) One yellow ball.

(b) A bag contains 3 white and 4 black balls. Two balls are drawn at random find the probability that

(i) One is white and other is black

(ii) Both are white

(iii) At least one is white

7.10 (a) A bag contains 3 black, 4 white and 5 red balls. If 3 balls are drawn at random from the bag. Find the probability that

(i) One black, one white and one red ball

(ii) Two red and one black ball

(iii) All balls are black

(b) A box contains 3 red, 4 white and 5 black balls. Three balls are drawn from the box. Find the probability that they may be

(i) all of different colours

(ii) all are of the same colours

7.11 From a well shuffled pack of 52 playing cards, a card is drawn at random. What is the probability that it is

(i) Red card (ii) Kind of Diamond

(iii) Card is Queen (iv) Card is Heart

7.12 (a) Two cards are drawn at random from a well shuffled pack of playing cards. Find the probability that

(i) Both are red. 4

(ii) One is heart and other is diamond

(iii) Both are of the same colour

(b) A digit is selected at random from first 15 natural numbers. Find the probability that the selected digit is

(i) An odd number

(ii) An even number

(iii) Less than 5

(iv) Complete square

(v) Divisible by 5

7.16 (a) State and prove addition law of probabilities for not mutually exclusive events.

(b) A fair die is rolled. What is the probability that the outcome is either an odd number or a prime number?

HINT

Odd Number

A number which is not divisible by 2

i.e. 1, 3, 5, ...

Prime Number

A number which is divisible by 1 and itself only

i.e. 2, 3, 5, ...

7.14 (a) An integer is chosen at random from the first positive 100 integers. Find the probability that the chosen integer is divisible by 8 or 12.

(b) A class contains 10 men and 20 women out of which half men and half women have brown eyes. Find the probability that the person chosen is a women or has brown eyes.

7.15 (a) A card is drawn from a well shuffled pack of 52 playing cards. What is the probability that card is either king or diamond?

(b) The probability that a boy will pass the examination is 0.70 and that for girl is 0.60. What is the probability that at least one of them will pass the examination?

.6 (a) Define addition law of probability for mutually exclusive events.

(b) A card is drawn from a well shuffled pack of 52 playing cards. Find the probability that the card is either king or queen.

7.17 (a) A ball is drawn at random from a bag containing 5 red, 6 white, 4 blue and 3 orange balls. Find the probability that the ball is either red or blue.

(b) Two balance dice are rolled. Find the probability of getting a total of either 7 or 9.

7.18 (a) What do you meant by conditional probability?

(b) A fair die is rolled. Find the conditional probability that the face is odd given that the face is less than 4.

7.19 (a) A man tosses two fair dice. What is the conditional probability that the sum of two top sides is 7, given that

(i) the sum is odd

(ii) the sum is greater than 6.

(iii) the two dice had the same outcome.

(b) Three fair coins are tossed. What is the conditional probability that 2 head appears given that at least one head.

7.20 (a) State and prove multiplication law of probabilities.

(b) Define independent and dependent events.

(c) Two cards are drawn from a well-shuffled pack of 52 playing cards. Find the probability that they are both aces if the first card is

(i) replaced (ii) not replaced

7.21 (a) Find $P(A \cap B) = ?$

Given that $P(A) = 0.25$, $P(B) = 0.60$

Assuming that A and B are independent events.

(b) given $P(A) = 0.60$

$P(B) = 0.40$, $P(A \cap B) = 0.24$

Find $P(A/B)$, $P(B/A)$ and $P(A \cup B)$

7.22 Two friends A and B appear in an interview for two posts in an office. The probability that A will be selected is $\frac{1}{3}$ and the probability that B will be selected is $\frac{1}{4}$. What is the probability that

(i) Both will be selected (ii) Only A will be selected

(iii) Only B will be selected (iv) At least one will be selected

(v) Neither will be selected.

7.23 Two coins are tossed. Show that the probability of getting at least one head is ¾.

7.24 A fair dice is rolled twice. Find the probability that sum of numbers appearing is

(i) at most five

(ii) at least 10

7.25 An unbiased coin is tossed twice. Find the probability that

(i) one tail appears

(ii) at least lone head

7.26 A card is drawn from well shuffled pack of 52 cards. What is the prob that it is

(i) Black card (ii) A face card

7.27 EACH QUESTION HAS FOUR POSSIBLE ANSWERS. SELECT THE CORRECT ONE AND ENCIRCLE IT.

(i) **The probability of an event always lies between**

(a) −1 and 0 (b) 0 and 1

(c) −1 and +1 (d) 1 and 100

(ii) **A letter is chosen as random, from the word "SAMPLE", the probability of vowel is**

(a) $\frac{3}{6}$ (b) $\frac{2}{3}$

(c) $\frac{1}{3}$ (d) $\frac{1}{6}$

(iii) **The probability of an impossible event is always equal to**

(a) 1 (b) 0

(c) −1 (d) none of these

(iv) **Two coins are tossed. The probability of getting 2 heads is**

(a) $\frac{1}{4}$ (b) $\frac{1}{2}$

(c) $\frac{3}{4}$ (d) 1

(v) **If A and B are independent events then $P(A \cap B)$ is**

(a) $P(A) + P(B)$ (b) $P(A) \cdot P(B/A)$

(c) $P(A) \cdot P(B)$ (d) $P(A) \cdot P(A/B)$

(vi) **The probability of drawing a king from pack of 52 cards is**

(a) $\frac{1}{52}$ (b) $\frac{1}{13}$

(c) $\frac{2}{13}$ (d) $\frac{2}{52}$

(vii) **A tail appears when a fair coin is tossed**

(a) $\frac{1}{2}$ (b) $\frac{1}{3}$

(c) $\frac{2}{3}$ (d) 1

(viii) **An ace comes up, when a die is rolled**

(a) $\frac{2}{3}$ (b) $\frac{1}{3}$

(c) $\frac{1}{6}$ (d) $\frac{3}{6}$

(ix) When two coins are tossed, the possible outcomes are

(a) 1 (b) 2

(c) 4 (d) 6

(x) Two events A and B are called mutually exclusive if

(a) $A \cup B = \phi$ (b) $A \cap B = \phi$

(c) $A \cap B = S$ (d) $A \cup B = S$

(xi) An event that contains at least two sample points is called

(a) Simple event (b) Compound event

(c) Multiple event (d) Equally likely event

(xii) If two dice are rolled the possible outcomes are

(a) 6 (b) 12

(c) 18 (d) 36

(xiii) If A and B are mutually exclusive events, then

(a) $P(A \cup B) = P(A) + P(B)$

(b) $P(A \cup B) = P(A) \,.\, P(B)$

(c) $P(A \cup B) = P(A) \,.\, P(B/A)$

(d) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(xiv) The probability of drawing a picture card from a pack of 52 playing cards is

(a) $\frac{1}{52}$ (b) $\frac{1}{32}$

(c) $\frac{1}{2}$ (d) $\frac{12}{52}$

(xv) The probability of getting a red card from a pack of cards is

(a) $\frac{1}{52}$ (b) $\frac{4}{52}$

(c) $\frac{26}{52}$ (d) $\frac{3}{52}$

(xvi) If two events A and B are not independent, then

(a) $P(A \cap B) = P(A) \,.\, P(B)$ (b) $P(A \cap B) = P(A) \,.\, P(B/A)$

(c) $P(A \cap B) = P(A) + P(B/A)$ (d) $P(A \cap B) = P(A) + P(A/B)$

(xvii) A digit is selected from first ten natural numbers, the probability of even number is

(a) $\frac{1}{10}$ (b) $\frac{1}{5}$

(c) $\frac{1}{3}$ (d) $\frac{1}{2}$

(xviii) In case of objective approach, the definitions of probability are

(a) 2 (b) 3

(c) 4 (d) 5

(xix) In conditional probability P(A/B) is equal to

(a) $\frac{P(A \cup B)}{P(B)}$ (b) $\frac{P(A \cap B)}{P(B)}$

(c) $\frac{P(A \cap B)}{P(A)}$ (d) $\frac{P(A \cup B)}{P(A)}$

(xx) $\binom{n}{r}$ **is equal to**

(a) $\frac{n!}{(n-r)!}$ (b) $\frac{n!}{n!(n-r)!}$

(c) $\frac{n!}{r!(n-r)!}$ (d) $\frac{n!}{(n+r)!}$

(xxi) 0! is equal to

(a) 0 (b) 1

(c) 2 (d) 3

(xxii) If $P(A \cap B) = 0.25$, $P(A) = 0.75$ then P(B/A) is:

(a) 1 (b) 0

(c) $\frac{1}{3}$ (d) $\frac{2}{3}$

(xxiii) An orderly arrangement of objects is called:

(a) Combination (b) Permutation

(c) Sample space (d) Factorial

(xxiv) If A and B are mutually exclusive events then $P(A \cap B)$ is:

(a) 1 (b) 0.5

(c) 0 (d) None of these

ANSWERS 7.27

(i)	(b)	**(ii)**	(c)	**(iii)**	(b)	**(iv)**	(a)	**(v)**	(c)
(vi)	(b)	**(vii)**	(a)	**(viii)**	(c)	**(ix)**	(c)	**(x)**	(b)
(xi)	(b)	**(xii)**	(d)	**(xiii)**	(a)	**(xiv)**	(d)	**(xv)**	(c)
(xvi)	(b)	**(xvii)**	(d)	**(xviii)**	(b)	**(xix)**	(b)	**(xx)**	(c)
(xxi)	(b)	**(xxii)**	(c)	**(xxiii)**	(b)	**(xxiv)**	(c)		

SHORT QUESTION ANSWERS

1. Define set

Ans. A set is a well-defined collection of distinct objects. The objects of the set are called elements.

2. What is the difference between permutation and combination?

Ans. An orderly arrangement of distinct objects is called permutation while non-orderly arrangement of objects is called combination.

3. Define outcome and sample space.

Ans. **Outcome**

The result obtained from an experiment is called outcome.

Sample Space

The set of all possible outcomes of a random experiment is called sample space *e.g.* when we toss a coin the possible outcomes are head and tail, therefore

$S = \{H, T\}$

4. What do you meant by Random experiment?

Ans. An experiment which give different results even though it is repeated a large number of times under similar conditions is called random experiment *e.g.* tossing a coin, rolling a die etc.

5. What is the difference between simple event and compound event?

Ans. **Simple Event**

An event that contains only one sample point is called simple event *e.g.* occurrence of 3, when a die is rolled.

i.e. $A = \{3\}$

Compound Event

An event that contains more than one sample points is called compound event. *e.g.* occurrence of even numbers when a die is rolled.

i.e. $A = \{2, 4, 6\}$

6. Define Sure Event.

Ans. An event that is certain to occur is called sure event. For example, sample space is a sure event.

7. Define impossible event.

Ans. An event which can never occur is called an impossible event. For example, empty set is an impossible event.

8. Define equally likely events.

Ans. Two events A and B are said to be equally likely, if they have same chances of occurrence. For example, when we toss a coin, head and tail are called equally likely events.

9. What is meant by Mutually Exclusive events?

Ans. The events A and B are mutually exclusive if and only if they cannot both occur at the same time for example when a coin is tossed, we get either head or tail, but not both. Head and tail are mutually exclusive events.

10. How we define probability in terms of objective approach?

Ans. In terms of objective approach there are three definitions of probability namely,

(a) Classical or A priori definition

(b) Relative frequency or A posteriori definition

(c) The axiomatic definition of probability

11. Write down the classical definition of probability.

Ans. If a random experiment can produce 'n' mutually exclusive and equally likely outcomes and m of which are favourable to the occurrence of event A, then the probability of event A is given by

$$P(A) = \frac{m}{n} = \frac{\text{No. of favourable outcomes}}{\text{Total no. of outcomes}}$$

12. Give the relative frequency definition of probability.

Ans. If a random experiment is repeated a large number of times say n under uniform conditions and if an event A occurs m times, then the probability of A is defined as the limit of relative frequency $\frac{m}{n}$ as n tends to infinity

$$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

13. What is the axiomatic definition of probability?

Ans. Let S be the sample space with sample points $E_1, E_2, \ldots, E_i, \ldots E_n$. Each sample point is assigned a real number like $P(E_1) \ldots P(E_2) \ldots P(E_i) \ldots P(E_n)$, which are called probabilities of E_i, that must satisfy the following axioms.

(a) $0 \leq P(E_i) \leq 1$

(b) $P(S) = 1$

(c) If A and B are mutually exclusive events,

then, $P(A \cup B) = P(A) + P(B)$

14. Define subjective approach of probability.

Ans. In this approach probability is based on the degree of confidence or belief that a reasonable person has in the occurrence of event A.

15. Define Probability of an event.

Ans. $P(A) = \dfrac{\text{Number of sample points in A}}{\text{Number of sample points in S}}$

i.e. $P(A) = \dfrac{n(A)}{n(S)}$

16. Define collectively exhaustive events.

Ans. Two mutually exclusive events are said to be collectively exhaustive events if their union constitute the entire sample space.

17. What is conditional probability?

Ans. If A and B are two events in a sample space S then conditional probability of event A, given that event B has occurred, denoted by P(A/B) is defined as

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \; P(B) \neq 0$$

18. What is the difference between independent and dependent events?

Ans. Two events A and B are said to be independent if the occurrence of one does not effect the other.

Mathematically,

Two events A and B are independent iff

(a) $P(A/B) = P(A)$

(b) $P(B/A) = P(B)$

(c) $P(A \cap B) = P(A) \cdot P(B)$

Otherwise events are dependent.

19. Define addition law for not mutually exclusive events.

Ans. If A and B are two not mutually exclusive events, then the probability that at least one of them occurs is given by sum of separate probabilities of event A and B minus the probability of the joint event $A \cap B$.

i.e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

20. Define addition law for mutually exclusive events.

Ans. If A and B are mutually exclusive events.

then $P(A \cup B) = P(A) + P(B)$

CHAPTER # 8

RANDOM VARIABLES

8.1 RANDOM VARIABLES

An experiment is a well define action of obtaining an observation. Experiments in which the outcomes vary from trial to trial are called random experiment. The individual outcomes of a random experiment are said to be sample points. The set of all possible outcomes of a random experiment is called sample space.

Generally, we are not interested associated with each sample point, but only some number associated with these points. For example, when a coin is tossed two times, there are four possible outcomes, the sample space of the experiment is

$S = \{HH, HT, TH, TT\}$

If we are interested in the number of head appears, then a numerical value of 0, 1, 2 will be assigned to each sample point. Such a numerical quantity whose value is determined by the outcome of a random experiment, is called a random variable. In other words we can say A random variable is a function which assigns a numerical value to each simple event in a sample space.

A random variable is also called a chance variable, a stochastic variable and is abbreviated as r.v. The random variables are usually denoted by capital letters such as X, Y. A random variable can be either discrete or continuous.

8.2 DISCRETE RANDOM VARIABLE AND ITS PROBABILITY DISTRIBUTION

A random variable x is said to be discrete if it can assume only a limited number of values, which can be listed. For example, the number of heads obtained in coin tossing experiments, the number of fatal accidents etc.

Let X be a discrete r.v. taking value $x_1, x_2, \ldots, x_n$. Then the function, denoted by P(x) and defined by

$P(x) = f(x) = P(X = x_i)$

For $i = 1, 2, \ldots, n$ is called the probability function of the r.v.x.

A probability distribution must satisfy the following two properties.

(i) $f(x_i) \geq 0$

(ii) $\Sigma f(x_i) = 1$

EXAMPLE 8.1

Find the probability distribution for the number of heads, when 3 coins are tossed. Also draw probability Histogram and line graph.

SOLUTION

Total number of outcomes $= 2^3 = 8$

S = {HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}

Let X be the r.v. that denotes the number of heads. Then x can take values 0, 1, 2 and 3 and their probabilities are

$$f(0) = P(X = 0) = P[\{TTT\}] = \frac{1}{8}$$

$$f(1) = P(X = 1) = P[\{HTT, THT, TTH\}] = \frac{3}{8}$$

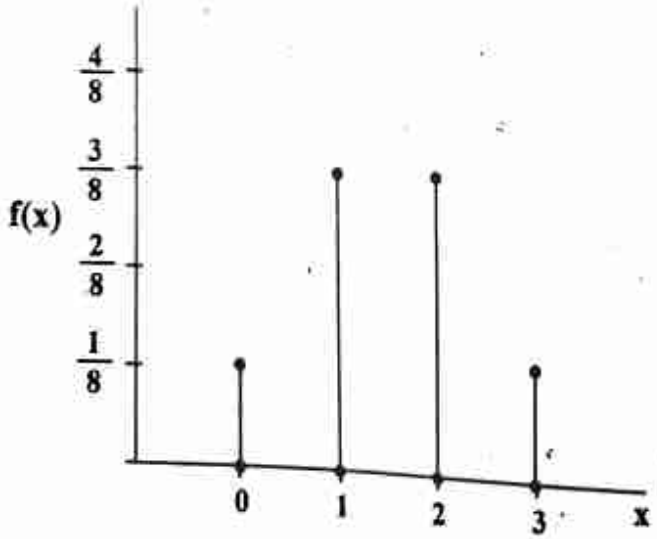
$$f(2) = P(X = 2) = P[\{HHT, HTH, THH\}] = \frac{3}{8}$$

$$f(3) = P(X = 3) = P[\{HHH\}] = \frac{1}{8}$$

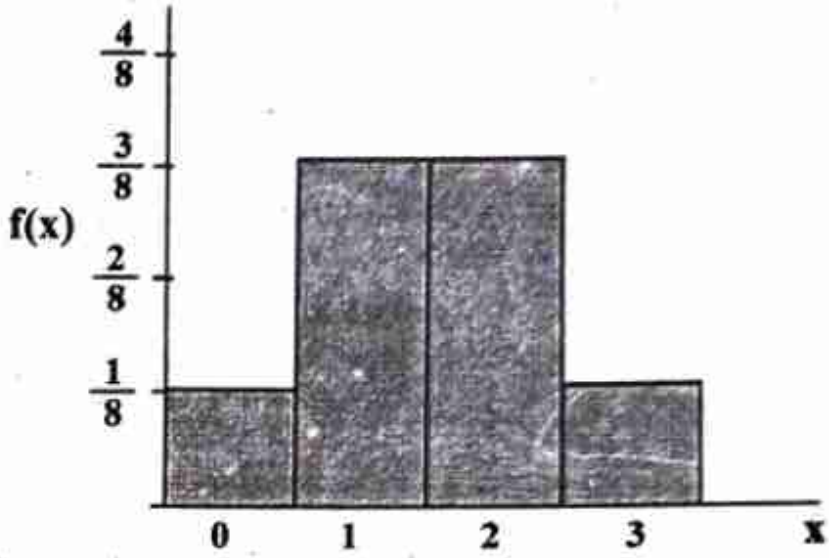
Probability distribution of No. of heads.

No. of Heads (x_i)	0	1	2	3
Probability $f(x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

(i) **Line Graph of Probability Distribution**



(ii) Histogram of Probability Distribution



EXAMPLE 8.2

Find the probability distribution of the sum of the dots when two fair dice are rolled.

SOLUTION

Total number of outcomes = $6^2 = 36$

The sample space of this experiment is

$$S = \left\{ \begin{matrix} (1,1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ (2,1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ (3,1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ (4,1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ (5,1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ (6,1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \end{matrix} \right\}$$

Let X be the random variable that denotes the sum of dots. Then x can take values 2, 3, 4, ..., 12.

$$f(2) = P(X = 2) = P[\{1, 1\}] = \frac{1}{36}$$

$$f(3) = P(X = 3) = P[\{(1, 2), (2, 1)\}] = \frac{2}{36}$$

$f(4) = P(X = 4) = P[\{(1, 3), (2, 2), (3, 1)\}] = \frac{3}{36}$

$f(5) = P(X = 5) = P[\{(1, 4), (2, 3), (3, 2), (4, 1)\}] = \frac{4}{36}$

$f(6) = P(X = 6) = P[\{(1,5), (2,4), (3,3), (4,2), (5,1)\}] = \frac{5}{36}$

$f(7) = P(X = 7) = P[\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}] = \frac{6}{36}$

$f(8) = P(X = 8) = P[\{(2,6), (3,5), (4,4), (5,3), (6,2)\}] = \frac{5}{36}$

$f(9) = P(X = 9) = P[\{(3, 6), (4, 5), (5, 4), (6, 3)\}] = \frac{4}{36}$

$f(10) = P(X = 10) = P[\{(4, 6), (5, 5), (6, 4)\}] = \frac{3}{36}$

$f(11) = P(X = 11) = P[\{(5, 6), (6, 5)\}] = \frac{2}{36}$

$f(12) = P(X = 12) = P[\{6, 6\}] = \frac{1}{36}$

Probability distribution of the r.v. X

x_i	2	3	4	5	6	7
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$
x_i	8	9	10	11	12	
$f(x_i)$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

EXAMPLE 8.3

Three balls are drawn from a bag containing 5 white, 3 black balls. If x denotes the number of white balls drawn. Find the probability distribution of x.

SOLUTION

Total No. of outcomes $= \binom{8}{3} = 56$

Let X be a.r.v that denotes the no. of white balls, then x can take values 0, 1, 2, 3.

$$f(0) = P(X = 0) = \frac{\binom{5}{0}\binom{3}{3}}{\binom{8}{3}} = \frac{1 \times 1}{56} = \frac{1}{56}$$

$$f(1) = P(X = 1) = \frac{\binom{5}{1}\binom{3}{2}}{\binom{8}{3}} = \frac{5 \times 3}{56} = \frac{15}{56}$$

$$f(2) = P(X = 2) = \frac{\binom{5}{2}\binom{3}{1}}{\binom{8}{3}} = \frac{10 \times 3}{56} = \frac{30}{56}$$

$$f(3) = P(X = 3) = \frac{\binom{5}{3}\binom{3}{0}}{\binom{8}{3}} = \frac{10 \times 1}{56} = \frac{10}{56}$$

Probability distribution of r.v.X

x	0	1	2	3
f(x)	$\frac{1}{56}$	$\frac{15}{56}$	$\frac{30}{56}$	$\frac{10}{56}$

8.3 CONTINUOUS RANDOM VARIABLE AND ITS PROBABILITY DENSITY FUNCTION

A random variable X is said to be continuous if it can assume any value within a given range. For example, weight of a person, the amount of rainfall, the temperature at a certain place etc.

The function f(x) is called probability density function abbreviated as p.d.f or simply density function of the r.v.x.

A function f(x) which satisfy the following conditions is called probability density function.

(i) $f(x) \geq 0$

(ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1$

EXAMPLE 8.4

A continuous random variable x can assume values between 0 and 2 have a density function.

$$f(x) = A(2 - x)(2 + x)$$

Find (i) The value of A

(ii) $P\left(x = \frac{1}{2}\right)$

(iii) $P(x \le 1)$

(iv) $P(x \ge 2)$

(v) $P(1 \le x \le 2)$

SOLUTION

(i) To find the value of A, we must have

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{0}^{2} A(2 - x)(2 + x)\,dx$$

$$A\int_{0}^{2} (4 - x^2)\,dx = 1$$

$$A\left|\frac{4x^1}{1} - \frac{x^3}{3}\right|_{0}^{2} = 1$$

$$A\left[\left\{4(2) - \frac{(2)^3}{3}\right\} - \left\{4(0) - \frac{(0)^3}{3}\right\}\right] = 1$$

$$A\left[\left(8 - \frac{8}{3}\right) - (0)\right] = 1$$

$$A\left[\frac{24 - 8}{3}\right] = 1$$

$$16A = 3$$

$$A = \frac{3}{16}$$

After substituting the value of A, function f(x) becomes

$$f(x) = \frac{3}{16}(2 - x)(2 + x)\ ,\ 0 \le x \le 2$$

(ii) $P\left(X = \frac{1}{2}\right) = 0$, (as probability at a particular value is 0)

(iii) $P(x \le 1) = \frac{3}{16}\int_0^1 (4 - x^2)\,dx$

$$= \frac{3}{16}\left[4x - \frac{x^3}{3}\right]_0^1$$

$$= \frac{3}{16}\left[4(1) - \frac{(1)^3}{3}\right] = \frac{3}{16}\left(4 - \frac{1}{3}\right)$$

$$= \frac{3}{16}\left[\frac{12 - 1}{3}\right]$$

$$= \frac{3}{16} \times \frac{11}{3} = \frac{11}{16}$$

(iv) $P(x \ge 2) = 0$, (as outside the range 0 to 2)

(v) $P(1 \le x \le 2) = \frac{3}{16}\int_1^2 (4 - x^2)\,dx$

$$= \frac{3}{16}\left[4x - \frac{x^3}{3}\right]_0^1$$

$$= \frac{3}{16}\left[\left\{4(2) - \frac{(2)^3}{3}\right\} - \left\{4(1) - \frac{(1)^3}{3}\right\}\right]$$

$$= \frac{3}{16}\left[\left\{8 - \frac{8}{3}\right\} - \left\{4 - \frac{1}{3}\right\}\right]$$

$$= \frac{3}{16}\left[\left\{\frac{24 - 8}{3}\right\} - \left\{\frac{12 - 1}{3}\right\}\right]$$

$$= \frac{3}{16}\left[\frac{16}{3} - \frac{11}{3}\right]$$

$$= \frac{3}{16}\left(\frac{16 - 11}{3}\right)$$

$$= \frac{3}{16} \times \frac{5}{3} = \frac{5}{16}$$

EXAMPLE 8.5

A continuous random variable x with P.d.f. is given below.

$f(x) = K(x + 30),\ 2 \le x \le 8$

Where K is constant, find

(i) Value of K

(ii) $P(3 \le x \le 6)$

(iii) $P(x \le 6)$

(iv) $x \ge 4)$

SOLUTION

(i) To find the value of K, we must have

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{2}^{8} K(x + 30) = 1$$

$$K\left|\frac{x^2}{2} + 30x\right|_{2}^{8} = 1$$

$$K\left[\left\{\frac{(8)^2}{2} + 30(8)\right\} - \left\{\frac{(2)^2}{2} + 30(2)\right\}\right] = 1$$

$$K\left[\left\{\frac{64}{2} + 240\right\} - \left\{\frac{4}{2} + 60\right\}\right] = 1$$

$$K\,[\{32 + 240\} - \{2 + 60\}] = 1$$

$$K\,[272 - 62] = 1$$

$$210\,K = 1$$

$$K = \frac{1}{210}$$

$$f(x) = \frac{1}{210}(x + 30) \qquad 2 \le x \le 8$$

(ii) $$P(3 \le x \le 6 = \int_3^6 \frac{1}{210}(x + 30)\,dx$$

$$= \frac{1}{210}\int_3^6 (x + 30)\,dx$$

$$= \frac{1}{210}\left|\frac{x^2}{2} + 30x\right|_3^6$$

$$= \frac{1}{210}\left[\left\{\frac{(6)^2}{2} + 30(6)\right\} - \left\{\frac{(3)^2}{2} + 30(3)\right\}\right]$$

$$= \frac{1}{210}\left[(18 + 180) - \left(\frac{9}{2} + 90\right)\right]$$

$$= \frac{1}{210}\left[198 - \left(\frac{9 + 180}{2}\right)\right]$$

$$= \frac{1}{210}\left[198 - \frac{189}{2}\right]$$

$$= \frac{1}{210}\left[\frac{396 - 189}{2}\right]$$

$$= \frac{1}{210}\left[\frac{207}{2}\right] = \frac{207}{420}$$

(iii) $$P(x \le 6) = \int_2^6 \frac{1}{210}(x + 30)\,dx$$

$$= \frac{1}{210}\int_2^6 (x + 30)\,dx$$

$$= \frac{1}{210}\left|\frac{x^2}{2} + 30x\right|_2^6$$

$$= \frac{1}{210}\left[\left\{\frac{(6)^2}{2} + 30(6)\right\} - \left\{\frac{(2)^2}{2} + 30(2)\right\}\right]$$

$$= \frac{1}{210}\ [(18 + 180) - (2 + 60)]$$

$$= \frac{1}{210}\ [198 - 62]$$

$$= \frac{1}{210}\ (136) = \frac{136}{210} = \frac{68}{105}$$

(iv) $$P(x \geq 4) = \frac{1}{210} \int_4^8 (x + 30)\, dx$$

$$= \frac{1}{210} \left| \frac{x^2}{2} + 30x \right|_4^8$$

$$= \frac{1}{210} \left[\left\{ \frac{(8)^2}{2} + 30(8) \right\} - \left\{ \frac{(4)^2}{2} + 30(4) \right\} \right]$$

$$= \frac{1}{210}\ [(32 + 240) - (8 + 120)]$$

$$= \frac{1}{210}\ [272 - 128]$$

$$= \frac{1}{210}\ (144) = \frac{144}{210} = \frac{72}{105}$$

8.4 MATHEMATICAL EXPECTATION OF A RANDOM VARIABLE

Let X be a discrete random variable taking values $x_1, x_2, \ldots x_n$ with corresponding probabilities $f(x_1), f(x_2), \ldots, f(x_n)$. Then the expectation of a random variable X is denoted by E(X) and defined as

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots, + x_n f(x_n)$$

$$E(X) = \Sigma x_i f(x_i) \quad \text{(For discrete variable)}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{(For continuous variable)}$$

8.4.1 Properties of Expected Values

(i) If a is constant, then $E(a) = a$

(ii) $E(ax) = a\,E(x)$

(iii) $E(ax + c) = a\,E(x) + c$

(iv) $E(x \pm y) = E(x) \pm E(y)$

(v) $E(x\,y) = E(x) \,.\, E(y)$ (If x and y are independent)

EXAMPLE 8.6

What is the mathematical expectation of the number of heads when two coins are tossed.

SOLUTION

Total number of outcomes $= 2^2 = 4$

$S = \{HH, HT, TH, TT\}$, $n(s) = 4$

Let X be the r.v. that denotes the no. of heads, then x can take values 0, 1 and 2.

$f(0) = P(X = 0) = P[\{TT\}] = \frac{1}{4}$

$f(1) = P(X = 1) = P[\{HT, TH\}] = \frac{2}{4}$

$f(2) = P(X = 2) = P[\{HH\}] = \frac{1}{4}$

X	f(x)	xf(x)
0	$\frac{1}{4}$	0
1	$\frac{2}{4}$	$\frac{2}{4}$
2	$\frac{1}{4}$	$\frac{2}{4}$
Σ	1	$\frac{4}{4} = 1$

$E(x) = \Sigma x\, f(x) = \frac{4}{4} = 1$

EXAMPLE 8.7

If $f(x) = \frac{6 - |7 - x|}{36}$ for $x = 2, 3, \ldots, 12$, then find the mean and variance of the random variable x.

SOLUTION

Calculation of mean and variance

x	$f(x) = \frac{6 - \lvert 7 - x \rvert}{36}$	x f(x)	x^2 f(x)
2	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{4}{36}$
3	$\frac{2}{36}$	$\frac{6}{36}$	$\frac{18}{36}$
4	$\frac{3}{36}$	$\frac{12}{36}$	$\frac{48}{36}$
5	$\frac{4}{36}$	$\frac{20}{36}$	$\frac{100}{36}$
6	$\frac{5}{36}$	$\frac{30}{36}$	$\frac{180}{36}$
7	$\frac{6}{36}$	$\frac{42}{36}$	$\frac{294}{36}$
8	$\frac{5}{36}$	$\frac{40}{36}$	$\frac{320}{36}$
9	$\frac{4}{36}$	$\frac{36}{36}$	$\frac{324}{36}$
10	$\frac{3}{36}$	$\frac{30}{36}$	$\frac{300}{36}$
11	$\frac{2}{36}$	$\frac{22}{36}$	$\frac{242}{36}$
12	$\frac{1}{36}$	$\frac{12}{36}$	$\frac{144}{36}$
Σ	1	$\frac{252}{36} = 7$	$\frac{1974}{36}$

$$\text{Mean} = E(x) = \Sigma x\, f(x) = 7$$

$$\text{Var}(x) = E(x^2) - [E(x)]^2$$

$$= \Sigma x^2 f(x) - [E(x)]^2$$

$$= \frac{1974}{36} - 49$$

$$= \frac{1974 - 1764}{36} = \frac{210}{36} = 5.83$$

EXAMPLE 8.8

Construct a probability distribution based on the following frequency distribution.

Outcome	102	105	108	111	114	117
Frequency	10	20	45	15	20	15

Also compute the expected value of the outcome.

SOLUTION

Outcome (x)	Frequency	f(x)	x f(x)
102	10	$\frac{10}{125} = 0.08$	8.16
105	20	$\frac{20}{125} = 0.16$	16.80
108	45	$\frac{45}{125} = 0.36$	38.88
111	15	$\frac{15}{125} = 0.12$	13.32
114	20	$\frac{20}{125} = 0.16$	18.24
117	15	$\frac{15}{125} = 0.12$	14.04
Σ	125		109.44

Expected outcome = E(x) = Σx f(x) = 109.44

EXAMPLE 8.9

Given the following discrete probability distribution.

x	0	1	2	3	4	5
f(x)	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Compute mean, variance, S.D and C.V.

SOLUTION

x	f(x)	x f(x)	x^2 f(x)
0	$\frac{6}{36}$	0	0
1	$\frac{10}{36}$	$\frac{10}{36}$	$\frac{10}{36}$
2	$\frac{8}{36}$	$\frac{16}{36}$	$\frac{32}{36}$
3	$\frac{6}{36}$	$\frac{18}{36}$	$\frac{54}{36}$
4	$\frac{4}{36}$	$\frac{16}{36}$	$\frac{64}{36}$
5	$\frac{2}{36}$	$\frac{10}{36}$	$\frac{50}{360}$
Σ	1	$\frac{70}{36}$	$\frac{210}{36}$

$$\text{Mean} = E(x) = \Sigma x\, f(x) = \frac{70}{36} = 1.94$$

$$E(x^2) = \Sigma x^2 f(x) = \frac{210}{36} = 5.83$$

$$\text{Var}(x) = E(x^2) - [E(x)]^2$$

$$= 5.83 - (1.94)^2 = 2.07$$

$$\text{S.D.} = \sqrt{\text{var}(x)} = 1.44$$

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100$$

$$= \frac{1.44}{1.94} \times 100 = 74.16\%$$

EXAMPLE 8.10

A variable x has P.d.f.

$f(x) = 20x^3(1 - x) \qquad 0 < x < 1$

Find mean and variance.

SOLUTION

$$\text{Mean} = E(x) \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$= 20 \int_0^1 x \,.\, x^3 (1 - x)\, dx$$

$$= 20 \int_0^1 (x^4 - x^5)\, dx$$

$$= 20 \left| \frac{x^5}{5} - \frac{x^6}{6} \right|_0^1$$

$$= 20\left[\frac{1}{5} - \frac{1}{6}\right] = \frac{20}{30} = \frac{2}{3}$$

$$E(x^2) = \int_{-\infty}^{\infty} x^2\, f(x)\, dx$$

$$= 20 \int_0^1 x^2 \,.\, x^3 (1 - x)\, dx$$

$$= 20 \int_0^1 (x^5 - x^6)\, dx$$

$$= 20 \left| \frac{x^6}{6} - \frac{x^7}{7} \right|_0^1$$

$$= 20\left[\frac{1}{6} - \frac{1}{7}\right]$$

$$= \frac{20}{42} = \frac{10}{21}$$

$$\text{Var}(x) = E(x^2) - [E(x)]^2$$

$$= \frac{10}{21} - \left(\frac{2}{3}\right)^2 \quad \Rightarrow \quad = \frac{10}{21} - \frac{4}{9} = \frac{2}{63}$$

EXAMPLE 8.11

If it rains, an umbrella saleman can earn Rs. 200 per day. If it is fair he can lose Rs. 50 per day. What is his expectation if the probability of rain is 0.4?

SOLUTION

Let x denote the number of rupees the saleman earns. Then x is a r.v. with possible values 200 and –50, with corresponding probabilities are 0.4 and 0.6 respectively.

$$E(x) = (200 \times 0.4) + (-50 \times 0.6)$$
$$= 80 - 30 = 50$$
$$= \text{Rs. } 50 \text{ per day}$$

EXAMPLE 8.12

A and B throw with one die for a prize of Rs. 55, which is to be won by the player who first throws 1. If A has the first throw, what are their expectations?

SOLUTION

The probability of throwing a 1 with one die is $\frac{1}{6}$ and that of not throwing a 1 is $\frac{5}{6}$.

As A has the first throw, therefore he can win in the first, third, fifth, ... throws.

$$P(A) = \frac{1}{6} + \left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right) + \ldots\ldots$$

$$a = \frac{1}{6}$$

$$r = \frac{\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)}{\left(\frac{1}{6}\right)} = \left(\frac{5}{6}\right)^2$$

$$S_\infty = \frac{a}{1-r}$$

$$= \frac{\frac{1}{6}}{1 - \frac{25}{36}} = \frac{\frac{1}{6}}{\frac{11}{36}}$$

$$= \frac{1}{6} \times \frac{36}{11} = \frac{6}{11}$$

The probability that B wins = 1 – P(A)

$$= 1 - \frac{6}{11} = \frac{5}{11}$$

Hence A's expectation $= \frac{6}{11} \times 55 =$ Rs. 30

B's expectation $= \frac{5}{11} \times 55 =$ Rs. 25

EXAMPLE 8.13

A random variable "x" has the following probability distribution.

x_i	0	1	2	3	4	5
$f(x_i)$	0.10	0.25	0.30	0.20	0.10	0.05

Find (i) E(x), $E(x^2)$ and var (x)

(ii) Construct probability distribution y = 3x + 2 then find E(y), $E(y^2)$ and var (y).

(iii) What is the relation between E(x) and E(y) and var (x) and var (y)

SOLUTION

(i)

x_i	$f(x_i)$	$x_i \cdot f(x_i)$	$x_i^2 f(x_i)$
0	0.10	0	0
1	0.25	0.25	0.25
2	0.30	0.60	1.20
3	0.20	0.60	1.80
4	0.10	0.40	1.60
5	0.05	0.25	1.25
Σ	1.00	2.10	6.10

$E(x) = \Sigma x\, f(x) = 2.10$

$E(x^2) = \Sigma x^2 f(x) = 6.10$

$Var(x) = E(x^2) - [E(x)]^2$

$= 6.10 - (2.10)^2 = 1.69$

(ii)

$y = 3x + 2$	$f(y)$	$y\,f(y)$	$y^2 f(y)$
2	0.10	0.20	0.40
5	0.25	1.25	6.25
8	0.30	2.40	19.20
11	0.20	2.20	24.20
14	0.10	1.40	19.60
17	0.05	0.85	14.45
Σ	**1.10**	**8.30**	**84.10**

$E(y) = \Sigma y\, f(y) = 8.30$

$E(y^2) = \Sigma\, y^2 f(y) = 84.10$

$var(y) = E(y^2) - [E(y)]^2$

$= 84.10 - (8.30)^2$

$= 15.21$

(iii) $y = 3x + 2$

$E(y) = 3E(x) + 2$

$= 3(2.10) + 2$

$= 6.30 + 2.0 = 8.30$

$y = 3x + 2$

$var\,(y) = 9\,var\,(x)$

$= 9(1.69) = 15.21$

SUMMARY

APPLICATION	FORMULA
Probability function	$f(x) = P(x = x_i)$ Where $i = 1, 2, 3 \ldots\ldots n$
Properties of probability distribution	(i) $f(x) \geq 0$ (ii) $\Sigma f(x) = 1$
Probability density function	A function is said to be Pdf if it satisfy the following conditions (i) $f(x) \geq 0$ (ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1$
Mathematical Expectation i.e. (Mean) Variance	$E(X) = \Sigma x f(x)$ Variance $= E(x^2) - [E(x)]^2$ Where, $E(x) = \Sigma x f(x)$ $E(x^2) = \Sigma x^2 f(x)$

EXERCISES

8.1 (a) What do you meant by Random Variable?

(b) A die is rolled. Let x denote the number of dots. Write down the possible outcomes and the values assigned to random variable. Find its probability distribution.

8.2 (a) Define discrete random variable and its probability distribution. Write down two basic properties of probability distribution.

(b) From an urn containing 4 red and 6 white marbles. A man draws three marbles at random with out replacement. Find the probability distribution of number of red marbles.

8.3 A committed size 5 is to be selected at random from 3 women and 5 men. Find the probability distribution of number of women on the committee.

8.4 Given the discrete probability distribution

$$f(x) = \binom{4}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x} \quad ; \; x = 0, 1, 2, 3, 4$$

(i) Find probability distribution.

(ii) Calculate mean and variance.

8.5. Three coins are tossed. Find the mean and variance of number of heads.

8.6 Two unbiased dice are rolled. Find the probability distribution of sum of dots. Also calculate its mean and variance.

8.7 Two dice are rolled. Find the expected value of the absolute difference of the number on two dice.

8.8 Construct a probability distribution based on the following frequency distribution.

Outcome	2	4	6	8	10	12	14
Frequency	24	22	16	12	7	3	1

Compute Mean and variance of the distribution.

8.9 A coin is tossed 4 times. If x denote the number of heads? What is the probability distribution of x. Find the mean and variance of number of heads.

8.10 Find the probability distribution of number of girls in families with three children assuming equal probabilities for boys and girls.

8.11 (a) Define expectation. Write down its properties.

(b) Let x be a random variable with probability distribution.

x	−1	0	1	2	3
f(x)	0.125	0.50	0.20	0.05	0.125

(i) Find E(x) and var(x).

(ii) Find the P.d of r.v. $y = 2x + 1$, using this P.d of y, determine E(y) and var(y)

(iii) How E(x) and E(y) and var(x) and var(y) related?

8.12 A and B throw with one die for a prize of Rs. 66, which is to be won by the player who first throws 3. If A has the first throw, what are their respective expectations?

8.13 A, B and C cut a pack of card successively in the order mentioned. If the person who cuts a Heart first, receives Rs. 175? What are their expectations?

8.14 A man draws 2 balls from a bag containing 3 white and 5 black balls. If he receives Rs. 70.00 for every white ball, he draws and Rs. 7.00 for every black ball. Find his expectations.

8.15 A fair die is rolled. If x is the random variable, the spots on the top, then obtain the probability distribution of x and hence obtained the expected value of the random variable x.

8.16 (a) Define continuous random variable. What do you mean by probability density function?

(b) A continuous random variable has a P.d.f given below.

$$f(x) = \frac{2(1 + x)}{27} \; ; \; 2 \le x \le 5$$

(i) Show that $P(2 \le x \le 5) = 1$

(ii) Find $P(x < 4)$

(iii) Find $P(3 \le x \le 4)$.

8.17 Given the following function

$$f(x) = A(4x - 2x^2) \qquad 0 \le x \le 2$$
$$= 0 \quad \text{Otherwise}$$

Calculate the value of A. So as f(x) may be probability density function. Also compute E(x) and Var (x).

8.18 A variable x has P.d.f.

$$f(x) = \frac{3}{8}(x - 2)^2 \ , \ 0 \le x \le 2$$
$$= 0 \quad \text{Otherwise}$$

Find expected value of x and its standard deviation.

CHAPTER # 9

PROBABILITY DISTRIBUTIONS

9.1 BINOMIAL PROBABILITY DISTRIBUTION

Many experiments consists of repeated independent trials, each trial having only two possible outcomes. For example head and tail, success and failure, alive and dead etc. If the probability of each outcome remains the same for all trials, then such trials are called Bernoulli *trials*. The experiment having n Bernoulli trials is called Binomial experiment. A binomial experiment possesses the following properties.

(i) The outcome of each trial may be classified into two categories i.e. success and failure.

(ii) The probability of success, denoted by p, remains constant for all trials.

(iii) The repeated trials are independent.

(iv) The experiment is repeated a fixed number of times.

Suppose we have n independent trials for which the probability of success is p and probability of failure is q, where $q = 1 - p$. Then the probability of exactly x successes is given by

$P(X = x) = \binom{n}{x} p^x q^{n-x}$, $x = 0, 1, 2, \ldots, n$, is called binomial probability distribution.

It has two parameters n and p. It was discovered by Swiss Mathematician Jakob Bernoulli.

EXAMPLE 9.1

A Coin is tossed 4 times. Find the probabilities of obtaining number of heads.

SOLUTION

$n = 4, \qquad p = \frac{1}{2}, \qquad q = 1 - p = \frac{1}{2}$

$P(X = x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, 2, 3, 4$

$$P(X = x) = \binom{4}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{4-x}$$

$$P(X = 0) = \binom{4}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 = 1 \times 1 \times \frac{1}{16} = \frac{1}{16}$$

$$P(X = 1) = \binom{4}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = 4 \times \frac{1}{2} \times \frac{1}{8} = \frac{4}{16}$$

$$P(X = 2) = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = 6 \times \frac{1}{4} \times \frac{1}{4} = \frac{6}{16}$$

$$P(X = 3) = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1 = 4 \times \frac{1}{8} \times \frac{1}{2} = \frac{4}{16}$$

$$P(X = 4) = \binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = 1 \times \frac{1}{16} \times 1 = \frac{1}{16}$$

EXAMPLE 9.2

If 20% of the bolt produced by a machine are defective, determine the probability that out of 4 bolts chosen at random (i) Zero (ii) 2 bolts are defective

SOLUTION

$$n = 4, \qquad p = 20\% = \frac{20}{100} = \frac{1}{5}, \quad q = \frac{4}{5}$$

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \; x = 0, 1, 2, \ldots, n, \qquad = \binom{4}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{4-x}$$

where x be the defective bolts

(i) $$P(X = 0) = \binom{4}{0}\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^4 = 1 \times 1 \times \frac{256}{625} = \frac{256}{625}$$

(ii) $$P(X = 2) = \binom{4}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^2 = 6 \times \frac{1}{25} \times \frac{16}{25} = \frac{96}{625}$$

9.1.1 Binomial Frequency Distribution

If the binomial probability distribution is multiplied by the number of experiments (N), then the distribution is called binomial frequency distribution. This the expected frequency of x successes in N experiment is $N.\binom{n}{x} p^x q^{n-x}$.

EXAMPLE 9.3

Out of 800 families with 5 children each. How many would you expect to have
(i) 3 boys (ii) at least 4 boys
(iii) atmost one day

SOLUTION

$$n = 5, \qquad N = 800, \qquad p = \frac{1}{2}, \qquad q = \frac{1}{2}$$

(i) $p(3 \text{ boys}) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = 10 \times \frac{1}{8} \times \frac{1}{4} = \frac{10}{32}$

Hence expected number of families with 3 boys $= 800 \times \frac{10}{32} = 250$

(ii) p(at least 4 boys).

$$p(x \geq 4) = p(x = 4) + p(x = 5)$$

$$= \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0$$

$$= (5 \times \frac{1}{16} \times \frac{1}{2}) + (1 \times \frac{1}{32} \times 1)$$

$$= \frac{5}{32} + \frac{1}{32} = \frac{6}{32}$$

Hence expected no. of families with at least 4 boys $= \frac{6}{32} \times 800 = 150$

(iii) p(atmost one boy)

$$p(x \leq 1) = p(x = 1) + p(x = 0)$$

$$= \binom{5}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^4 + \binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5$$

$$= (5 \times \frac{1}{2} \times \frac{1}{16}) + (1 \times 1 \times \frac{1}{32})$$

$$= \frac{5}{32} + \frac{1}{32} = \frac{6}{32}$$

Expected no. of families $= 800 \times \frac{6}{32} = 150$

9.1.2 Properties of the Binomial Distribution

(i) Mean = np

(ii) Variance = npq

(iii) The shape of the binomial distribution depends on the values of two parameters n and p. We observe that, when $p < \frac{1}{2}$, the distribution is positively skewed and when $p > \frac{1}{2}$, the distribution becomes negatively

skewed. When $p = \frac{1}{2}$, the distribution is always symmetrical. As n increases to ∞, $\beta_1 \to 0$ and $\beta 2 = 3$, then binomial probability distribution is symmetrical and mesokurtic.

EXAMPLE 9.4

In a binomial distribution n = 10 and p = 2/5. Find mean, variance and standard deviation of the distribution.

SOLUTION

Here $n = 10$, $p = \frac{2}{5}$, $q = 1 - p = \frac{3}{5}$

We know that

$$\text{Mean} = nP$$

$$= 10 \times \frac{2}{5} = 4$$

$$\text{Variance} = nPq$$

$$= 10 \times \frac{2}{5} \times \frac{3}{5} = \frac{60}{25} = 2.2$$

$$S.D = \sqrt{\text{Variance}} = \sqrt{2.2} = 1.4832$$

EXAMPLE 9.5

In a binomial distribution, mean and variance were found to be 3.20 and 1.152 respectively. Find n and p.

SOLUTION

$$\text{Mean} = np = 3.20 \text{ ..... (i)}$$

$$\text{Variance} = npq = 1.152 \text{ ..... (ii)}$$

Dividing equation (ii) by (i), we get

$$\frac{npq}{np} = \frac{1.152}{3.20}$$

$$q = 0.36$$

$$p = 1 - q = 1 - 0.36 = 0.64$$

Putting the value of p = 0.64 in eq (i)

$$np = 3.20$$

$$n(0.64) = 3.20$$

$$n = \frac{3.20}{0.64} \quad \Rightarrow n = 5$$

EXAMPLE 9.6

Is it possible to have a binomial distrubiton with mean = 5 and S.D = 4

SOLUTION

We know that

$$\text{Mean} = nP = 5$$

$$S.D = \sqrt{nPq} = 4$$

$$\text{Variance} = nPq = 16$$

$$\frac{nPq}{nP} = \frac{16}{5} \quad \Rightarrow q = 3.2$$

Since p ro q can not be greater than one. So it is not possible to have a binomial distribution.

9.2 HYPERGEOMETRIC EXPERIMENT

When the successive trials are dependent and the probability of success does not remain constant for all trials. Such experiments are called hypergeometric experiments. Hypergeometric experiment has the following properties.

(i) The out come of an experiment can be classified into to categories, success or failure.

(ii) The successive trials are dependent

(iii) The probability of success changes from each trial.

(iv) The experiment is repeated a fixed number of times.

9.2.1 Hypergeometric Probability Distribution

The number of successes, x in a hypergeometric experimetn is called hypergeometric random variable and its probability distribution is called hypergeometric distribution. It is given as

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \; ; \; x = 0,1,2,3, \ldots\ldots n \text{ or } k \text{ (whichever is less)}$$

The hypergeometric distribution has 3 parameters n,k and N.

9.2.2 Properties of Hypergeometric Distribution

1. $\text{Mean} = \frac{nk}{N}$

2. $\text{Variance} = \frac{nk}{N}\left(\frac{N-k}{N}\right)\left(\frac{N-N}{N-1}\right)$

3. If N becomes very large, hypergeoemtric distribution tends to binomial probability distribution.

EXAMPLE 9.7

An urn contains 4 white and 7 black balls. A sample of 5 balls is selected at random. Determine the probability distribution of white balls.

SOLUTION

$N = 4 + 7 = 11,\ n = 5$

Taking white ball as success i.e $k = 4$

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}\ ;\ x = 0,1,2,3,4,5$$

$$P(X = x) = \frac{\binom{4}{x}\binom{7}{5-x}}{\binom{11}{5}}$$

$$P(X = 0) = \frac{\binom{4}{0}\binom{7}{5}}{\binom{11}{5}} = \frac{1 \times 21}{462} = \frac{21}{462}$$

$$P(X = 1) = \frac{\binom{4}{2}\binom{7}{3}}{\binom{11}{5}} = \frac{4 \times 35}{462} = \frac{140}{462}$$

$$P(X = 2) = \frac{\binom{4}{2}\binom{7}{4}}{\binom{11}{5}} = \frac{6 \times 35}{462} = \frac{210}{462}$$

$$P(X = 3)\ \frac{\binom{4}{3}\binom{7}{2}}{\binom{11}{5}} = \frac{4 \times 21}{462} = \frac{84}{462}$$

$$P(X = 4)\ \frac{\binom{4}{4}\binom{7}{1}}{\binom{11}{5}} = \frac{1 \times 7}{462} = \frac{7}{462}$$

Hence hypergeometric probability distribution is given by

X	0	1	2	3	4
P(x)	21/462	140/462	210/462	84/462	7/462

9.3 THE POISSON DISTRIBUTION

Experiments yielding observations of rare events i.e. events having a very small chance of occurrence when the experiment is repeated a very large number of times during given unit of time or region are called Poisson experiments. Such experiments might be the number of typing errors on a page, number of deaths by horse kicks per year, number of traffic accidents at a busy crossing per month; etc.

Poisson distribution is the limiting form of the binomial distribution when n is very large and p, the probability of success is very small.

The probability distribution of Poisson random variable is called Poisson probability distribution. Its function is given as

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}; \qquad x = 0,1,2,\ldots\ldots \infty.$$

Where e = 2.718

It has only one parameter u.

9.3.1 Properties of Poisson Distribution

1. It is a discrete probability distribution
2. Its range is 0 to ∞
3. Mean and variance are equal

EXAMPLE 9.8

A clerk makes 2 errors per page on the average. What is the probability that on the next page he makes (i) no error (ii) 3 errors (iii) 4 or more errors.

SOLUTION

Let X be the r.v, that denotes of the number of errors per page. Then x has a Poisson distribution with parameter u = 2, then

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$$

$$P(X = x) = \frac{e^{-2}(2)^x}{x!}, \text{ where } e = 2.718$$

Now (i) $P(x = 0) = \frac{e^{-2}(2)^0}{0!} = 0.1353$

(ii) $P(x = 3) = \frac{e^{-2}(2)^3}{3!} = \frac{0.1353 \times 8}{6} = 0.1805$

(iii) $P(x \geq 4) = 1 - [P(3) + P(2) + P(1) + (0)$

$$= 1 - \left[\frac{e^{-2}(2)^3}{3!} + \frac{e^{-2}(2)^2}{2!} + \frac{e^{-2}(2)^1}{1!} + \frac{e^{-2}(2)^0}{0!}\right]$$

$$= 1 - [0.1805 + 0.2786 + 0.2786 + 0.1353]$$

$$= 1 - 0.8651 \Rightarrow \quad 0.1349$$

9.4 NEGATIVE BINOMIAL DISTRIBUTION

In the binomial experiments the number of successes varies and number of trials is fixed. But there are experiments in which the number of successes is fixed and the number of trials varies to produce the fixed number of successes. Such experiments are called negative binomial distribution.

Let the random variable x denote the number of trials untill k successes occur then negative binomial distribution is given by

$$P(X = x) = \binom{x-1}{k-1} p^k q^{x-k} \quad x = k, k+1, k+2, \ldots$$

9.4.1 Properties of Negative Binomial Distribution

i. It is a discrete distribution.

ii. It has two parameter k and p

iii. Its mean is $\frac{kq}{p}$

iv. Its variance is $\frac{kq}{p^2}$

EXAMPLE 9.9

The probability that a person will install a black telephone in a residence is estimated to be 0.3. Find the probability that the 10th phone installed in a division is the 5th black phone.

SOLUTION

Here $P = 0.3$, $q = 1-P = 0.7$

$X = 5$ 5th success and $x = 10$ (10th telephone)

We know that

$$P(X = x) = \binom{x-1}{k-1} p^k q^{x-k}$$

$$P(X = 10) = \binom{10-1}{5-1} (0.3)^5 (0.7)^5$$

$$= \binom{9}{4} (0.3)^5 (0.7)^5$$

$$= 126 \times 0.00243 \times 0.16807$$

$$= 0.0515$$

9.5 GEOMETRIC PROBABILITY DISTRIBUTION

When an experiment consists of repeated independent trials with probability P of success and the trials are repeated until the first success occurs, is called geometric experiment.

A random variable x is defined to have geometric distribution if its density is given by

$$P(X = x) = q^{x-1}p, \quad x = 1,2,3,\text{-----},\infty.$$

It has only one parameter p. Geometric distribution is a special case of negative binomial distribution when k = 1

9.5.1 Properties of Geometric distribution

(i) It is a discrete distribution.

(ii) It has only one parameter i.e. p.

(iii) Its mean is $\frac{1}{p}$.

(iv) Its variance is $\frac{q}{p^2}$.

(v) It is a positively skewed.

EXAMPLE 9.10

Tossing an unbiased coin, determine the probability that the first head occurs on the third trial.

SOLUTION

Here $P = \frac{1}{2}$, the head occurs on the third trial i.e the first success occurs on the 3rd trial, so it is a geometric distribution with x = 3

$$P(X = x)\ Pq^{x-1}$$

$$P(X = 3) = \frac{1}{2}\left(\frac{1}{2}\right)^{3-1}$$

$$= \frac{1}{2}\left(\frac{1}{2}\right)^{2}$$

$$= \frac{1}{2} \times \frac{1}{4}$$

$$= \frac{1}{8}$$

9.6 NORMAL DISTRIBUTION

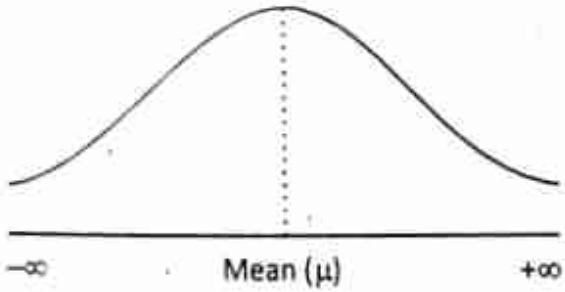
The normal distribution is the most important of the continuous probability distribution. It was first discovered by Abraham DeMoivre as a limiting form of binomial distribution. The name of P.S. Laplae is also associated with the derivation of normal distribution. German mathematician Carl. F.Gauss derived its equation mathematically as a probability distribution. Thus the normal distribution is often called the Gaussian distribution. It was Karl Pearson who is 1893 called its normal distribution and is best known by this name today.

A normal distribution is defined by the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < X < +\infty \text{ and } \sigma > 0.$$

where μ is the mean, σ is the standard deviation and $\pi(=3.1416)$ and $e(=2.7183)$ are constants. Obviously a normal distribution is characterised by two parameters μ and σ, its mean and standard deviation.

The behaviour of many of the real-life situations can be modelled as normal distribution. Some examples which follow normal distribution are monthly salary of employees in a locality, marks of students in an entrance test, height of employees in a company, internal diameter of bearings produced in a factory etc.



9.6.1 Standard Normal Distribution

The variable $Z = \frac{X-\mu}{\sigma}$ which measures the deviations of x from mean in standard deviation unit is called standard normal variable. It has mean 0 and standard deviation 1.

If the observations of real-life problems follow the normal distribution with mean μ and variance σ^2, then its random variable can by converted into a standard normal variable using the following transformation with mean 0 and standard deviation 1.

$$Z = \frac{X-\mu}{\sigma}$$

The P.d. of z, denoted by $\phi(z)$, becomes

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{for } -\infty < z < \infty$$

9.6.2 Properties of Normal Destruction

The main properties of the normal distribution are given below.

(i) It is continuous distribution ranges from $-\infty$ to ∞.

(ii) It has two parameter μ and σ.

(iii) Its mean is μ and variance is σ^2.

(iv) Its graph is bell shaped and unimodel.

(v) Total area under the normal curve is unity.

(vi) It is symmetrical distribution therefore,
Mean = Median = Mode

(vii) Mean deviation of the normal distribution is equal to $\frac{4}{5}$ of its standard deviation i.e M.D = $\frac{4}{5}\sigma$,

(viii) The two tails of the normal distribution extend indefinitely and never touch the horizontal axis.

(ix) In normal distribution, the limits $\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$ contains respectively 68.26%, 95.45% and 99.73% values /area

(x) The normal curve has point of inflection at $\mu - \sigma$ and $\mu + \sigma$.

(xi) In a normal distribution
$Q_1 = \mu - 0.6745\,\sigma$ and $Q_3 = \mu + 0.6745\,\sigma$

EXAMPLE 9.11

In a survey with a sample of 200 peoples, the monthly income follows normal distribution with mean and standard deviation Rs.10,000 and Rs.2000 respectively.

(a) What is the probability that the monthly income is less that Rs.11,000? Also find the number of persons having income less than Rs.11,000.

(b) What is the probability that the monthly income is more than Rs.12,000?

(c) Find the probability that the monthly income is in between Rs.7,000 and Rs.11,200? Also find the number of persons having income in between Rs.7,000 and 11,200.

SOLUTION

Let x be a random variable representing the monthly income of persons.

$\mu = 10{,}000$ and $\sigma = 2{,}000$

To find the desired probabilities, we transform x values to z values, using

$$Z = \frac{X-1000}{2000}$$

(a) $P(x \le 11{,}000) = ?$

At x = 11,000, we get $Z = \frac{11000 - 10000}{2000}$

$Z = 0.5$

$P(x \le 11000) = P(z \le 0.5)$

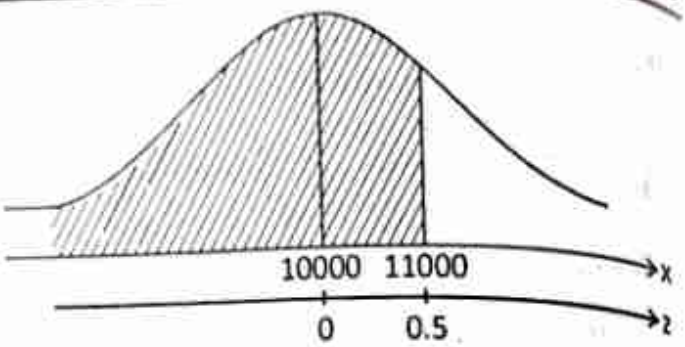
Hence using area table, we get

$= 0.5 + P(0 \le z \le 0.5)$

$= 0.5 + 0.1915$

$= 0.6915$

Number of persons having income less that Rs.11000 = 200 × 0.6915 = 138.3 ≈ 138



(b) $P(x \ge 12000) = ?$

At x = 12000, then

$$Z = \frac{12000 - 10000}{2000}$$

$Z = 1.0$

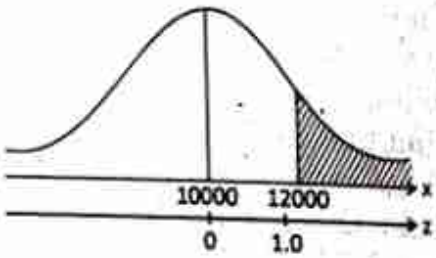
$P(x \ge 12000) = P(Z \ge 1.0)$

From area table, we have

$= 0.5 - P(0 \le Z \le 1.0)$

$= 0.5 - 0.3413$

$= 0.1587$



(c) $P(7000 \le x \le 11200) = ?$

At X = 7000, $Z = \frac{7000 - 10000}{2000}$

$Z = -1.5$

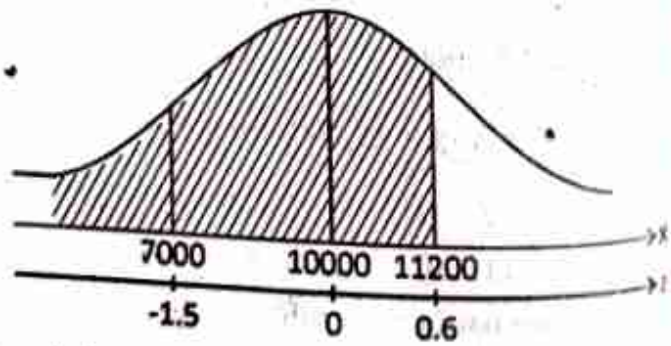
for x = 11200, $Z = \frac{11200 - 10000}{2000}$ $\Rightarrow Z = 0.6$

$P(7000 \le x \le 11200) = P(-1.5\ z \le 0.6)$

$= P(-1.5 \le z \le 0) + P(0 \le z \le 0.6)$

$= 0.4332 + 0.2257$

$= 0.6589$



Number of persons having income in between Rs.7000 and Rs.11,200 = 200 × 0.6589 = 131.8 ≈ 132

EXAMPLE 9.12

In a normal distribution with $\mu = 13.5$ and $\sigma = 3.6$, find two points such that a single observation has 95% chance of falling between them.

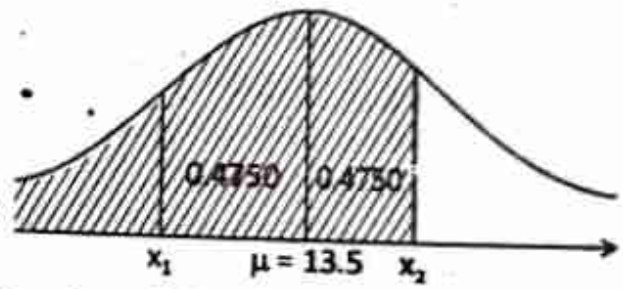
SOLUTION

Let x_1 and x_2 be the two points between which the probability of an observation falling is 0.95. As the curve is symmetric, so half of 0.95 i.e 0.4750 is the area lying on either side of μ.

Using normal table inversly, we find

$$(z/p = 0.4750) = 1.96$$

The point x_1 is on the left side of μ, so



$z = -1.96$ and x_2 which is to the right of μ, $z = 1.96$, therefore,

$$x_1 = \mu + z\sigma = 13.5 + (-1.96)(3.6) = 6.4$$

$$x_2 = \mu + z\sigma = 13.5 + 1.96(3.6) = 20.6$$

SUMMARY

Probability distributions with mean, variance and their parameters are given below

Sr. No.	Prob. Distribution	Prob. density function	Parameters	Mean	Variance
1.	Binomial distribution (discrete)	$P(X = x) = \binom{n}{x} p^x q^{n-x}$ $x = 0,1,2,.....n$	n,P	nP	nPq
2.	Hypergeometric distribution (discrete)	$P(X = x) \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$ $x = 0,1,2,.....n$ or K	n,K,N	$\frac{nK}{N}$	$\frac{nK}{N}\left(\frac{N-k}{N}\right)\left(\frac{N-n}{N-1}\right)$
3.	Poisson distribution (discrete)	$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$ $x = 0,1,2,.....\infty$	μ	μ	μ
4.	Negative Binomial distribution (discrete)	$P(X = x) = \binom{x-1}{k-1} p^k q^{x-k}$, $x = k, k+1, k+2,.....$	K,P	$\frac{Kq}{P}$	$\frac{Kq}{P^2}$
5.	Geometric distribution (discreet)	$P(X = x) = Pq^{x-1}$ $x = 1,2,3,.....\infty$	P	$\frac{1}{P}$	$\frac{q}{P^2}$
6.	Normal distribution (continuous)	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $-\infty < X < \infty$	μ and σ	μ	σ^2
7.	Standardized normal (continuous)	$Z = \frac{X-\mu}{\sigma}$	μ	0	1

EXERCISES

9.1 (a) What is meant by binomial experiment and what are its properties?

(b) A coin is tossed 4 times. Find the probabilities of obtaining various number of heads.

9.2 (a) Let x have a binomial distribution with n = 4 and $P = \frac{1}{3}$, find P(x = 1), $P(x = \frac{3}{2})$, P(x = 3), P(x = 6) and $P(x \leq 2)$.

(b) A coin is tossed 10 times; what is the probability of having (i) 0 head (ii) 3 heads (iii) atmost 2 heads (iv) at least 3 heads?

9.3 (a) The incidence of an occupational disease in an industry is such that the workers have 20% chance of suffering from it. What is the probability that out of 6 workers.

(i) Not more than 2 will catch the disease

(ii) 4 or more will catch the disease

(b) Suppose that 70% adults with allergies report relief with a specific medication. If the medication is given to 7 new patients with allergies, what is the probability that it is effective in exactly four?

9.4 (a) The experience of a house agent indicates that he can provide suitable accommodation for75% of the clients who come to him. If on a particular occasion 6 clients approach him. Calculate the probability that (i) less than 4 clients (ii) exactly 4 clients (iii) at least 5 clients, will get satisfactory accommodation.

(b) Based on past experience, that chief of organization has estimated that the probability of completing each project in time is 0.7. The company is planning to execute 12 such projects. Find the probability of completing (i) no project (ii) four projects in time (iii) atmost 3 projects in time.

9.5. (a) In a binomial distribution n = 20 and P = 3/5. Find the mean and variance.

(b) In a binomial distribution, mean and standard deviation were found to be 38 and 5.6 respectively, find the parameters n and p.

9.6 Five dice are rolled 96 times. Find the expected frequencies when throwing of a 4,5 or 6 is regarded as success.

9.7 (a) What is hypergeometric experiment and what are its properties?

(b) Three balls are drawn from a bag containing 5 white and 3 black balls. If X denote the number of white balls drawn from the bag, then find the probability distribution.

9.8 (a) A random variable X follows hypergeometric distribution having n = 5, N = 12 and k = 3 find (i) $P(x \le 1)$ (ii) $P(x > 1)$

(b) There are seven people who work in an office. Of the seven, four would like to be transferred. If three people from this office are randomly selected for transfer, what is the probability that

(i) All three will want to be transferred?

(ii) Atmost one will want to be transferred?

9.9 A bag contains 4 red balls and 6 black balls. A sample of 4 balls is selected from the bag. Let X be the number of red balls contained in a sample, then find the probability distribution of x.

9.10 A committee of size 3 is selected from 4 men and 2 women. Find the probability distribution for the number of men on the committee.

9.11 (a) Define the Poisson distribution.

(b) Given $\mu = 4.2$, for a Poisson distribution, find (i) $P(x \le 2)$ (ii) $P(x \ge 5)$ (iii) $P(x = 8)$

9.12 (a) A secretary makes 2 errors per page. What is the probability that on the next page she makes (i) 4or more errors (ii) no error.

(b) The arrival rate of customers arriving at a bank counter follows Poisson distribution with mean arrival 4 per 10 minutes interval. Find the probability that

(i) Exactly no customer will arrive in 10 minutes.

(ii) Exactly two customers will arrive in 10 minutes interval.

(iii) Atmost two customers will arrive in 10 minutes.

(iv) At least three customers will arrive in 10 minutes interval.

9.13 The quality control assistant takes a sample of 25 units per hour interval at a particular workstation of a production line and inspects them one by one. Based on the past experience, he has estimated that the probability of one unit will be defective is 0.04. Find the probability that

(i) No piece in the sample is defective

(ii) 3 pieces in the sample will be defective

(iii) Atmost 2 pieces will be defective

(iv) At least 3 pieces will be defective

9.14 Suppose that the number of insurance claims closely approximates a Poisson distribution with $\mu = 0.05$. Find the probability of (i) no claim and (ii) 1 or fewer claims.

9.15 If 3 percent of the electric blubs manufactured by a company are defective. Find the probability that in a sample of 100 bulbs

(i) Exactly 5 will be defective

(ii) Between 1 and 3 inclusive are defective

(iii) Atmost 2 bulbs are defective.

9.16 (a) What is negative binomial distribution?

(b) A person throws a pair of dice. What is the probability that he will get a total of 7 on the 3rd time on the seventh throw?

9.17 (a) The probability that a swimmer will succeed across the lake is 0.5. What is the probability that the eighth swimmer is the fourth one to cross a lake?

(b) A person throw a die. What is the probability that he will get number 6 on the 4th time on the 9th throw?

9.18 (a) Define geometric distribution.

(b) When throwing a die, determine the probability that number 1 occurs on the 6th times.

(c) Tossing an unbiased coin, determine the probability that first tail occurs on the second trial.

9.19 (a) Define the normal distribution. Also write down its properties.

(b) A random variable x is normally distributed with mean = 50 and variance = 25. Find the probability that (i) larger than 54 (ii) smaller than 57 (iii) it will fall between 0 and 40 (iv) 55 and 100.

9.20 (a) The achievement score for a college entrance examination are normally distributed with a mean of 75 and standard deviation of 10. What fractions of scores would you expect to lie between 70 and 90?

(b) The wrist circumference of adult males is normally distributed with a mean of $\mu = 6.85''$ and a standard deviation of $\sigma = 0.40''$. One man is selected at random. Determine the probabilities for the following possible wrist sizes

(i) Less than 8.00"

(ii) Between 5.75" and 6.75"

9.21 The average rain fall in Islamabad for the month of March is 9.22 cm. Assuming a normal distribution with standard deviation of 2.83 cm. Find the probability that next March Isalamabad receives (i) less than 8.00 cm of rain (ii) more than 5.00 cm but not over 7.00 cm rain (iii) more than 13.8 cm.

9.22 A training program has been designed to upgrade the supervisory skills of production supervisors. A study of past participant, indicate, that the mean length of time spent on program is 500 hours and that this normally distributed random variable has a standard derivation of 100 hours. Find the probability that

(i) Participant will require more than 500 hours.

(ii) Candidate will take between 500 and 650 hours.

(iii) Participant will take more than 700 hours.

(iv) Candidate selected will require fewer than 580 hours to complete the training.

9.23 If X ~ N (60,100), where X indicates marks obtained by student. Find the probability that a student selected at random obtain marks,

(i) Less than 56

(ii) More than 50

9.24 The height of boys at college follow a normal distribution with $\mu = 150.3$ cm and $\sigma = 5.0$ cm. Find the probability that the selected student has age

(i) Less than 153 cm

(ii) More than 145 cm

(iii) Between 146 cm and 152 cm

9.25 (a) In a normal distribution $\mu = 30$ and $\sigma = 5$. Find two points containing middle 95% of area.

(b) A random variable is normally distributed with mean 500 and standard deviation 100. What is the 95th percentile of the distribution.

(c) The heights of applicants to the police force are normally distributed with mean 170 cm and standard deviation 3.8 cm. If 30% of applicants are rejected because they are two small, what is the minimum acceptable height for the police force?

9.26 If X ~ N (60,100) find (i) a point that has 15% area below it (ii) a point that has 28% area above it.

9.27 **SELECT THE CORRECT ANSWER**

(i) **The mean of binomial distribution is:**

(a) nP (b) nPq

(c) nq (d) $\sqrt{nPq}$

(ii) **The binomial distribution has ..... parameters**

(a) 1 (b) 2

(c) 3 (d) 4

(iii) **The standard deviation of binomial distribution is:**

(a) $\sqrt{nP}$ (b) $\sqrt{nq}$

(c) Pq (d) $\sqrt{nPq}$

(iv) **Binomial distribution is symmetrical if:**

(a) $P = 1$ (b) $P = \frac{1}{2}$

(c) $P > \frac{1}{2}$ (d) $P < \frac{1}{2}$

(v) **If n = 20, P = 0.6, then variance of binomial distribution is equal to:**

(a) Zero (b) 1

(c) 4.8 (d) 12

(vi) **For a binomial distribution with parameters n and P, mean and variance are related as:**

(a) Mean = Variance (b) Mean > Variance

(c) Mean < Variance (d) None of these

(vii) **The binomial distribution is positively Skewed when:**

(a) $P = q$ (b) $P > q$

(c) $P < q$ (d) None of these

(viii) **The hypergeometric distribution has ..... parameters**

(a) 1 (b) 2

(c) 3 (d) 4

(ix) **The mean of hypergeometric distribution is:**

(a) $\frac{nN}{K}$ (b) $\frac{n+K}{N}$

(c) $\frac{nK}{N}$ (d) $\frac{Nk}{N}$

(x) **In a hypergeometric distribution N = 6, n = 4 and k = 3, then the mean is equal to:**

(a) 2 (b) 4

(c) 6 (d) 24

(xi) **The Poisson distribution has parameters:**

(a) 1 (b) 2

(c) 3 (d) 4

(xii) For a Poisson distribution, the mean and variance are related by:

(a) Mean = Variance (b) Mean > Variance

(c) Mean < Variance (d) None of above

(xiii) Negative binomial distribution has parameters:

(a) 1 (b) 2

(c) 3 (d) 4

(xiv) Mean of the negative binomial distribution is:

(a) $\frac{Kq}{P}$ (b) $\frac{nq}{P}$

(c) $\frac{Kq}{P^2}$ (d) $\frac{Pq}{K}$

(xv) Geometric distribution has parameters:

(a) 1 (b) 2

(c) 3 (d) 4

(xvi) Variance of Geometric distribution is:

(a) $\frac{1}{P}$ (b) $\frac{q}{P^2}$

(c) $\frac{1}{q}$ (d) $\frac{Pq}{n}$

(xvii) Geometric distribution is:

(a) Symmetrical (b) Negatively Skewed

(c) Positively Skewed (d) None of these

(xviii) Normal distribution has ..... parameters.

(a) 1 (b) 2

(c) 3 (d) 4

(xix) The mean of the standard normal distribution is:

(a) 0 (b) 1

(c) 100 (d) ∞

(xx) In a normal distribution mean deviation is:

(a) $\frac{3}{4}\sigma$ (b) $\frac{4}{5}\sigma$

(c) $\frac{3}{5}\sigma$ (d) $\frac{2}{5}\sigma$

(xxi) If X ~ N (40,25), the median is:

(a) 15 (b) 25

(c) 40 (d) 1000

(xxii) The range of the normal distribution is:

(a) 0 to ∞ (b) 0 to n

(c) – ∞ to ∞ (d) –∞ to 0

(xxiii) In a standard normal distribution, variance is:

(a) 0 (b) 1

(c) 2 (d) 10

(xxiv) Total area under the normal curve is:

(a) –1 (b) 0

(c) 1 (d) 100

(xxv) In normal distribution if mode = 20, then the value of mean is:

(a) 10 (b) 20

(c) 30 (d) 40

(xxvi) Normal distribution has parameters:

(a) μ (b) μ, σ^2

(c) σ (d) μ, σ, P

(xxvii) In a normal distribution mean, median and mode are:

(a) Positive (b) Negative

(c) Equal (d) Not equal

(xxviii) Normal distribution is a:

(a) Unimodal (b) Bimodal

(c) Trimodal (d) None of these

Answers 9.27

(i)	(a)	(ii)	(b)	(iii)	(d)	(iv)	(b)	(v)	(c)
(vi)	(b)	(vii)	(c)	(viii)	(c)	(ix)	(c)	(x)	(a)
(xi)	(a)	(xii)	(a)	(xiii)	(b)	(xiv)	(a)	(xv)	(a)
(xvi)	(b)	(xvii)	(c)	(xviii)	(b)	(xix)	(a)	(xx)	(b)
(xxi)	(c)	(xxii)	(c)	(xxiii)	(b)	(xxiv)	(c)	(xxv)	(b)
(xxvi)	(b)	(xxvii)	(c)	(xxviii)	(a)				

SHORT QUESTION ANSWERS

1. What is Binomial experiment?

Ans. A trial that gives only two possible outcomes is called a Bernoulli trial. For example success and failure, alive and dead etc. If the probability of each outcome remains the same for all trials, then such trials are Bernoulli trials. The experiment having n Bernoulli trials is called Binomial experiment.

2. Write the properties of Binomial experiment?

Ans. A binomial experiment possesses the following properties.

(i) The outcome of each trial may be classified into two categories i.e success and failure.

(ii) The probability of success remains constant for all trials.

(iii) The repeated trials are independent.

(iv) The experiment is repeated a fixed number of times.

3. Define Binomial probability distribution.

Ans. A distribution that gives the probability of x successes for fixed number of independent trials. The probability distribution function is.

$$P(X = x) = \binom{n}{x} p^x q^{n-x}; x = 0,1,2,.....n$$

It has two parameters n and p.

4. Define hypergeometric probability distribution.

Ans. Suppose a population consists of N items which are classified as K successes and K-K failure. If we select a sample of n elements from the population without replacement in such a way that x successes are selected from K successes and n-x failure from N-K failures. The probability distribution defined in this situation is called hypergeometric distributing given by

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, x = 0,1,2,.....n \text{ or } k$$

The hypergeometric distribution has three parameters N, n and K.

5. What are the properties of hypergeometric experiment?

Ans. Hypergeometric experiment has the following properties

(i) The successive trials are dependent

(ii) The probability of success does not remains constant for all trials.

(iii) The experiment is repeated a fixed number of times.

(iv) The outcome of an experiment can be classified into two categories.

6. **Define Poisson distribution.**

Ans. The probability distribution of Poisson random variable is called Poisson probability distribution: Its function is given as

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!} ; x = 0,1,2,.....\infty$$

Where e = 2.718

It has only one parameter "μ"

7. **Define Negative binomial distribution.**

Ans. Let the random variable x denote the number of trials untill k successes occur, then negative binomial distribution is given by

$$P(X = x) = \binom{x-1}{k-1} P^k q^{x-k}$$

$$x = k, k + 1, k + 2,.....$$

It has two parameters k and p.

8. **What is geometric distribution?**

Ans. When an experiment consists of repeated independent trials with probability P of success and the trials are repeated untill the first success occur is called geometric distribution. Its Pdf is given by

$$P(X = x) = Pq^{x-1} ; x = 1,2,3,.....\infty$$

9. **Define normal distribution.**

Ans. A normal distribution is defined by the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < X < \infty$$

Where $\pi = \frac{22}{7}$ and e = 2.718

It has two parameters μ and σ^2

10. **What is standardized normal distribution?**

Ans. If the observations of real life problems follow the normal distribution with mean μ and variance σ^2, then its random variable can be converted into standard normal variable using the following transformation with mean 0 and $\sigma = 1$.

$$z = \frac{x - \mu}{\sigma}$$

The P.d of z, denoted by $\phi(z)$, becomes

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{for } -\infty < z < \infty$$

11. Write down the properties of normal distribution.

Ans. The main properties of the normal distributions are

(i) It is continuous distribution ranges from $-\infty < z < \infty$.

(ii) It has two parameters μ and σ^2

(iii) Its mean is μ and variance is σ^2

(iv) It is symmetrical distribution i.e Mean = Median = Mode

(v) MD = $4/5\ \sigma$

(vi) It is a unimodel and bell shaped.

CHAPTER # 10

SAMPLING AND SAMPLING DISTRIBUTIONS

10.1 INTRODUCTION

Sampling is used in almost every field of life. A cook taste a bit of cooked food to find whether it has been properly cooked or not, A food inspector takes a sample of food or items like milk, flour etc to find whether they are pure or not. In medical sciences, few drops of blood are taken and tested in a laboratory to know whether the blood contains some abnormalities or not. Before discussing sampling briefly we first define terms population and a sample.

10.2 POPULATION AND SAMPLE

The total number of objects under consideration is called population. For example, Number of students in a college, the number of books in a library etc.

Population may be finite or infinite. When the total number of objects are known, then it is called finite population. For example, number of colleges in Punjab, number of teachers in a school etc. When the total number of objects are not known *i.e.* they can not be counted, then it is called infinite population. For example, number of stars on the sky, number of fish in a lake etc.

Sample is the small part of population which represents the whole population. For example, if we select 10 students from a class consisting of 90 students. 10 selected students constitute a sample.

10.3 SAMPLING

The process of selecting a sample from the population is called sampling. There are two basic aims of sampling.

(i) To get maximum information about a population without studying each and every unit of the population.

(ii) To get the reliability of the estimates.

10.3.1 Advantages of Sampling

The important advantages of sampling are as under.

(i) Reduced Cost

A sample survey will usually less costly than a complete census *i.e.* It is cheaper to collect information from 2000 people than from two millions.

(ii) Time Saving

Sampling saves a lot of time and energy. The data can be collected and processed more quickly in a sample than a complete enumeration.

(iii) Greater Accuracy

The results obtained from a sample can be carefully planned and are expected to be more accurate.

(iv) Measure of Reliability

The most important advantage of sampling is that it provides a valid measure of reliability for the sample estimates.

(v) Feasibility

Sampling is essential to obtaining the data, when the measurement process physically damage or destroyed the sampling units under consideration. For example, whole pot of soap can not be tested to determine if it has an acceptable flavor. If we want to investigate the average life time of bulbs from a lot, we will put a few light bulbs on test. In such cases complete enumeration would neither by physically possible nor practically feasible.

10.3.2 Limitations of Sampling

In sampling a careful, proper plan is needed, otherwise incorrect and misleading results will be obtained. Secondly if the information about each and every item of the population is required, then the complete enumeration is better. e.g. the list of voters is prepared to include the name of each and every voter.

10.3.3 Sample Design and Sample Survey

In sampling studies we have to make a plan regarding the size of the sample, selection of the sample, collection of the sample data and preparation of the final results based on sample study. The whole procedure involved is called the sample design. The collection of detailed information is known as survey. When a survey is carried out by a sampling design, it is called sample survey.

10.3.4 Sampling Frame

A sampling frame is the complete list of all elements in the population. For example, A list of household in a city, telephone directory. The requirement of a good frame are that the frame should

(i) not contain in accurate sampling units:

(ii) be complete and exhaustive

(iii) be free from errors and duplication of sampling units.

(iv) be as up-to-data as possible at the time of use.

Most of the frames used for sampling do not meet all these requirements.

10.3.5 Parameter and Statistic

Parameter

A numerical quantity, which is calculated from the population is called parameter. For example, population mean "μ", standard deviation "σ".

Statistic

A numerical quantity, which is calculated from the sample is called statistic. For example, "$\bar{x}$", standard deviation S.

It is important to note that parameters are fixed number, *i.e.* constant on the other hand statistic is a variable, because they vary from sample to sample for the same situation.

10.3.6 Sampling with and Without Replacement

Samples may be selected with replacement or without replacement. When any sampling unit is drawn from a finite population and then returned to the population before another unit is drawn is called **sampling with replacement.**

The total number of samples in case of with replacement can be obtained by using the formula N^n, where N is the population size and n is the sample size.

When any sampling unit is drawn from the population and it is not replaced before the selection of next unit, then the sampling is said to be sampling without replacement.

The total number of samples in case of without replacement can be obtained by using the relation $\binom{N}{n}$.

10.3.7 Sampling and Non sampling Errors

Sampling Error

Sampling error is the difference between the value of the sample and the true value of the population parameter.

i.e. $\bar{x} - \mu$

As the sample size increases, the sampling error is reduced.

Non Sampling Error

These errors are not due to sampling techniques but they occur during the process of collection of data.

The basic causes of these errors are due to

(i) In proper sampling procedures

(ii) Misreporting facts

(iii) In Perfect Selection

These errors can be avoided through the proper selection of questionnaires, proper training of the investigators etc.

10.4 PROBABILITY AND NON-PROBABILITY SAMPLING

Sampling methods are classified as probability sampling and non-probability sampling. When each unit in a population has known probability of being included in the sample then sampling is said to be probability sampling or random sampling. Examples of probability sampling are simple random sampling, stratified random sampling, systematic sampling and cluster sampling.

A process in which the personal judgement determines which unit of the population is selected for a sample is called non probability sampling. Judgement and Quota sampling are the examples of non-probability sampling.

10.5 TYPES OF PROBABILITY SAMPLING

The main types of probability sampling are given below.

10.5.1 Simple Random Sample

A sample is said to be a simple random sample if each unit in the population has an equal probability of being included in the sample.

Suppose a finite population contains N units and a sample of size n is to be selected. if we take sample with replacement, then the number of possible samples of size n is N^n, as the first unit of the sample can be selected in N different ways and so on.

The main advantage of simple random sampling is that it provides unbiased estimates of population mean.

10.5.2 Stratified Random Sampling

When the population is heterogeneous, then the simple random sampling fails to represent the whole population. In such cases the population is divided into homogenous groups. This process of dividing the population into groups is called stratification. Each group is called a stratum. We select a simple random sampling from each stratum and combine these samples into a single sample, then the sampling is called stratified random sampling.

10.5.3 Systematic Random Sampling

This is a technique in which samples are drawn according to some pre-determined pattern. For example, A population consisting of N values is numbered 1 to N. Now a sample of size n is to drawn from the population. Some times the initial starting unit is determined by selecting at random a unit between 1 and k, where k is known as sampling interval. Such a process is called systematic random sampling.

composed by ARFAN ALI MATHEMATICIAN 03148100197

Suppose a population consists of 100 values which are numbered 1 to 100. Suppose that the "k", class interval is equal to 10. The first unit is selected at random between 1 and 10. Let a selected unit will be 7, then the other samples are

$$7, 7 + k, 7 + 2k, 7 + 3k, \ldots\ldots$$

i.e. 7, 17, 27, 37, ..., 97

Systematic random sampling saves lot of time, cost and efforts.

10.5.4 Cluster Sampling

In cluster sampling, we divide the population into groups, or clusters and there select a simple random samples of these clusters. We assume that these individual clusters are representative of the population as a whole. If a market research team is attempting to determine by sampling the average number of air conditioner per household in a large city, they could use a city map to divide the city into blocks and then choose a certain number of blocks (clusters) for interviewing. Every household in each of these blocks would be interviewed.

The difference between a cluster and a stratum is that a stratum is expected to be as homogenous as possible, while cluster must be a hetrogenous as possible.

10.6 TYPES OF NON PROBABILITY SAMPLING

The important types of non-probability sampling are judgment sampling and Quota sampling.

10.6.1 Judgment Sampling

In this method, the sampling is not based on probability theory. But the personal judgment is involved in the selection of samples. It is not an objective method for measuring the reliability of the sample results.

The judgment or purposive sampling is preferred over probability sampling in several situations. For example, when taking a sample of melons from a truck load, the observer selects a few melons using his own judgment.

10.6.2 Quota Sampling

In this sampling the information is collected purposively from a segment of population (the quotas) the quota of urban and rural, Men and women etc. It is widely used in public opinions, polls and market research survey.

10.7 SAMPLING DISTRIBUTION

The probability distribution of a statistic is called sampling distribution. For example, the distribution of sample mean is a sampling distribution of mean.

The standard deviation of the sampling distribution of a statistic is called standard error of the statistic.

10.7.1 Sampling Distribution of Mean

The sampling distribution of the mean is the probability distribution of the mean $\bar{x}$ of all possible random samples of the same size that could be selected from a given population.

The mean of the distribution is denoted by $\mu_{\bar{x}}$ and the standard error of the mean $\sigma_{\bar{x}}$.

The sampling distribution of the mean has the following properties.

(i) $\mu_{\bar{x}} = \mu$

(ii) $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ (When the sampling is done with replacement)

$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\left(\dfrac{N-n}{n-1}\right)$ (When the sampling is done W.O.R)

If x is a random variable which follows normal distribution with mean μ and variance σ^2, $\bar{x}$ is also a random variable which follows normal distribution with mean μ and variance σ^2/n. The standard normal variable then becomes

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

If sampling is with out replacement and sample size n is 5% or greater than 5% of population size N, then z values are obtained by the formula

$$z = \frac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}}}$$

EXAMPLE 10.1

A population consists of 4 numbers 2, ,4 6 and 8. Considering all possible samples of size 2 which can be drawn with replacement from the population. Find

(i) Sampling Distribution of $\bar{x}$

(ii) Verify the result

$\mu_{\bar{x}} = u$

and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

SOLUTION

Total number of samples = $4^2 = 16$

Sample	$\bar{x}$	Sample	$\bar{x}$
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

Sampling Distribution of $\bar{x}$

$\bar{x}$	Tally bar	f	$f(\bar{x})$	$\bar{x}\, f(\bar{x})$	$\bar{x}^2\, f(\bar{x})$
2	|	1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{4}{16}$
3	||	2	$\frac{2}{16}$	$\frac{6}{16}$	$\frac{18}{16}$
4	|||	3	$\frac{3}{16}$	$\frac{12}{16}$	$\frac{48}{16}$
5	||||	4	$\frac{4}{16}$	$\frac{20}{16}$	$\frac{100}{16}$
6	|||	3	$\frac{3}{16}$	$\frac{18}{16}$	$\frac{108}{16}$
7	||	2	$\frac{2}{16}$	$\frac{14}{16}$	$\frac{98}{16}$
8	|	1	$\frac{1}{16}$	$\frac{8}{16}$	$\frac{64}{16}$
Σ	—	16	1	$\frac{80}{16}$	$\frac{440}{16}$

$$\mu_{\bar{x}} = \Sigma\bar{x}\, f(\bar{x}) = \frac{80}{16} = 5 \qquad \text{..... (i)}$$

$$\sigma^2_{\bar{x}} = \Sigma\bar{x}^2\, f(\bar{x}) - [\Sigma\bar{x}\, f(\bar{x})]^2$$

$$= \frac{440}{16} - (5)^2 = 2.5$$

$$\sigma_{\bar{x}} = \sqrt{2.5} = 1.58 \quad \text{..... (ii)}$$

x	x^2
2	4
4	16
6	36
8	64
20	**120**

Mean and Variance of Population

$$\mu = \frac{\Sigma x}{N} = \frac{20}{3} = 5 \quad \text{..... (iii)}$$

$$\sigma^2 = \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

$$= \frac{120}{4} - (5)^2$$

$$= 30 - 25$$

$$\sigma^2 = 5$$

$$\sigma = 2.236$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{\sqrt{2}} = 1.58 \quad \text{..... (iv)}$$

From eq. (i) and (iii), we have

$$\mu_{\bar{x}} = \mu = 5$$

From equation (ii) and (iv), we have

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 1.58 \text{ proved}$$

EXAMPLE 10.2

A population consists of 6 numbers 0, 3, 6, 12, 15 and 18. How many samples of size n = 3 can be drawn, without replacement from this population? Compute the sampling distribution the mean. Compute the mean and variance of this distribution and verify that

$$\mu_{\bar{x}} = \mu$$

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

SOLUTION

Total number of samples of size 3 = $\binom{6}{3}$ = 20

Samples	$\bar{x}$	Samples	$\bar{x}$
0, 3, 6	3	3, 6, 12	7
0, 3, 12	5	3, 6, 15	8
0, 3, 15	6	3, 6, 18	9
0, 3, 18	7	3, 12, 15	10
0, 6, 12	6	3, 12, 18	11
0, 6, 15	7	3, 15, 18	12
0, 6, 18	8	6, 12, 15	11
0, 12, 15	9	6, 12, 18	12
0, 12, 18	10	6, 15, 18	13
0, 15, 18	11	12, 15, 18	15

Sampling Distribution of $\bar{x}$

$\bar{x}$	Tally	f	$f(\bar{x})$	$\bar{x}\, f(\bar{x})$	$\bar{x}^2\, f(\bar{x})$
3	|	1	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{9}{20}$
5	|	1	$\frac{1}{20}$	$\frac{5}{20}$	$\frac{25}{20}$
6	||	2	$\frac{2}{20}$	$\frac{12}{20}$	$\frac{72}{20}$
7	|||	3	$\frac{3}{20}$	$\frac{21}{20}$	$\frac{147}{20}$
8	||	2	$\frac{2}{20}$	$\frac{16}{20}$	$\frac{128}{20}$
9	||	2	$\frac{2}{20}$	$\frac{18}{20}$	$\frac{162}{20}$
10	||	2	$\frac{2}{20}$	$\frac{20}{20}$	$\frac{200}{20}$

11	|||	3	$\frac{3}{20}$	$\frac{33}{20}$	$\frac{363}{20}$
12	||	2	$\frac{2}{20}$	$\frac{24}{20}$	$\frac{288}{20}$
13	|	1	$\frac{1}{20}$	$\frac{13}{20}$	$\frac{169}{20}$
15	|	1	$\frac{1}{20}$	$\frac{15}{20}$	$\frac{225}{20}$
Σ	—	**20**	**1**	$\frac{180}{20}$	$\frac{1788}{20}$

$$E(\bar{x}) = \mu = \Sigma\bar{x}\, f(\bar{x})$$

$$= \frac{180}{20} = 9 \qquad \text{..... (i)}$$

$$\sigma^2_{\bar{x}} = \Sigma\bar{x}^2 f(\bar{x}) - [\Sigma\bar{x}\, f(\bar{x})]^2$$

$$= \frac{1788}{200} - (9)^2$$

$$\sigma^2_{\bar{x}} = 89.4 - 81$$

$$= 8.4 \qquad \text{..... (ii)}$$

Verification

x	x^2
0	0
3	9
6	36
12	144
15	225
18	324
54	**738**

$$\mu = \frac{\Sigma x}{N} = \frac{54}{6} = 9 \qquad \text{..... (iii)}$$

$$\sigma^2 = \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

$$= \frac{738}{6} - (9)^2$$

$$= 123 - 81 = 42$$

$$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

$$= \frac{42}{3}\left(\frac{6-3}{6-1}\right)$$

$$= \frac{42}{3}\left(\frac{3}{5}\right) = 8.4 \qquad \text{..... (iv)}$$

From equation (i) and (iii), we have

$$\mu_{\bar{x}} = \mu = 9$$

From equation (ii) and (iv), we get

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = 8.4 \quad \textbf{Proved}$$

EXAMPLE 10.3

A population consists of six values 2, 4, 6, 8, 10, 12. We have to consider all possible samples of size four with out replacement. Find

(i) Mean and Standard Deviation of population

(ii) Number of samples

(iii) Mean and Standard deviation of sampling distribution of mean.

SOLUTION

x	x^2
2	4
4	16
6	36
8	64
10	100
12	144
42	364

(i) $\text{Mean} = \mu = \frac{\Sigma x}{N} = \frac{42}{6} = 7$

$$\sigma^2 = \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

$$= \frac{364}{6} - (7)^2$$

$$= 60.67 - 49 = 11.67$$

$$\sigma = \sqrt{11.67} = 3.416$$

(ii) Number of samples = $\binom{N}{n} = \binom{6}{4} = 15$

Samples	$\bar{x}$	Samples	$\bar{x}$
2, 4, 6, 8	5.0	2, 6, 10, 12	7.5
2, 4, 6, 10	5.5	2, 8, 10, 12	8.0
2, 4, 6, 12	6.0	4, 6, 8, 10	7.0
2, 4, 8, 10	6.0	4, 6, 8, 12	7.5
2, 4, 8, 12	6.5	4, 6, 10, 12	8.0
2, 4, 10, 12	7.0	4, 8, 10, 12	8.5
2, 6, 8, 10	6.5	6, 8, 10, 12	9.0
2, 6, 8, 12	7.0		

Sampling distribution of mean ($\bar{x}$)

$\bar{x}$	Tally	f	$f(\bar{x})$	$\bar{x}\, f(\bar{x})$	$\bar{x}^2\, f(\bar{x})$
5.0	|	1	$\frac{1}{15}$	$\frac{5}{15}$	$\frac{25}{15}$
5.5	|	1	$\frac{1}{15}$	$\frac{5.5}{15}$	$\frac{30.25}{15}$
6.0	||	2	$\frac{2}{15}$	$\frac{12}{15}$	$\frac{72}{15}$
6.5	||	2	$\frac{2}{15}$	$\frac{13}{15}$	$\frac{84.5}{15}$
7.0	|||	3	$\frac{3}{15}$	$\frac{21}{15}$	$\frac{147}{15}$
7.5	||	2	$\frac{2}{15}$	$\frac{15}{15}$	$\frac{112.5}{15}$
8.0	||	2	$\frac{2}{15}$	$\frac{16}{15}$	$\frac{128}{15}$
8.5	|	1	$\frac{1}{15}$	$\frac{8.5}{15}$	$\frac{72.25}{15}$
9.0	|	1	$\frac{1}{15}$	$\frac{9}{15}$	$\frac{81}{15}$
Σ	—	15	1	$\frac{105}{15}$	$\frac{752.5}{15}$

$$\mu_{\bar{x}} = E(\bar{x}) = \Sigma\bar{x}\ f(\bar{x}) = \frac{105}{5} = 7.0$$

$$\sigma^2_{\bar{x}} = \Sigma\bar{x}^2\ f(\bar{x}) - [\Sigma\bar{x}\ f(\bar{x})]^2$$
$$= \frac{752.5}{15} - (7)^2$$
$$= 50.17 - 49 = 1.17$$
$$\sigma_{\bar{x}} = \sqrt{1.17} = 1.08$$

EXAMPLE 10.4

The random variable x has the following probability distribution.

x	0	1	2	3
f(x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Find the mean $\mu_{\bar{x}}$ and variance $\sigma^2_{\bar{x}}$ of the mean $\bar{x}$ for a sample of 16.

SOLUTION

x	f(x)	x f(x)	x^2 f(x)
0	$\frac{1}{8}$	0	0
1	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$
2	$\frac{3}{8}$	$\frac{6}{8}$	$\frac{12}{8}$
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{9}{8}$
Σ	1	$\frac{12}{8}$	$\frac{24}{8}$

$$\mu = E(x) = \Sigma x\ f(x)$$
$$= \frac{12}{8} = 1.5$$
$$\sigma^2 = Var\ (x) = \Sigma x^2\ f(x) - [\Sigma x\ f(x)]^2$$
$$\sigma^2 = \frac{24}{8} - (1.5)^2$$
$$= 3 - 2.25$$
$$= 0.75$$

We know that from sampling distribution of mean

$$\mu_{\bar{x}} = \mu = 1.5$$
$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{0.75}{16} = 0.047$$

EXAMPLE 10.5

If the size of simple random sample from an infinite population is 36 and the standard error of the mean is 2. What must the size of the sample become if the standard error is to be reduced to 1.2.

SOLUTION

$n = 36, \sigma_{\bar{x}} = 2$

We know that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$2 = \frac{\sigma}{\sqrt{36}}$$

$$2 = \frac{\sigma}{6}$$

$$\sigma = 12$$

If standard error $\sigma_{\bar{x}}$ is reduced to 1.2 then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$1.2 = \frac{12}{\sqrt{n}}$$

$$12 = 1.2\sqrt{n}$$

$$\sqrt{n} = \frac{12}{1.2}$$

Squaring both sides, we have

$n = 100$

EXAMPLE 10.6

A random sample of size 70 is taken from normal population which has its mean and variance as 200 and 750 respectively what is the probability that the sample mean is greater than 205?

SOLUTION

We have given

$n = 70, \mu = 200, \sigma^2 = 750$

$P(\bar{x} > 205) = ?$

We know that

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{\bar{x} - 200}{27.386/\sqrt{70}}$$

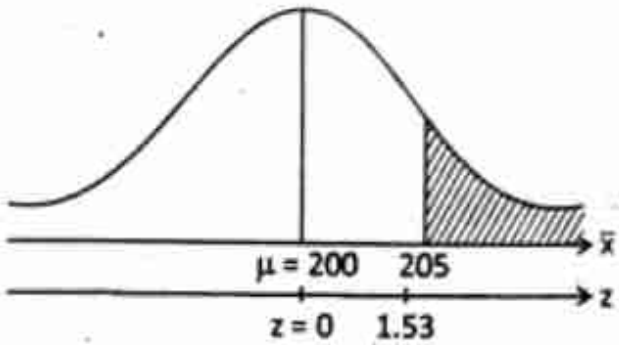
$$z = \frac{\bar{x} - 200}{3.273}$$

When $\bar{x} = 205$, then

$$z = \frac{205 - 200}{3.273} = 1.53$$

$$P(z > 1.53) = P(0 \text{ to } \infty) - P(0 \text{ to } 1.53)$$

$$= 0.5 - 0.4370 = 0.063$$



EXAMPLE 10.7

The age of employees in a company follows normal distribution with mean and variance 40 and 121 years respectively. If a random sample of 64 employees is taken from a finite population size 1000, what is the probability that a sample mean is (a) less than 45 (b) greater than 42.

SOLUTION

We have given that

$$\mu = 40, \sigma^2 = 121, n = 64, N = 1000$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$$

(Since sample size n = 64 is greater than 5% of N = 1000, so we use f.p.c, the factor $\sqrt{\frac{N-n}{N-1}}$)

$$z = \frac{\bar{x} - 40}{11/8\left(\frac{1000-64}{1000-1}\right)}$$

$$z = \frac{\bar{x} - 40}{1.29}$$

(a) $P(\bar{x} < 45) = ?$

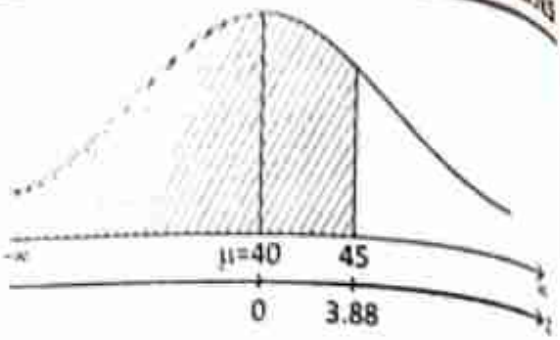
When $\bar{x} = 45$, then $z = \frac{45 - 40}{1.28} = 3.88$

$P(z < 3.88) = ?$

$= P(-\infty \text{ to } 0) + P(0 \text{ to } 3.88)$

$= 0.5 + 0.4999$

$= 0.9999$

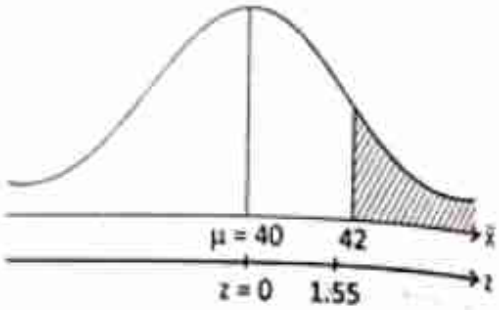


(b) $P(\bar{x} > 42) = ?$

When $\bar{x} = 42$, then $z = \frac{42 - 40}{1.29} = z = 1.55$

$= 0.5 - 0.4394$

$= 0.0606$



10.8 CENTRAL LIMIT THEOREM

If a variable x from a population has mean μ and finite variance σ^2, then the sampling distribution of sample mean $\bar{x}$ approaches a normal distribution with mean μ and variance σ^2/n as the sample size n approaches infinity.

This theorem is useful to sample a population when its distribution is unknown or non normal.

EXAMPLE 10.8

A random sample of size 36 is selected from a Poisson distribution with mean = 4. Find using the central limit theorem, the probability that a sample mean is less than 3.5.

SOLUTION

Let x denote the Poisson distribution with mean = 4, then Var(X) = 4

By central limit theorem, $\bar{x}$ is approximately $N\left(4, \frac{4}{36}\right)$

We need $P(\bar{x} < 3.5)$

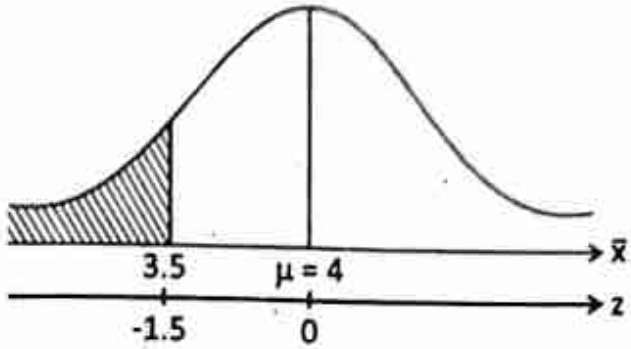
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$z = \frac{3.5 - 4}{2/\sqrt{36}} \quad \Rightarrow \quad z = \frac{-0.5}{0.333} = -1.50$$

$P(z < -1.50) = P(-\infty \le z < 0) - P(-1.5 \le z\ 0)$

$= 0.5 - 0.4332$

$= 0.0668$



10.9 SAMPLING DISTRIBUTION OF SAMPLE PROPORTION

In a sample of n successive trials, let the probability of success in a trial be p and x be the number of successes. Then x follows binomial distribution which has mean and variance nP and nPq respectively.

For large value of n, the binomial distribution can be approximated to normal distribution as $x \sim N(\mu, \sigma^2)$. Let the proportion of success be $\hat{P}$ and it is given by

$$\hat{P} = \frac{x}{n}$$

The sampling distribution of $\hat{P}$ has the following important properties

(i) $\mu_{\hat{P}} = P$

(ii) $\sigma_{\hat{P}} = \sqrt{\frac{Pq}{n}}$ (when the sampling is done with replacement)

$\sigma_{\hat{P}} = \sqrt{\frac{Pq}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ (when the sampling is done W.O.R)

The standard normal variable in case of proportion is given by

$$z = \frac{\hat{P} - P}{\sqrt{\frac{Pq}{n}}}$$

EXAMPLE 10.9

A population consists of N = 5 values 1,2,3,4,5. Draw all possible sample of size 3 with out replacement from the population and find the proportion of odd numbers in the samples. Construct the sampling distribution of sample proportion and verify that

(i) $\mu_{\hat{P}} = P$ (ii) $Var(\hat{P}) = \frac{Pq}{n}\left(\frac{N-n}{N-1}\right)$

SOLUTION

The number of possible samples $= \binom{N}{n} = \binom{5}{3} = 10$

Sample No.	Samples	$\hat{P}$
1	1,2,3	2/3
2	1,2,4	1/3
3	1,2,5	2/3
4	1,3,4	2/3
5	1,3,5	3/3
6	1,4,5	2/3
7	2,3,4	1/3
8	2,3,5	2/3
9	2,4,5	1/3
10	3,4,5	2/3

The sampling distribution of sample proportion is given below

$\hat{P}$	Tally bar	f	$f(\hat{P})$	$\hat{P}f(\hat{P})$	$\hat{P}^2f(\hat{P})$
1/3	III	3	3/10	3/20	3/90
2/3	~~IIII~~ I	6	6/10	12/30	24/90
3/3	I	1	1/10	3/30	9/90
Σ	–	10	1	18/30	36/90

$$\mu\hat{P} = \Sigma\hat{P}f(\hat{P}) = \frac{18}{30} = 0.60 \text{ and}$$

$$Var(\hat{P}) = \Sigma\hat{P}^2f(\hat{P}) - (\mu\hat{P})^2$$

$$= \frac{36}{90} - \left(\frac{18}{30}\right)^2 = 0.4 - 0.36 = 0.04$$

Now

$$P = \frac{X}{N} = 3/5 = 0.6\ ,\ q = 1 - P = 0.4$$

$$\frac{Pq}{n}\left(\frac{N-n}{N-1}\right) = \frac{3}{5}\cdot\frac{2}{5}\cdot\frac{1}{3}\left(\frac{5-3}{5-1}\right) = 0.04 = Var(\hat{P})$$

Hence proved.

EXAMPLE 10.10

The manager of a company feels that 52% of the employees will have enhanced skill after attending the training programme. A sample of 49 employees who attended the training on skill development, reveals that only 248 them are having enhanced skill. Find the probability that a sample of employees who attended the training have enhanced their skill.

SOLUTION

We have $P = 0.52$, $n = 49$, $q = 1 - P = 0.48$

$$\hat{P} = \frac{x}{n} = \frac{24}{49} = 0.49$$

$$z = \frac{\hat{P} - P}{\sqrt{\frac{Pq}{n}}} = z = \frac{0.49 - 0.52}{\sqrt{\frac{0.52\,(0.48)}{49}}}$$

$$z = -0.42$$

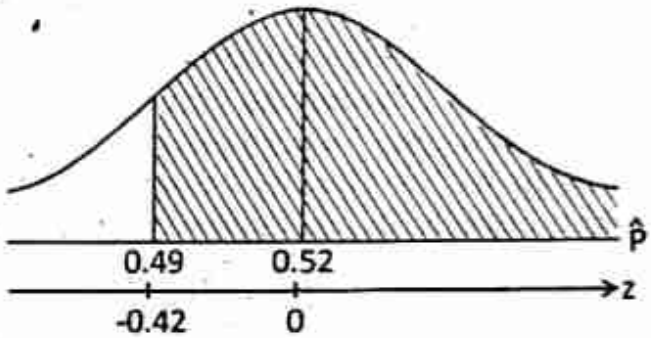
$P(\hat{P} > 0.49) = ?$

$P(z > -0.42) = ?$

$= P(-0.42 \text{ to } 0) + P(0 \text{ to } \infty)$

$= 0.1628 + 0.5000$

$= 0.6628$



The probability that the sample of employees who attended the training have enhanced their skill is 0.6628

10.10 STATISTICAL INFERENCE: ESTIMATION

The art of drawing conclusions or inferences about the population from the limited information contained in the sample is called statistical inference. Statistical inference is divided into two major areas: estimation of parameters and testing of hypothesis.

Estimation is a technique by which we obtain an estimate of the true but unknown value of population parameter by using the sample observations from the population. For example we may estimate the mean and variance of population by calculating the mean and variance of sample drawn from the population. Estimation is further divided into point estimation and interval estimation.

10.11 POINT ESTIMATION

The object of point estimation is to obtain a single value from the sample which will represent the unknown value of the parameter. A statistic used to estimate a parameter is called an estimator, the numerical value which we obtain for an estimator in a given problem is called estimate. Let x be the random variable denoting of the weight of each student. A sample of n students is drawn and mean weight of sample ($\bar{x}$) is competed. The statistic $\bar{x}$ is a point estimator of population mean (μ).

10.12 INTERVAL ESTIMATION

A point estimator calculated from the sample data provides a single value as an estimate of the parameter. A point estimator can not be expected to be exactly equal to the population parameter. For example, the mean of the sample taken from population may assume different values for different samples. We therefore, estimate an interval of values within which the population parameter is expected to occur.

10.13 CONFIDENCE INTERVAL

A confidence interval is a range of values within which the population parameter is expected to occur. The probability attach to the confidence interval is called confidence coefficient or level of confidence. This probability indicates how confident we are that the interval estimate will include the population parameter. A higher probability means more confidence. In estimation the most commonly used confidence levels are 90 percent, 95 percent and 99 percent, but we are free to apply any confidence level according to situation of the problem/research.

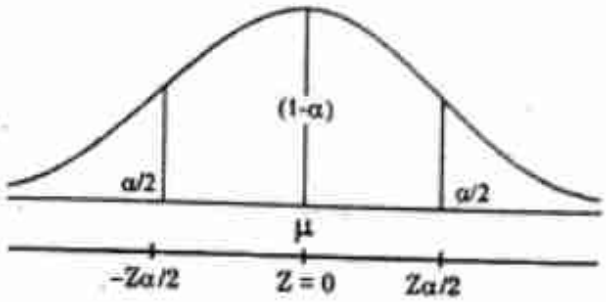
A confidence interval estimate is in the form $[L < \theta < U]$, where L and U depend upon the value of statistic of a random sample selected from the population and the sampling distribution of the statistic. The interval (L, U) is called a 100 $(1-\alpha)$ percent confidence interval for the unknown parameters θ. The difference U-L i.e. width of the confidence interval is called the precision of the estimate. The precision may be increased either by decreasing the confidence level or increasing the sample size.

10.13.1 Confidence Interval Estimate of a Population Mean when σ known

Let a random sample $x_1, x_2, \ldots\ldots, X_n$ of size n be drawn from a normal population whose mean (μ) is unknown and variance (σ^2) is known. Then the sampling distribution of the mean $\bar{x}$ will be normal with mean μ and standard deviation $\sigma/\sqrt{n}$; and the variable $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ will be exactly standard normal.

The normal distribution tells us that the probability that a value of z will fall in the interval from $-Z_{\frac{\alpha}{2}}$ and $Z_{\frac{\alpha}{2}}$ is equal to 1-α. That is we make the following statement

$$P\left[-Z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$



(i) We multiply all terms inside the brackets by $\frac{\sigma}{\sqrt{n}}$ we get

$$-Z\alpha/_2 \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$$

(ii) We subtract $\bar{x}$ from each term we have

$$-\bar{x} - Z\alpha/_2 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$$

(iii) We multiply all terms by (−1), we get

$$\bar{x} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}} > \mu > \bar{x} - Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$$

Which is equal to

$$\bar{x} - Z\alpha/_2 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$$

(iv) We substitute this result in the probability form

$$P[\bar{x} - Z\alpha/_2 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}}] = 1-\alpha$$

Hence, for a particular sample size n, a 100 (1 − α) percent confidence interval for μ is given by

$$(\bar{x} - Z\alpha/_2 \frac{\sigma}{\sqrt{n}}, \bar{x} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}}) = \bar{x} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$$

Confidence interval when σ is unknown and the sample size is sufficiently large (n ≥ 30) is

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

EXAMPLE 10.11

The diameter of shafts produced in a machine follows normal distribution with a variance of 81 mm. A random sample of 36 shafts taken from a production has its mean diameter of 30 mm. Find 95 percent confidence interval for the diameter of shafts.

SOLUTION

We have

$$\sigma^2 = 81, \bar{x} = 30, n = 36, \alpha = 0.05$$

and $Z_{\alpha/2} = 1.96$

The 95% confidence interval for μ is

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$30 \pm 1.96\frac{9}{\sqrt{36}}$$

30 ± 2.94 or 27.06 mm and 32.94 mm

The 95% confidence interval for μ is (27.06, 32.94).

EXAMPLE 10.12

A random sample of 100 workers with children in day care shows a mean day-care cost of Rs.5000 and a standard deviation of Rs.1000. Calculate 95 percent confidence interval.

SOLUTION

We have

$$n = 100, \bar{x} = 5000, s = 1000, \alpha\ 0.05$$

$$Z_{\alpha/2} = 1.96$$

The 95% confidence interval for μ is

$$\bar{x} \pm Z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$5000 \pm 1.96\left(\frac{1000}{\sqrt{100}}\right)$$

$5000 \pm 1.95\ (100)$

5000 ± 196 or $5000 - 196, 5000 + 196$

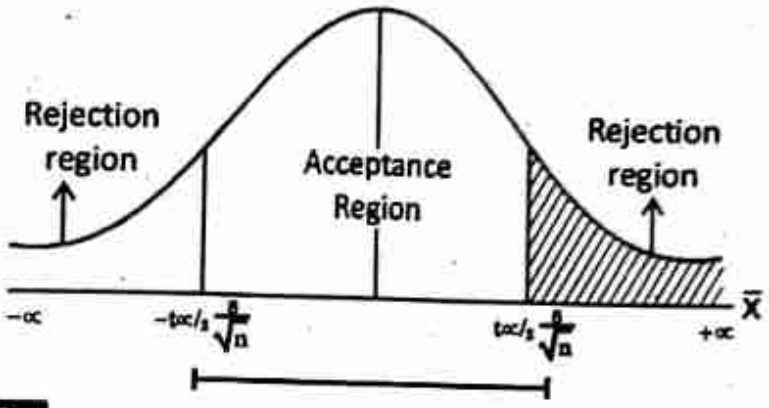
4804, 5196

The 95% confidence interval for μ is (Rs.4804, Rs.5196)

10.13.2 Confidence Interval Estimation of Population Mean when σ unknown and Sample Size is Less than 30 (n < 30)

Consider a normal population whose mean and standard deviation are unknown. If a small sample of size n (n < 30) is selected from the population, then the confidence interval for population mean (μ) is.

$$\bar{x} \pm t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}} \quad t^{\alpha}_{2}{}_{(n-1)}$$



EXAMPLE 10.13

A sample of 5 measurements of the reaction time of a certain stimulus gave a mean 0.298 seconds and a standard deviation 0.024 seconds. Find 95 percent confidence interval for the actual mean reaction time.

SOLUTION

$n = 5,\ \bar{x} = 0.298$ seconds, $s = 0.024$ seconds

$\alpha = 0.05,\ t_{\alpha/2}(\upsilon) = t_{0.025}(4) = 2.78 \qquad \therefore\ = n - 1 = 4$ d.f

95% confidence interval for μ is

$$\bar{X} \pm t_{\alpha/2}(\upsilon) \frac{s}{\sqrt{n}}$$

$$0.298 \pm 2.78\left(\frac{0.024}{\sqrt{5}}\right)$$

$$0.298 \pm 0.030$$

i.e. 0.268, 0.328 seconds

10.13.3 Confidence Interval for Population Proportion

Let a random sample of size n (n > 30) be selected from a binomial population with an unknown proportion of success P and let the sample proportion be $\hat{P} = \frac{x}{n}$.

We know that sampling distribution of sample proportion $\hat{P}$ is approximately normal with a mean P and standard deviation of $\sqrt{\frac{Pq}{n}}$ when P is not too close to 0 or 1. Thus, a 100 (1 – α)1/2 confidence interval for P is

$$\hat{P} \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{q}}{n}}$$

EXAMPLE 10.14

In a random sample of 160 people, 50 are found to be smokers obtain 95 percent confidence interval for the true proportion of smokers in a population.

SOLUTION

An approximate 95% confidence interval for P is.

$$\hat{P} \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{q}}{n}}$$

Where $\hat{P} = \frac{x}{n} = \frac{50}{160} = 0.3125$, $\hat{q} = 1 - \hat{P} = 0.6875$

$\alpha = 0.05$, $Z_{\alpha/2} = 1.96$

Substituting these values, we get

$$0.3125 \pm 1.96\sqrt{\frac{(0.3125)(0.6875)}{160}}$$

$$0.3125 \pm 1.96\,(0.0366)$$

or 0.3125 ± 0.0718

or 0.2407, 0.3843

Hence 95% confidence interval for the true proportion μ of smokers in a locality is (0.2407, 0.3843)

10.14 DETERMINATION OF SAMPLE SIZE

If the sample size is very large, then the error in making decision will be very small. But it will be very costly exercise to do sampling for a large sample size. In sampling, the right size of sample minimizes the error in making decisions. Hence before performing sampling, it is general practice to determine the sample size for a desired accuracy based on the population variance.

10.14.1 Sample Size for Estimating Population Mean

The 100(1 –α) percent confidence interval for μ is given by

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu\ \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Which may be written as

$$|\bar{x} - \mu| = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The quantity $|\bar{x} - \mu|$ is called the error of estimator $\bar{x}$ and denoted by e.

The above equation may be written as

$$e = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or $$\sqrt{n} = Z_{\alpha/2}\frac{\sigma}{e}$$

or $$n = \left(\frac{Z_{\alpha/2}\,\sigma}{e}\right)^2$$

EXAMPLE 10.15

The axial length of pistons produced follows normal distribution with the variance of 64mm. Determine the sample size such that the mean length of pistons is within plus or minus 2 mm with a confidence level of 95%

SOLUTION

Here $\sigma 2 = 64\text{mm}, \quad \sigma = 8\text{mm}$

$\alpha = 0.05, Z_{\alpha/2} = 1.96; e = 2\text{mm}$

So the required sample size is

$$n = \left(\frac{Z_{\alpha/2}\,\sigma}{e}\right)^2$$

$$n = \left[\frac{1.96\,(8)}{2}\right]^2 = 61.47 \approx 62$$

10.14.2 Sample Size for Estimating Population Proportion

The 100(1 – α) percent large sample confidence interval for population proportion P is given by

$$\hat{P} \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{q}}{n}}$$

The error of estimation e will be given by

$$e = Z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{q}}{n}}$$

$$\sqrt{n} = Z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{q}}{n}}$$

$$n = \frac{(Z_{\alpha/2})^2 \hat{P}\hat{q}}{(e)^2}$$

EXAMPLE 10.16

In a random sample of 200 persons having their lunch at a college canteen on a meatless day it was observed that 30 percent preferred vegetables dishes. How large a sample size is needed if we want be to 95% confidence that our estimate of P within 0.02?

SOLUTION

Here $\hat{P} = 0.30, \hat{q} = 1 - \hat{P} = 0.70$

$n = 200, \alpha = 0.05, Z_{\alpha/2} = 1.96$ and $e = 0.02$

We know that

$$n = \frac{(Z_{\alpha/2})^2 \hat{P}\hat{q}}{(e)^2}$$

$$n = \frac{(1.96)^2 (0.30)(0.70)}{(0.02)^2}$$

$$n = 2016.84 \approx 2017$$

SUMMARY

The brief summary of this chapter is given below

APPLICATION	FORMULA
Sampling with replacement	N^n
Sampling without replacement	$\binom{N}{n}$
Sampling distribution of Mean	$\mu_{\bar{x}} = \mu$, $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$ (W.R) $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ (W.O.R) where $\sigma\bar{x} = \Sigma\bar{x}\, f(\bar{x})$ and $\sigma^2_{\bar{x}} = \Sigma\bar{x}^2 f(\bar{x}) - (\mu_{\bar{x}})^2$
Sampling distribution of proportion	$\mu_{\hat{P}} = P$ $Var(\hat{P}) = \frac{Pq}{n}$, (W-R) $Var(\hat{P}) = \frac{Pq}{n}\left(\frac{N-n}{N-1}\right)$, (W.O.R) Where $\mu_{\hat{P}} = \Sigma\hat{P}\, f(\hat{P}), \hat{P} = \frac{x}{n}$ $\hat{q} = 1 - \hat{P}$ $Var(\hat{P}) = \Sigma\hat{P}^2 f(P) - (-\mu\hat{P})^2$
Confidence interval for mean, when σ known or σ unknown and $n > 30$	$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Confidence for mean when σ unknown and sample size is less than 30 ($n < 30$)	$\bar{x} \pm t_{\alpha/2}(\upsilon)\frac{s}{\sqrt{n}}$

	Where $s=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ With $\upsilon = (n-1)$ degree of freedom
Confidence interval for proportion of large samples	$\hat{P} \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}\hat{q}}{n}}$
Estimation of sample size in case of mean	$n = \left(\frac{Z_{\alpha/2}\,\sigma}{e}\right)^2$
Estimation of sample size in case of proportion	$n = (Z_{\alpha/2})^2\,\hat{P}\hat{q}/(e)^2$

EXERCISES

10.1 (a) What do you meant by population, sample and sampling?

(b) Distinguish between parameter and statistic.

10.2 (a) Explain sampling and non sampling errors.

(b) Describe the advantages of sampling over complete enumeration.

10.3 (a) Define probability and non probability sampling.

(b) Explain the following concepts

(i) Simple random sampling

(ii) Stratified random sampling

(iii) Systematic sampling

(iv) Cluster Sampling

10.4 What do you understand by sampling with and without replacement.

10.5 (a) Define sampling distribution. Write down the properties of sampling distribution of mean.

(b) A population consists of 6 values 1, 3, 5, 7, 9, 11. Take all possible samples of size 2 which can be drawn with replacement from this population. Find the sample means and form a sampling distribution of mean. Verify that

(i) $\mu_{\bar{x}} = \mu$

(ii) $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

10.6 A sample of size n = 3 is to be randomly selected without replacement from a population 0, 3, 6, 9 and 12.

(i) Find the sampling distribution of sample mean, $\bar{x}$

(ii) Calculate mean and standard deviation of $\bar{x}$, and verify that

$\mu_{\bar{x}} = \mu$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$$

10.7 A population consists of values 0, 3, 6. Take all possible samples of size 3 with replacement from this population. From sampling distribution of mean and calculate mean and variance. Also calculate mean and variance of the population and verify the relation.

$\mu_{\bar{x}} = \mu$

$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

10.8 A population consists of the numbers 2, 4, 6, 6, 8, 10. Calculate the sample mean for all possible random samples of size 2 that can be drawn with replacement. From a sampling distribution of mean and find the mean and variance of this distribution and verify that

(i) $\mu_{\bar{x}} = \mu$

(ii) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

10.9 A sample of size n = 3 is to be randomly selected without replacement from a population that has N = 5, whose values are 0, 3, 6, 9, 12.

(i) Find the sampling distribution of $\bar{x}$

(ii) Calculate the mean and standard deviation of the sampling distribution of $\bar{x}$ and verify that

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

10.10 For a population consisting of 3, 6, 9, 12, 15, 18 draw all possible samples of size 3 without replacement. Find the means of these samples and make the frequency distribution of sample means. Calculate mean and standard deviation of population and sample means.

10.11 A population consists of six numbers 6,9, 12, 15, 18 and 21. Consider all possible samples of size = 3, which can be drawn without replacement from this population. Calculate

(i) The mean of population

(ii) The standard deviation of population

(iii) The mean of the sampling distribution of means

(iv) The standard error.

10.12 Draw all possible random samples of size 2 with replacement from the population 2, 4, 6, 8, 10. Make sampling distribution of sample means.

(i) Find mean and variance of sampling distribution of mean.

(ii) Find population mean and variance and verify the results.

10.13 Draw all possible random samples of size 2 without replacement from the population 0, 2, 4, 6, 8, 10. Make sampling distribution of sample means.

(i) Find mean and variance of sampling distribution of mean.

(ii) Find population mean and variance and verify the results.

10.14 A finite population consists of the numbers 2, 4 and 6.

(a) Form the sampling distribution of $\bar{x}$, when random samples of size 4 are drawn with replacement.

(b) Verify that $\mu_{\bar{x}} = \mu$, $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

10.15 The random variable x has the following probability distribution.

x	2	3	4	5	6	7
f(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$
x	8	9	10	11	12	
f(x)	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

Find the mean $\mu_{\bar{x}}$ and $\sigma^2_{\bar{x}}$ of the mean $\bar{x}$ of a random sample of 40.

10.16 Given the following population distribution.

x	1	2	3	4
f(x)	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{2}{7}$	$\frac{1}{7}$

Find sampling distribution of mean if a sample of three numbers is taken without replacement. How does the variance of the sampling distribution compare with the population variance.

10.17 (a) If the size of the simple random sample from population is 55, the variance of the sample mean is 27. Find the standard error if n = 165.

(b) If the size of simple random sample from an infinite population is 36 and standard error of the mean is 4. What must be the size of sample becomes if the standard error is reduced to 1.5.

10.18 (a) The annual income of employees of an industry follows normal distribution with mean and variance Rs.4 lakhs and Rs.1 lakh, respectively. A random sample of size 48 is taken. What is the probability that the sample mean is greater than 4.25 lakhs?

(b) A non-normal distribution representing the number of trips performed by trucks per week in a coal field has a mean of 100 trips and variance 121 trips. A random sample of 36 trucks in taken from the population what is the probabilities that the sample mean is

(i) Greater than 105 trips

(ii) Less than 102 trips

(iii) Between 101 and 103 trips

10.19 (a) What is meant by sampling distribution of sample proportion? Describe its important properties.

(b) Draw all possible samples of size 3 without replacement from the population 3,4,5,6,7. Find the proportion of even numbers in the sample. Construct the sampling distribution of sample proportion and verify

(i) $\mu_{\hat{P}} = P$ (ii) $Var(\hat{P}) = \frac{Pq}{n}\left(\frac{N-n}{N-1}\right)$

10.20 A population consists of 4 members. The marital status of each member is given below

Member	1	2	3	4
Martial Status	M	S	M	S

Where M stands for married and S for single.

Determine the proportion of single members in the population. Select all possible samples of two members with replacement from the populating.

Compute the $\mu\hat{P}$ and $Var(\hat{P})$ and verify that

(i) $\mu\hat{P} = P$ (ii) $Var(\hat{P}) = \frac{Pq}{n}$

10.21 A population consists of N = 6 values 1,3,6,8,9 and 12. Draw all possible samples of size 3 without replacement from the population and find the proportion of even numbers in the samples. Construct the sampling distribution of sample proportion and verify that

(i) $\mu\hat{P} = P$ (ii) $Var(\hat{P}) = \frac{Pq}{n} \cdot \left(\frac{N-n}{N-1}\right)$

10.22 (a) It has been observed that 20 percent of the people are smokers. What is the probability that in a random sample of 500 people more than 25 percent are smokers in a certain locality?

(b) Forty percent of all the workers in a factory are married. If 100 workers are selected at random. What is the probability that in a sample of100 workers, (i) less than 43 percent are married (ii) between 32% and 47% are married.

10.23 It has been found that 3% of the energy savers bulbs manufactured by a factory are defective. What is the probability that in a lot of 500 bulbs (i) 5% or more are defective (ii) 2% or less are defective.

10.24 (a) Explain what is meant by confidence interval, confidence limits and confidence coefficient?

(b) The diameter of shafts produced in a machine follows normal distribution with variance 121mm. A random sample of 49 shafts taken from the production has its mean diameter of 50mm. Find 95 percent confidence interval for the diameter of shafts.

10.25 (a) In an automotive safety test conducted by Highway safety research centre, the average tire pressure in a sample of 62 tires was found to be 24 pounds per square inch, and the standard deviation was 2.1 pounds per square inch. Calculate 95 percent confidence interval for the population mean.

(b) From a population 540, a sample of 60 individuals is taken. From this sample, the mean was found to 6.2 and the standard deviation 1.368. (i) Find the estimated standard error (ii) construct 95% confidence interval for the mean.

10.26 (a) A sample of 12 had a mean of 62 and a standard deviation of 10. Construct 95% confidence interval for the population mean.

(b) Seven homemakers were randomly selected, and it was determined that the distance, they travelled in their house work had an average of 39.2 kilometers per week and a sample standard deviation of 3.2 kilometers. construct 95% confidence interval for population mean.

10.27 (a) A social psychologist, surveyed 150 top executives and founds that 42 percent of them were drug addict. Construct 99 percent confidence interval for the true proportion of executives who had drug addict.

(b) When a sample of 70 salesman was surveyed regarding the poor performance, 66 percent believed that decreased sales were due to hot temperature in summer months. Find 95 percent confidence interval for the population proportion.

10.28 In a random sample of 1000 houses in a city, 900 own T.V sets. Construct 95 percent confidence interval for the true proportion.

10.29 (a) What is the importance of choosing an appropriate sample size in a research?

(b) Given a population with standard deviation of 8.6, what sample size is needed to estimate the mean of the population within ±0.5 with 99 percent confidence?

10.30 (a) For a test market, find the sample size needed to estimate the true proportion of consumers satisfaction with α certain new product within ± 0.04 at the 90 percent confidence level. The proportion of satisfaction of the consumer is around 0.60.

(b) It knows that from past experience that standard deviation of the annual earnings of entire population (1000) of these graduates is about Rs.15000. How large a sample size should the university take in order to estimate the mean annual earnings of last year class within Rs.5000 and at a 95 percent confidence level?

10.31 SELECT THE CORRECT ANSWER:

(i) **A numerical quantity calculated from the sample is called**
- (a) Census
- (b) Parameter
- (c) Statistic
- (d) None of these

(ii) **All the possible shoes made in service factory are the example of:**
- (a) Finite population
- (b) Infinite population
- (c) Both (a) and (b)
- (d) None of these

(iii) **The latest Census in Pakistan was held in:**
- (a) 1981
- (b) 1991
- (c) 1998
- (d) 2018

(iv) **Each and every unit of population is enumerated in:**
- (a) Census
- (b) Sampling
- (c) Both (a) and (b)
- (d) None of these

(v) **The difference between statistic and parameter is called:**
- (a) Bias
- (b) Sampling error
- (c) Standard error
- (d) Precision

(vi) **If the sample size is increased, then standard error will be:**
- (a) Increased
- (b) Decreased
- (c) No effect
- (d) None of these

(vii) **The collection of all the elements in a group is called:**
- (a) Population
- (b) Data
- (c) Sample
- (d) Registration

(viii) **A small representative part of the population is called:**
- (a) Population
- (b) Secondary Data
- (c) Sample
- (d) None of these

(ix) **A sample is usually selected by:**
- (a) With replacement
- (b) Without replacement
- (c) Both (a) and (b)
- (d) None of these

(x) **Pakistan cricket team can be selected as:**
- (a) Simple random sampling
- (b) Cluster sampling
- (c) Stradified sampling
- (d) Judgment sampling

(xi) **Draws of prize bond uses:**
- (a) Simple random sampling
- (b) Systematic sampling
- (c) Cluster sampling
- (d) Judgment sampling

(xii) **A sampling unit can be selected more than once in:**

(a) With replacement (b) Without replacement

(c) Both (a) and (b) (d) Quota sampling

(xiii) **Standard deviation of the sampling distribution of statistic is called:**

(a) Sampling error (b) Standard error

(c) Bias (d) Non sampling error

(xiv) **An other name of Probability sampling is:**

(a) Random Sampling (b) Judgment sampling

(c) Quota Sampling (d) None of these

ANSWERS 10.31

(i)	(c)	(ii)	(a)	(iii)	(d)	(iv)	(a)	(v)	(b)
(vi)	(b)	(vii)	(a)	(viii)	(c)	(ix)	(c)	(x)	(d)
(xi)	(a)	(xii)	(a)	(xiii)	(b)	(xiv)	(a)		

SHORT QUESTION ANSWERS

1. **Define population.**

Ans: The total number of objects under consideration is called population. For example number of students in a college, the number of books in a library etc.

2. **Define sample.**

Ans: Sample is the small part of population which represents the whole population.

3. **What do you understand by sampling?**

Ans: The process of selecting a sample from the population is called sampling.

4. **What are the advantages of sampling?**

Ans: The important advantages of sampling are as under:

(i) A sample survey will usually less costly than complete enumeration.

(ii) It saves a lot of time and energy.

(iii) The results obtained from a sample can be carefully planned and are expected to be more accurate.

(iv) It provides a valid measure of reliability for the sample estimates.

5. **Define sampling frame.**

Ans: A sampling frame is the complete list of all the elements in the population. e.g. A list of households in a city, telephone directory etc.

6. What is the difference between parameter and statistic?

Ans: Parameter:

A numerical quantity which is calculated from the population is called parameter. e.g. mean (μ), standard deviation (σ).

Statistic:

A numerical quantity, which is calculated from the sample is called statistic. e.g. mean ($\overline{x}$), variance (s).

7. Differentiate between sampling with and sampling without replacement.

Ans: Sampling with Replacement.

When any sampling unit is drawn from the population and then returned to the population before another unit is drawn is called sampling with replacement.

Sampling without Replacement:

When any sampling unit is drawn from the population and it is not replaced before the selection of next unit, then the sampling is called sampling without replacement.

8. What is the difference between sampling and non-sampling errors?

Ans: Sampling Error:

Sampling error is the difference between the value of the sample and true value of the population parameter.

i.e. $\overline{x} - \mu$

Non-sampling Error:

These errors are not due to sampling techniques, but they occur during the process of collection of data.

9. Define probability sampling.

Ans: When each unit in a population has known probability of being included in the sample, then sampling is said to be probability sampling.

Major types of probability sampling are:

(i) Simple Random Sampling

(ii) Stratified Random Sampling

(iii) Systematic sampling

(iv) Cluster sampling

10. Define Non-probability sampling.

Ans: A process in which the personal judgment determines which unit of the population is selected for a sample is called non-probability sampling.

Define sampling distribution.

Ans: The probability distribution of the statistic $(\bar{x}, \hat{P})$ is called sampling distribution.

12. What are the properties of sampling distribution of the mean?

Ans: (i) $\mu_{\bar{x}} = \mu$ (ii) $Var(\bar{x}) = \frac{\sigma^2}{n}$ (W.R)

$$Var(\bar{x}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \text{ (W.O.R)}$$

13. What are the properties of sampling distribution of sample proportion?

Ans: (i) $\mu_{\hat{P}} = P$ (ii) $Var(\hat{P}) = \frac{Pq}{n}$ (W.R)

$$Var(\hat{P}) = \frac{Pq}{n}\left(\frac{N-n}{N-1}\right) \text{ (W.O.R)}$$

14. Define Estimation.

Ans: The art of drawing conclusions or inferences about the population from the limited information contained in the sample is called estimation.

15. Define confidence interval.

Ans: A confidence interval is a range of values within which the population parameter is expected to occur. The probability attached to the confidence interval is called confidence coefficient or level of confidence.

CHAPTER # 11

HYPOTHESIS TESTING

11.1 INTRODUCTION

Hypothesis testing begins with an assumption called a hypothesis that we make about a population parameter. Then we collect data, produce sample statistics, and use this information to decide how likely it is that our hypothesized population parameter is correct. For example, we assume a certain value for a population mean. To test the validity of our assumption, we gather sample data and determine the difference between the hypothesized value and actual value of the sample mean. Then we judge whether the difference is significant or insignificant. The smaller the difference, the greater the likelihood that our hypothesized value for the mean is correct. The larger the difference, the smaller the likelihood.

First we shall discuss some basic concepts about testing of hypothesis.

11.2 STATISTICAL HYPOTHESIS

Statement or an assumption which may or may not be true is called statistical hypothesis.

11.2.1 Hypothesis Testing

It is a procedure which enables us to decide on the basis of sample information whether to accept or reject a statement or an assumption about the value of population parameter.

11.2.2 The Null and the Alternative Hypothesis

Any hypothesis which is to be tested for possible rejection under the assumption that it is true is called Null hypothesis. It is denoted by H_o. For example, we suppose that average weight of college students is 60 kg. This statement is taken as a hypothesis and is written as,

$$H_o : \mu = 60 \text{ kg}.$$

Any hypothesis that is different from the null hypothesis is called an alternative hypothesis. It is denoted by H_1.

For example, if our null hypothesis is

$$H_0 : \mu = 60 \text{ kg}$$

then the alternative hypothesis will be three types such as

$$H_1 : \mu \neq 62$$

$$H_1 : \mu > 62$$

$$H_1 : \mu < 62$$

11.2.3 Simple and Composite Hypothesis

If a hypothesis specifies the value of the parameters is called simple hypothesis. For example, if the weight of college students are normally distributed with $\sigma = 3$, the hypothesis is that its mean μ is say 62 kg. The above hypothesis is a simple hypothesis, as the mean and variance together specify a normal distribution completely.

If a hypothesis does not specifies the value of the parameters is called composite hypothesis.

For example, $H : \mu > 62$ or $H : \mu < 62$

11.2.4 Level of Significance

It is the probability of rejecting the null hypothesis, H_0, when H_0 is assumed to be true. It is denoted by α.

11.2.5 Test Statistic

A statistic which provides a basis for testing the null hypothesis is called test statistic. The commonly used test statistic are Z, t, x^2 and F.

11.2.6 Acceptance or Rejection Region

The results obtained from the test statistic can be divided into two groups.

(i) The results on the basis of which we accept the null hypothesis.

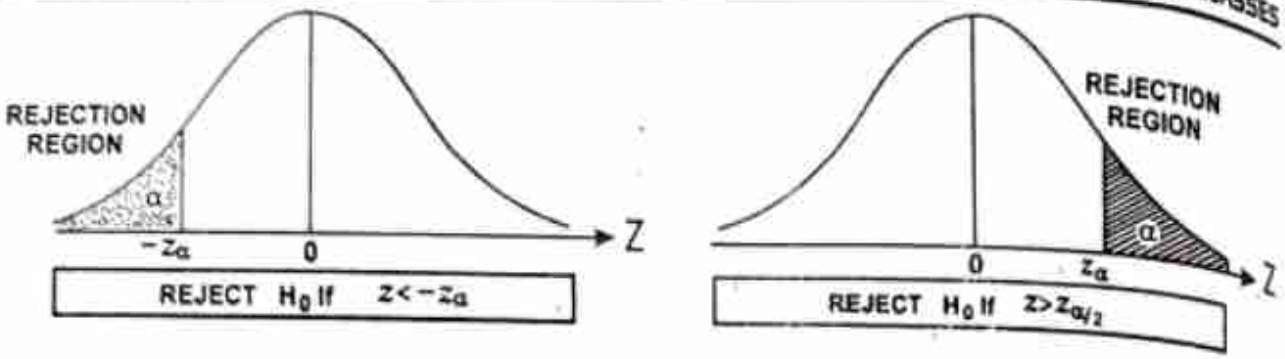
(ii) The results on the basis of which, we reject the null hypothesis.

The first group is called acceptance region and the second group is called rejection region or critical region.

11.2.7 One Tailed and Two Tailed Test

(a) One Tailed Test

A test for which the entire rejection is located at one side, either in the right tail or in the left tail of the sampling distribution of the test statistic is called one tailed test.



Fig

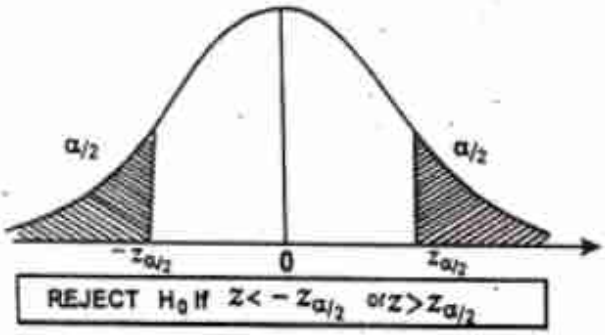
(b) Two Tailed Test

A test for which the rejection region is divided equally between the two tails of the sampling distribution of the test statistic is called two tailed or two sides test.

For example

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$



Fig

11.2.8 Type I and Type II Errors

If we reject the null hypothesis H_0, when it is in fact, true is called type I error.

i.e. Type I error : Reject H_0, when H_0 is true.

For example, an innocent person may be hauled up by the police, a deserving player may not be selected in the team etc.

If we accept the null hypothesis when it is in fact false is called an error of second type. For example, a weak student is passed by an examiner.

The decisions and the corresponding two types of errors may display in the tabular form as below.

True Situation	Decision	
	Accept H_0	Reject H_0 (or accept H_1)
H_0 is true	Correct decision (No error)	Wrong decision (Type I error)
H_0 is false	Wrong Decision (Type - II error)	Correct Decision (No error)

11.2.9 General procedure for testing hypothesis

(i) State your problem and formulate an approximate null and alternative hypothesis.

(ii) Decide upon a significance level, α.

(iii) Choose an appropriate test statistic.

(iv) Compute the value of test statistic from the sample data.

(v) Determine the rejection or critical region.

(vi) Formulate the decision rule.

(a) Reject the null hypothesis H_0, if the calculated of the test statistic falls in the rejection region and conclude that H_1 is true.

(b) Accept the null hypothesis H_0, if the calculated value of the test statistic falls in the acceptance region.

11.2.10 Critical Values of Z

Significance Level	Two-Tailed Test ($Z_{\alpha/2}$)	One Tailed Test (Z_α)
0.10	1.645	1.28
0.05	1.96	1.645
0.01	2.58	2.33

11.3 TESTING HYPOTHESIS ABOUT MEAN WHEN σ (POPULATION STANDARD DEVIATION) KNOWN

Suppose a random sample of size n is drawn from a normal population with mean μ and known standard deviation σ. The sample mean is given by $\bar{x}$. The testing procedure is given below.

(i) Formulate the null and alternative hypothesis. There are three possibilities.

(a) $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$

(b) $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$

(c) $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$

(ii) Decide on the significance level (α)

i.e. $\alpha = 0.05$

(iii) Test Statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(iv) Calculation

Calculate the value of test statistic.

(v) Determine the rejection region or critical region.

When the alternative hypothesis is	The rejection region will be
(a) $H_1 : \mu \neq \mu_0$	(a) $\lvert Z \rvert \geq Z_{\alpha/2}$
(b) $H_1 : \mu > \mu_0$	(b) $Z > Z_\alpha$
(c) $H_1 : \mu < \mu_0$	(c) $Z < -Z_\alpha$

(vi) Conclusion

Reject H_0, when calculated value of Z falls in the rejection region, otherwise accept H_0.

EXAMPLE 11.1

A sample of 400 male students is found to have a mean height of 67.47 inches. Can it be regarded as a sample from a large population with mean height 67.39 inches and standard deviation 1.3 inches. Use $\alpha = 0.05$.

SOLUTION

$n = 400,\ \bar{x} = 67.47,\ \mu = 67.39,\ \sigma = 1.3,\ \alpha = 0.05$

(i) $H_0 : \mu = 67.39$

$H_1 : \mu \neq 67.39$ (Two sides test)

(ii) $\alpha = 0.05$

(iii) $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(iv) $Z = \dfrac{67.47 - 67.39}{\dfrac{1.3}{\sqrt{400}}} \Rightarrow = \dfrac{0.08}{\dfrac{1.3}{20}} = \dfrac{0.08}{0.065} = 1.23$

(v) Critical region

$|Z| \geq Z_{\alpha/2}$

$|Z| \geq 1.96$

(vi) Conclusion

Since calculated value of $Z = 1.23$ is less than table value, so we accept H_o and conclude that $\mu = 67.39$ inches.

EXAMPLE 11.2

A random sample of size 36 is taken from a normal population with known variance $\sigma^2 = 36$. If the mean of the sample is $\bar{x} = 42.6$, test the null hypothesis $\mu = 45$ against the alternative hypothesis $\mu < 45$ with $\alpha = 0.05$.

SOLUTION

$n = 36,\ \sigma^2 = 36,\ \bar{x} = 42.6,\ \mu = 45,\ \alpha = 0.05$

(i) $H_o : \mu = 45$ (ii) $\alpha = 0.05$

$H_1 : \mu < 45$

(iii) $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

(iv) $Z = \dfrac{42.6 - 45}{\dfrac{6}{\sqrt{36}}} \Rightarrow = \dfrac{-2.4}{1} = -2.4$

(v) Critical region

$Z < -Z_\alpha \Rightarrow Z < -1.645$

(vi) Conclusion

Since calculated value of Z falls in the rejection region, so we reject H_o and accept H_1.

11.3.1 Testing Hypothesis about Mean of a Normal Population, when σ is unknown and $n \geq 30$

When the population standard deviation σ is not known, we use the sample standard deviation S instead of σ. For a large sample size ($n \geq 30$), the central limit theorem allows us to assume that the sampling distribution of $\bar{x}$ is approximately normal with a mean of μ and standard deviation of $\dfrac{S}{\sqrt{n}}$, the rest of the procedure is the same.

EXAMPLE 11.3

A random sample of 36 drinks from a soft drink machine has an average content 7.4 ounces with a standard deviation of 0.48 ounces. Test the hypothesis $\mu = 7.5$ ounces against the alternative hypothesis $\mu < 7.5$ at the 0.01 level of significance.

SOLUTION

$n = 36,\ \bar{x} = 7.4,\ S = 0.48,\ \mu = 7.5,\ \alpha = 0.01$

(i) $H_o : \mu = 7.5$ (ii) $\alpha = 0.01$

$H_1 : \mu < 7.5$

(iii) $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(iv) $$Z = \frac{7.4 - 7.5}{\frac{0.48}{\sqrt{36}}} \Rightarrow = \frac{-0.1}{0.08} = -1.25$$

(v) Critical region

$Z < -Z_\alpha \Rightarrow Z < -2.33$

(vi) Conclusion

Since calculated value of Z is greater than table value, so we accept H_o.

11.3.2 Testing Hypothesis about mean, when population standard deviation (σ) is unknown and $n < 30$.

(Small sample test of mean)

The procedure for testing the hypothesis is given below.

(i) (a) $H_o : \mu = \mu_o$, and $H_1 = \mu \neq \mu_o$ (Two sided test)

(b) $H_o : \mu \leq \mu_o$, and $H_1 = \mu > \mu_o$ (One sides test)

(c) $H_o : \mu \geq \mu_o$, and $H_1 = \mu < \mu_o$ (One sided test)

(ii) Decide upon the significance level, α

(iii) Test Statistic

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}} \quad \text{with } \upsilon = (n - 1) \text{ degree of freedom}$$

(iv) Calculate the t value.

(v) Critical Region

For (a) $|t| \geq t_{\alpha/2}(\upsilon)$

For (b) $t \geq t_{\alpha}(\upsilon)$

For (c) $t \leq -t_{\alpha}(\upsilon)$

(vi) Conclusion

Reject H_o, if the calculated value of t falls in the critical region, otherwise accept H_o.

EXAMPLE 11.4

Given a sample mean of 83, a sample standard deviation of 12.5, and a sample size of 22, test the hypothesis that the value of population mean is 70 against the alternative that is more than 70 use 5% level of significance.

SOLUTION

$\bar{x} = 83, n = 22, s = 12.5, \mu = 70, \alpha = 0.05$

(i) $H_o : \mu = 70$ (ii) $\alpha = 0.05$

$H_1 : \mu > 70$

(iii) $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

With $\upsilon = n - 1$ degree of freedom

i.e. $\upsilon = 22 - 1 = 21$, d.f

(iv) $t = \dfrac{83 - 70}{\dfrac{12.5}{\sqrt{22}}} \Rightarrow = \dfrac{13}{2.67} = 4.878$

(v) Critical region

$t > t_{\alpha(\upsilon)}$

$t > t_{0.05(21)} \Rightarrow t > 1.721$

(vi) Conclusion

Since calculated value of t = 4.878 falls in the critical region, so we reject H_o.

EXAMPLE 11.5

Nine individuals are chosen at random from a normal population and their heights in inches are found to be 63, 64, 66, 67, 67, 68, 69, 69, 70. In the light of the data, test the hypothesis that the mean height of the population is 66 inches. Using 5% level of significance

SOLUTION

(i) $H_0 : \mu = 66$

$H_1 : \mu \neq 66$

(ii) $\alpha = 0.05$

(iii) $t = \dfrac{\overline{X} - \mu}{s / \sqrt{n}}$

with $\upsilon = n - 1$ degree of freedom

i.e. $\upsilon = 9 - 1 = 8$, d.f

x	x^2
63	3969
64	4096
66	4356
67	4489
67	4489
68	4624
69	4761
69	4761
70	4900
603	**40445**

(iv) Calculation

$$\bar{x} = \frac{\Sigma x}{n} = \frac{603}{9} = 67$$

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Or $$s^2 = \frac{1}{n - 1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]$$

$$= \frac{1}{8}\left[40445 - \frac{(603)^2}{9}\right] \Rightarrow = \frac{1}{8}[40445 - 40401]$$

$$= 5.5$$

$$s = \sqrt{5.5} = 2.345$$

$$t = \frac{67 - 66}{\frac{2.345}{\sqrt{9}}} = \frac{1}{0782} = 1.279$$

(v) Critical Region

$|t| \geq t_{\alpha/2\,(v)}$

$|t| \geq t_{0.025\,(8)}$

$|t| \geq 2.306$

(vi) Conclusion

Since the calculated value of t falls in the acceptance region. So we accept H_o, we therefore conclude that the mean height of the population is 66 inches.

11.4 TESTING HYPOTHESIS ABOUT DIFFERENCE BETWEEN TWO MEANS WITH σ_1 AND σ_2 KNOWN.

(i) Null and alternative Hypothesis

(a) $H_o = \mu_1 - \mu_2 = \Delta_o$ and $H1 = \mu_1 - \mu_2 \neq \Delta_o$

(b) $H_o = \mu_1 - \mu_2 \leq \Delta_o$ and $H1 = \mu_1 - \mu_2 > \Delta_o$

(c) $H_o = \mu_1 - \mu_2 \geq \Delta_o$ and $H1 = \mu_1 - \mu_2 < \Delta_o$

(ii) Level of significance, α

(iii) Test Statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_o}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(iv) Calculation

Calculate the value of Z.

(v) Critical Region

For (a) $|Z| \geq Z_{\alpha/2}$

For (b) $Z > Z_\alpha$

For (c) $Z. < -Z_\alpha$

(vi) Conclusion

Reject H_o, if the calculated of Z falls in the rejection region.

11.4.1 Testing Hypothesis about difference two means when σ_1 and σ_2 both unknown, but $n_1, n_2 \geq 30$ (Large sample size)

If the population standard deviation σ_1 and σ_2 are unknown we use sample standard deviation S_1 and S_2, provided both the sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$).

Then the test statistic becomes

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The rest of the procedure is the same.

EXAMPLE 11.6

Two independent samples of observation were collected. For the first sample of 60 elements the mean was 86 and the standard deviation is 6. The second sample of 75 elements had a mean 82 and a standard deviation of 9. Using $\alpha = 0.05$ test whether the two samples can reasonably be considered to have come from populations with the same mean.

SOLUTION

$n_1 = 60,\ \bar{x}_1 = 86,\ S_1 = 6,\ \alpha = 0.05$

$n_2 = 75,\ \bar{x}_2 = 82,\ S_2 = 9$

(i) $H_0 : \mu_1 = \mu_2$ (ii) $H_1 : \mu_1 \neq \mu_2$

(iii) $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

(iv) Calculation

$$Z = \frac{86 - 82 - (0)}{\sqrt{\frac{6^2}{60} + \frac{9^2}{75}}} \Rightarrow Z = \frac{86 - 82}{1.296} = \frac{4}{1.296} = 3.09$$

(v) Rejection Region

$|Z| \geq Z_{\alpha/2}$

$|Z| \geq 1.96$

(vi) Conclusion

Since calculated value of $Z = 3.09$ falls in the rejection region. So we reject H_0. It is reasonable to conclude that the two samples come from different populations.

11.4.1 Testing Hypothesis about difference of means of two normal populations when σ and σ_2 unknown and (n_1 and n_2 both ≤ 30) Small sample sizes

Testing Procedure

(i) Null and alternative hypothesis

(a) $Ho = \mu_1 - \mu_2 = \Delta_o$, $H_1 = \mu_1 - \mu_2 \neq \Delta_o$

(b) $Ho = \mu_1 - \mu_2 \leq \Delta_o$, $H_1 = \mu_1 - \mu_2 > \Delta_o$

(c) $Ho = \mu_1 - \mu_2 \geq \Delta_o$, $H_1 = \mu_1 - \mu_2 < \Delta_o$

(ii) Level of significance

$\alpha = 0.05$ or 0.01

(iii) Test Statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_o}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

With $\upsilon = (n_1 + n_2 - 2)$ degrée of freedom

(iv) Calculation

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1}, \quad \bar{x}_2 = \frac{\Sigma x_2}{n_2}$$

$$s^2_p = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

or $$s^2_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

or $$s^2_p = \frac{\left[\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}\right] + \left[\Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}\right]}{n_1 + n_2 - 2}$$

(v) Critical Region

For (a) $|t| \geq t_{\alpha/2(\upsilon)}$ For (b) $t > t_{\alpha(\upsilon)}$

For (c) $t < -t_{\alpha(\upsilon)}$

(vi) Conclusion

Reject H_o, if the calculated value of t, falls in the critical region, otherwise accept H_o.

EXAMPLE 11.7

The following data are the gains in pounds of babies from birth to age six months. All babies in both groups weighted approximately the same at birth. the babies in sample I were fed formula A and babies in sample II were fed formula B.

Sample I	5	7	8	9	6	7	10	8	6
Sample II	9	10	8	6	8	7	9	---	---

Test the 5% level of significance that the mean of population I equals mean of II.

SOLUTION

(i) $H_o : \mu_1 = \mu_2$ (ii) $\alpha = 0.05$

$H_1 : \mu_1 \neq \mu_2$

(iii) Test Statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

With $\upsilon = n_1 + n_2 - 2$ degree of freedom *i.e.* $9 + 7 - 2 = 14$ d.f.

x_1	x_2	x_1^2	x_2^2
5	9	25	81
7	10	49	100
8	8	64	64
9	6	81	36
6	8	36	64
7	7	49	49
10	9	100	81
8	---	64	---
6	---	36	---
66	**57**	**504**	**475**

(iv) Calculation

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{66}{9} = 7.333 \quad , \quad \bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{57}{7} = 8.143$$

$$s^2_p = \frac{\left[\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}\right] + \left[\Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}\right]}{n_1 + n_2 - 2}$$

$$= \frac{\left[504 - \frac{(66)^2}{9}\right] + \left[475 - \frac{(57)^2}{7}\right]}{9 + 7 - 2}$$

$$= \frac{(504 - 484) + (475 - 464)}{14}$$

$$= \frac{(20 + 10.857)}{14} = \frac{30.857}{14} = 2.204$$

$$s_p = \sqrt{2.204} = 1.4846$$

$$t = \frac{7.333 - 8.143}{1.4846\sqrt{\frac{1}{9} + \frac{1}{7}}}$$

$$= \frac{-0.810}{1.4846\sqrt{0.111 + 0.1429}}$$

$$= \frac{-0.810}{1.4846\,(0.5038)} = \frac{-0.81}{0.748} = -1.083$$

(v) Critical Region

$|t| \geq t_{\alpha/2}$

$|t| \geq t_{0.025(14)}$, $\quad |t| \geq 2.145$

(vi) Conclusion

Since calculated value of t does not fall in the critical region. So we accept H_o, i.e. the two sample means are equal.

11.4.3 Testing Hypothesis about two means with dependent samples (Paired Observations)

In testing hypothesis about two means, we have used independent samples, but there are many situations in which the two samples are not independent. This happens when the observations are found in pairs as two pairs are related to each other. For example, if we want to observe the effect of new diet on 15 animals the weights before the diet and after observing the effect of the diet observations in the two samples made on the same animals are related and hence form a pair. The testing procedure is given below.

(i)

(a) $H_o = \mu_D = 0$ and $H_1 = \mu_D \neq 0$

(b) $H_o = \mu_D \leq 0$ and $H_1 = \mu_D > 0$

(c) $H_o = \mu_D \geq 0$ and $H_1 \ \mu_D < 0$

(ii) Level of significance, α

(iii) Test Statistic

$$t = \frac{\bar{d} - \mu_D}{\frac{sd}{\sqrt{n}}}, \text{ with } \upsilon = (n-1) \text{ d.f}$$

(iv) Calculation

$$d = x_1 - x_2 \quad , \quad \bar{d} = \frac{\Sigma d}{n}$$

$$s^2_d = \frac{\Sigma(d - \bar{d})^2}{n - 1}$$

$$s^2_d = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]$$

(v) Critical Region

For (a) $|t| \geq t_{\alpha/2(\upsilon)}$

For (b) $t > t_{\alpha(\upsilon)}$

For (c) $t < -t_{\alpha(\upsilon)}$

(vi) Conclusion

Reject H_o, if calculated value of t falls in the rejection region, otherwise accept H_o.

EXAMPLE 11.8

Haemoglobin values were determined an six patients before starting and after 2 weeks on B_{12} therapy. The following data were obtained.

Number	Haemoglobin (gm) before therapy	Haemoglobin (gm) after therapy
1	12.2	13.0
2	11.3	13.4
3	14.7	16.0
4	11.4	13.6
5	11.5	14.0
6	12.7	13.8

Do the data indicate a significant improvement?

SOLUTION

(i) $H_0 : \mu_D \leq 0$ (ii) $\alpha = 0.05$

$H_1 : \mu_D > 0$

(iii) $$t = \frac{\bar{d} - \mu_D}{\frac{sd}{\sqrt{n}}}$$

With $\upsilon = (n - 1)$ degree of freedom

i.e. $\upsilon = 6 - 1 = 5$ d.f

x_1	x_2	$d = x_2 - x_1$	d^2
12.2	13.0	0.8	0.64
11.3	13.4	2.1	4.41
14.7	16.0	1.3	1.69
11.4	13.6	2.2	4.84
11.5	14.0	2.5	6.25
12.7	13.8	1.1	1.21
		10.0	**19.04**

(iv) Calculation

$$\bar{d} = \frac{\Sigma d}{n} = \frac{10.0}{6} = 1.667$$

$$s^2_d = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]$$

$$= \frac{1}{5}\left[19.04 - \frac{(10)^2}{6}\right] = \frac{1}{5}[19.04 - 16.667]$$

$$= \frac{2.373}{5} = 0.475$$

$$Sd = 0.6889$$

$$t = \frac{\bar{d} - \mu_D}{\frac{sd}{\sqrt{n}}} = \frac{1.667 - 0}{\frac{0.6889}{\sqrt{6}}} = \frac{1.667}{0.2813} = 5.927$$

(v) Critical Region

$t > t_{\alpha(v)}$

$t > t_{0.05(5)} \Rightarrow t > 2.015$

(vi) Conclusion

Since calculated value of t falls in the critical region. So we reject H_o and conclude that the data indicate a significant improvement.

11.5 TESTING HYPOTHESIS ABOUT POPULATION PROPORTION

Let $\hat{p}$ be the proportion of success in a sample of size n drawn from a binomial population having proportion P. If the sample size is sufficiently large then $\hat{p}$ will be approximately normally distributed with a mean P and standard deviation $\sqrt{\frac{P_o q_o}{n}}$ where $q = 1 - P_o$.

Testing procedure for proportion is given below.

(i) Null and alternative Hypothesis

(a) $H_o = P = P_o$; $H_1 = P \neq P_o$

(b) $H_o = P \leq P_o$; $H_1 = P > P_o$

(c) $H_o = P \geq P_o$; $H_1 = P < P_o$

(ii) Level of significance; α

(iii) Test Statistic

$$Z = \frac{\hat{p} - P_o}{\sqrt{\frac{P_o q_o}{n}}}$$

(iv) Calculation

We compute the value of Z.

$$\hat{p} = \frac{x}{n}$$

(v) Critical Region

For (a) $|Z| \geq Z_{\alpha/2}$

For (b) $Z > Z_\alpha$

For (c) $Z < -Z_\alpha$

(vi) Conclusion

Reject H_o if calculated value of Z falls in the critical region, otherwise accept H_o.

EXAMPLE 11.9

A coin is tossed 60 times and 32 heads are obtained. Discuss whether the coin may be an unbiased one.

SOLUTION

(i) $H_o : P = \frac{1}{2}$ (ii) $\alpha = 0.05$

$H_1 : P \neq \frac{1}{2}$

(iii) $$Z = \frac{\hat{p} - P_o}{\sqrt{\frac{P_o q_o}{n}}}$$

(iv) $$\hat{p} = \frac{x}{n} = \frac{32}{60} = 0.533$$

$$P_o = \frac{1}{2} = 0.5 \Rightarrow q_o = 1 - 0.5 = 0.5$$

$$Z = \frac{0.533 - 0.5}{\sqrt{\frac{0.5(0.5)}{60}}} \Rightarrow = \frac{0.033}{0.645} = 0.51$$

(v) Critical Region

$|Z| \geq Z_{\alpha/2}$

$|Z| \geq 1.96$

(vi) Conclusion

Since calculated value of Z does not fall in the rejection region. So we accept H_o and conclude that the coin is unbiased one.

11.5.1 Testing Hypothesis about Difference between two proportions.

Testing procedure for difference between two proportions are given below.

Case – I

If we wish to test the hypothesis

(i) $H_o : P_1 = P_2$

(a) $H_1 : P_1 \neq P_2$

(b) $H_1 : P_1 > P_2$

(c) $H_1 : P_1 < P_2$

(ii) $\alpha = 0.05$

(iii) $$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \quad \text{and} \quad \hat{p}_c = 1 - \hat{p}$$

Case - II

If P_1 and P_2 are different, then

$H_o = P_1 - P_2 = d_o$

(a) $H_1 = P_1 - P_2 \neq d_o$

(b) $H_1 = P_1 - P_2 > d_o$

(c) $H_1 = P_1 - P_2 < d_o$

Then test statistic becomes

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

(iv) Calculation

Calculate the value of Z

(v) Critical Region

For (a) $|Z| \geq Z_{\alpha/2}$

For (b) $Z > Z_\alpha$

For (c) $Z < -Z_\alpha$

(vi) Conclusion

Reject H_o, if calculated value of Z falls in the critical region, otherwise accept H_o.

EXAMPLE 11.10

A random sample of 200 tube lights manufactured by firm x showed 15 defective tube lights, while a random sample of 125 tube lights 10 tube lights are defective. Is there significant difference between proportions of two firms?

SOLUTION

(i) $H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

(ii) $\alpha = 0.05$

(iii) $$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(iv) Calculation

$$\hat{p}_1 = \frac{15}{200} = 0.075 \quad , \quad \hat{p}_2 = \frac{10}{125} = 0.080$$

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$$= \frac{200(0.075) + 125(0.080)}{200 + 125}$$

$$= \frac{15 + 10}{325} = \frac{25}{325} = 0.0769 \quad ; \quad \hat{q}_c = 1 - P_c = 0.9231$$

$$Z = \frac{0.075 - 0.080}{\sqrt{(0.0769)(0.9231)\left[\frac{1}{200} + \frac{1}{125}\right]}}$$

$$= \frac{-0.005}{\sqrt{0.0709\,(0.013)}} = \frac{-0.005}{0.0304} = -0.16$$

(v) Critical Region

$|Z| \geq Z_{\alpha/2}$

$|Z| \geq 1.96$

Since calculated value of Z does not fall in the critical region, so we accept H_0 and conclude that there is no difference between two firms.

11.6 TESTING HYPOTHESIS ABOUT VARIANCE

There are two sampling distributions for variance which are chi-square distribution and F distribution.

For tests concerning single population variance, Chi-square test should be used and for the tests concerning two population variances, F-test should be applied.

11.6.1 General Procedure for Testing about Variance

(i) Formulate the null and alternative hypothesis about σ^2. Three possible forms are

(a) $H_0 : \sigma^2 = \sigma_0^2$ and $H_1: \sigma^2 \neq \sigma_0^2$

(b) $H_0 : \sigma^2 \leq \sigma_0^2$ and $H_1: \sigma^2 > \sigma_0^2$

(c) $H_0 : \sigma^2 \geq \sigma_0^2$ and $H_1: \sigma^2 < \sigma_0^2$

(ii) Decide on significance level α. The commonly used values are at $\alpha = 0.05$ or $= 0.01$.

(iii) The test-statistic to be used is

$$\chi^2_{(n-1)} = \frac{nS^2}{\sigma_0^2} = \frac{\Sigma (X - \bar{X})}{\sigma_0^2}$$

which under H_0, has a chi-square distribution with (n-1) degrees of freedom.

(iv) Compute the value of Chi-square.

(v) Determine the critical region which depends on σ and the alternative hypothesis H_1.

(a) When H_1 is $\sigma^2 \neq \sigma_0^2$, the critical region is

$\chi^2 < \chi^2_{1-\alpha/2,(n-1)}$ and $\chi^2 > \chi^2_{\alpha/2,(n-1)}$ (Two-tailed test)

(b) When H_1 is $\sigma^2 > \sigma_0^2$, the critical region falls in the right tail and its value is $\chi^2 > \chi^2_{\alpha/(n-1)}$. (One-tailed test)

(c) When H_1 is $\sigma^2 < \sigma_0^2$, the critical region will be entirely in the left tail with the critical value $x^2_{1-\alpha,(n-1)}$. (One-tailed test)

(vi) **Conclusion:** Reject H_0 if calculated value of x^2 falls in the critical region.

EXAMPLE 11.11

A manufacture of car batteries claims that the life of his batteries have a standard deviation equal to 0.9 years. If a random sample of 10 of these batteries have a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ years.

SOLUTION

(i) H_0: $\sigma = 0.9$ and $H_1 = \sigma > 0.9$

(ii) The significance level is set at $\alpha = 0.05$

(iii) The test-statistic is

$\chi^2 = \frac{nS^2}{\sigma_0^2}$ has χ^2 distribution with (n-1) degrees of freedom

(iv) Calculation:

$n = 10, S = 1.2, \sigma_0 = 0.9$

$\chi^2 = \frac{10(1.2)^2}{0.9} = 17.78$

(v) The critical region is $\chi 2 > \chi^2_{0.05}(9) = 16.92$

(vi) **Conclusion:** Since calculated value of $\chi^2 = 17.78$ falls in the critical region, we therefore reject the null hypothesis.

EXAMPLE 11.12

The mean and variance of 10 boxes of barely are 14.16 and 0.092 respectively. Test the hypothesis that H_0; $\sigma^2 = 0.02$ against the alternative H_1: $\sigma^2 < 0.02$, using a 0.01 level of significance.

SOLUTION

(i) H_0: $\sigma^2 = 0.02$ and $H_1 = \sigma^2 < 0.02$

(ii) $\alpha = 0.01$

(iii) The test-statistic is $\chi^2 = \frac{nS^2}{\sigma_0^2}$, which has χ^2 distribution with (n-1) d.f.

(iv) Calculation:

$\chi^2 = \frac{10(0.092)}{0.02} = 46.0$

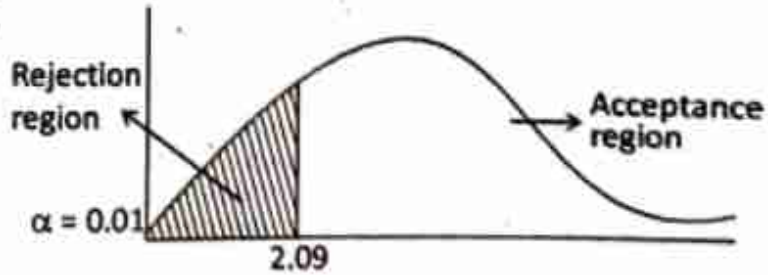
(v) The critical region is $\chi^2 = \chi^2_{(1-\alpha),(n-1)}$

$\chi^2 \leq \chi^2_{0.99}(9)$

$\chi^2 \leq 2.09$

(v) Conclusion:

Since calculated value of $\chi^2 = 46.0$ does not fall in the critical region, we therefore accept H_0



EXAMPLE 11.13

The following data pertain to weight increases (in kgs) of 7 babies: 5, 10, 8, 6, 7, 6, 4. Can these increases be reasonably regarded as a random sample from a normal population with variance 8?

SOLUTION

(i) $H_o = \sigma^2 = 8$ and $H_1 = \sigma^2 \neq 8$

(ii) $\alpha = 0.05$

(iii) Test statistic to be used

$\chi^2 = \frac{nS^2}{\sigma^2}$ which has χ^2 distribution with (n-1) degrees of freedom.

(iv) Calculation

$$S^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$$

$$= \frac{326}{7} - \left(\frac{46}{7}\right)^2$$

$$S^2 = 46.57 - 43.18 = 3.39$$

$$x^2 = \frac{7(3.39)}{8} = 2.97$$

X	X^2
5	25
10	100
8	64
6	36
7	49
8	36
4	16
46	326

(v) The critical region is $\chi^2 < \chi^2_{1-\alpha/2,(n-1)}$ and $\chi^2 > \chi^2_{\alpha/2,(n-1)}$

$x^2 < x^2_{0.795(6)}$ and $x^2 > x^2_{0.025(6)}$

$x^2 < 1.24$ and $x^2 > 14.45$

(v) Conclusion: Since calculated value does not fall in the critical region, so we accept H_o.

16.6.2 Confidence Interval Estimate of Variance

Two-sided confidence interval for σ^2 is

$$\frac{nS^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\alpha/2}}$$

Thus 95 percent confidence interval for σ^2 would be

$$\frac{nS^2}{\chi^2_{0.025}} < \sigma^2 < \frac{nS^2}{\chi^2_{0.975}}$$

EXAMPLE 11.14

Given a sample standard deviation of 6.4 from a sample of 15 observations. Construct a 90% confidence interval for the population variance.

SOLUTION

The confidence interval for σ^2 is

$$\frac{nS^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\alpha/2}}$$

90 percent confidence interval with 14 degrees of freedom is

$$\frac{15(6.4)^2}{\chi^2_{0.05(14)}} < \sigma^2 < \frac{15(6.4)^2}{\chi^2_{0.95(14)}}$$

$$\frac{614.40}{23.68} < \sigma^2 < \frac{614.40}{6.57}$$

$$25.95 < \sigma^2 < 93.52$$

11.7 TESTING HYPOTHESIS ABOUT THE EQUALITY OF TWO VARIANCES

The procedure for testing a hypothesis that a population variances σ_1^2 and σ_2^2 are equal, consisting of the following steps.

(i) Formulate the null hypothesis as H_o: $\sigma^2_1/\sigma^2_2 = 1$ (that is H_o: σ^2_1/σ^2_2). The alternative hypothesis may be

(a) H_1: $\sigma^2_1/\sigma^2_2 > 1$, or (b) H_1: $\sigma^2_1/\sigma^2_2 < 1$ or (c) H_1:$\sigma^2_1/\sigma^2_2 \neq 1$.

(ii) Decide on the significance level α.

(iii) The test-statistic to use is

$$F = \frac{s_1^2}{s^2_2}, \text{ where } s_1^2 \text{ is larger than } s^2_2.$$

which, if H_o is true has an F-distribution with υ_1 and υ_2 degrees of freedom.

(iv) Calculate the value of F from the sample data.

(v) Determine the critical region of size α from the right tail of F-distribution with υ_1 and υ_2 degrees of freedom.

(a) When H_1 is $\sigma^2_1/\sigma^2_2 > 1$ (i.e. H_1: $\sigma^2_1 > \sigma^2_2$), the critical region will be $F \geq F_{\alpha(\upsilon 1,\upsilon 2)}$.

(b) When H_1 is $\sigma^2_1/\sigma^2_2 < 1$ (i.e. H_1: $\sigma^2_1 < \sigma^2_2$), we interchange the role of two samples and use $F = s_2^2/s_1^2$, then the critical region will be $F \geq F_{\alpha(\upsilon 2,\upsilon 1)}$.

(c) When H_1 is $\sigma^2_1/\sigma^2_2 \neq 1$ (i.e. H_1: $\sigma^2_1 \neq \sigma^2_2$), the critical region will be $F \geq F_{\alpha/2(\upsilon 1,\upsilon 2)}$, when $s_1^2 > s^2_2$, or
$F \geq F_{\alpha/2(\upsilon 2,\upsilon 1)}$, when $s_2^2 > s_1^2$

(vi) Conclusion
Reject H_o, if the computed value of F falls in the critical region, otherwise accept H_o.

EXAMPLE 11.15

Two independent samples of students are taken from normal populations. The variance of marks of 10 students from class I is 64 and the variance of marks of 15 students of class 2 is 40. Check whether the variance of marks of the first sample is less than second class at 5% level of significance.

SOLUTION

$n_1 = 10$, $s_1^2 = 64$, $n_2 = 15$ and $s^2_2 = 40$

(i) $H_o = \sigma_1^2 \leq \sigma_2^2$

$H_1 = \sigma^2_1 > \sigma^2_2$

(ii) $\sigma = 0.05$

(iii) $F = \frac{s_1^2}{s^2_2}$ has F-distribution with υ_1, υ_2 d.f.

(iv) $F = \frac{65}{40} = 1.60$

(v) The critical region is $F \geq F_\alpha(\upsilon_1, \upsilon_2)$

$F \geq F_{0.05(9,14)} = 2.65$

(vi) Conclusion: Since calculated value of F = 1.60 does not fall in the critical region so we accept H_o.

EXAMPLE 11.16

Given the following information

$n_1 = 9$, $n_2 = 13$, $s_1^2 = 0.49$kg, $s_2^2 = 0.81$ kg

$\alpha = 0.01$

check the null hypothesis $H_o = \sigma_1^2 \geq \sigma_2^2$ against the alternative $H_1 = \sigma_1^2 < \sigma_2^2$.

SOLUTION

(i) $H_o = \sigma_1^2 \geq \sigma_2^2$, $H_1 = \sigma_1^2 < \sigma_2^2$

(ii) $\alpha = 0.01$

(iii) $F = \frac{s_1^2}{s_2^2}$ $(s_1^2 > s_2^2)$

has F-distribution with υ_1 and υ_2 d.f.

(iv) Calculation

Since s_2^2 is larger than s_1^2, we therefore interchange the role of two samples and use the test statistic $F = \frac{s_2^2}{s_1^2}$, with (υ_2, υ_1) degrees of freedom

$$F = \frac{0.81}{0.49} = 1.65$$

(v) The critical region is $F \geq F_{0.01(12,8)}$

$F \geq 5.67$ $(\because \upsilon_2 = 12 \quad \upsilon_1 = 8)$

(vi) Conclusion

Since calculated value of $F = 1.65$ does not fall in the critical region, so we accept the null hypothesis.

EXAMPLE 11.17

A random sample of 9 salt fish had a variance, s_1^2 in girth of 7.2, while a random sample of 8 fresh water fish had a variance, s_2^2 in girth of 3.6 (inches)^2. Test whether two different fish are equally variable.

SOLUTION

(i) $H_o = \sigma_1^2 = \sigma_2^2$ and $H_1 = \sigma_1^2 \neq \sigma_2^2$

(ii) $\alpha = 0.05$

(iii) $F = \dfrac{s_1^2}{s_2^2}$ has F-distribution with (υ_1, υ_2) d.f.

(iv) Calculation

$$F = \frac{7.2}{3.6} = 2.0$$

(v) The critical region is, $F \geq F_{\alpha/2}(\upsilon_1, \upsilon_2)$

$F \geq F_{0.025(8,7)} = 4.90$ $(\upsilon_1 = 9\text{-}1 = 8 \quad \upsilon_2 = 8\text{-}1 = 7)$

(vi) Conclusion

Since calculated value of $F = 2.0$ does not fall in the critical region, so we accept H_o and may conclude that there is no variability among two fish.

11.7.1 Confidence Interval for the Variance Ratio σ_1^2/σ_2^2

Let two independent random samples of size n_1 and n_2 be taken from two normal population with variances σ_1^2 and σ_2^2 and let s_1^2 and s_2^2 be the unbiased sample estimates of σ_1^2 and σ_2^2

Hence a $(1 - \alpha)$ 100 percent confidence interval for σ_1^2/σ_2^2 is given by

$$\left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2}(\upsilon_1, \upsilon_2)}, \quad \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2}(\upsilon_2, \upsilon_1) \right]$$

We can also find a confidence interval for σ_1/σ_2 by taking the square root of the endpoints of this interval.

EXAMPLE 11.18

Given two random samples of size $n_1 = 9$ and $n_2 = 13$ from two independent normal populations with $s_1 = 6$ and $s_2 = 5$. Find 95% confidence interval for σ_1^2/σ_2^2 and σ_1/σ_2.

SOLUTION

Given $n_1 = 9$, $n_2 = 13$, $s_1 = 6$ and $s_2 = 5$

The 95% confidence interval for σ_1^2/σ_2^2 is given by

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{0.025}(\upsilon_1, \upsilon_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} F_{0.025}(\upsilon_2, \upsilon_1) \qquad (\upsilon_1 = n_1 - 1 \quad \upsilon_2 = n_2 - 1)$$

Putting the values, we get

$$\frac{36}{25}\left(\frac{1}{3.51}\right) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{36}{25} \quad (4.20)$$

The confidence interval for σ_1/σ_2 is obtained by taking the square root of end points (0.41, 6.05).

We get

$0.64 < \sigma_1/\sigma_2 < 2.46$

SUMMARY

The formulae of testing the hypothesis are summarized below:

APPLICATION	FORMULA
1. Testing hypothesis about mean when population S.D. is known	1. $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
2. Testing hypothesis about difference between two means	2. $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n^2}}}$
3. Testing hypothesis about mean when Population standard deviation is unknown and n <3 0	3. $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ with $\upsilon = (n-1)$ d.f.
4. Testing hypothesis about difference b/w two means for small samples (σ_1, σ_2 both unknown)	4. $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ with $\upsilon = (n_1 + n_2 - 2)$ d. f
5. Testing hypothesis about proportion	5. $Z = \frac{\hat{P} - P}{\sqrt{\frac{Pq}{n}}}$
6. Testing hypothesis about difference b/w two proportions	6. $Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}_c \hat{q}_c \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
7. Testing hypothesis about two means (paired observations)	7. $t = \frac{\bar{d} - \mu d}{sd/\sqrt{n}}$
8. Testing hypothesis about variance	8. $\chi^2 = \frac{ns^2}{\sigma_0^2}$ with $\upsilon = (n-1)$ d.f
9. Testing hypothesis about the equality of two variances	9. $F = \frac{s_1^2}{s_2^2}$ with $\upsilon_1 = n_1 - 1$, $\upsilon_2 = n_2 - 1$ d.f

EXERCISES

11.1 (a) What do you meant by hypothesis testing?

(b) Explain with examples the difference between

(i) Null hypothesis and alternative hypothesis

(ii) Simple and Composite hypothesis

(iii) Acceptance region and rejection region

(iv) Type I error and Type II error

(v) One tailed and two tailed test.

11.2 What do you understand by level of significance, critical region and hypothesis testing?

11.3 (a) Describe the general procedure for testing about population mean.

(b) The average weight of a population of people is 140 lb and the standard deviation is 15 lb. 36 people are selected at random and their weight is 145 lb. Test the hypothesis that mean weight of a people are 140 lb.

11.4 (a) Newspapers press hypothesized that the average life of its press is 14,500 hours. They know that standard deviation of press life is 2100 hours. From a sample of 25 presses, the agency finds a sample mean of 13000 hours. At 0.01 significance level, should the company conclude that the average life of the press is less than the hypothesized 14500 hours?

(b) A random sample of 50 workers with children in day care centre shows the mean day-care cost Rs.2600 with a standard deviation of Rs.600. Verify the departments, claim that a mean exceeds Rs.2500 at the 0.05 level of significance.

11.5 (a) The mean life time of electric light bulbs produced by a company has in the past been 1120 hours with standard deviation of 125 hours. A sample of 100 electric bulbs recently chosen from a new supply produced bulbs should a mean life time of 1070 hours. Test the hypothesis that mean life time of bulb has not changed at 5% significance level.

(b) The average commission charge by brokerage firms on a sale of common stock is Rs.14400, and a standard deviation is Rs.5200. Company director has taken a random sample of 121 traders by his clients and determined that they paid an average commission of Rs.15100 at a 0.05 significance level, can director conclude that his clients commission are higher than industry average?

11.6 (a) A random sample of 49 drinks from a soft drink machine has an average contents 245 ml with a standard deviation of 7 ml. Test the hypothesis that an average contents 250 ml. Using 5% level of significance.

(b) A random sample of size 36 is taken from a normal population with known variance 25. If the mean of the sample is 42.6, test the hypothesis that population mean is less than 45.

11.7 (a) In what situations we use "t" distribution instead of Z-distribution?

(b) Ten individuals are chosen at random from a normal population and the heights are found to be in inches 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. In the light of these data, discuss the suggestion that the mean height in the population is 66 inches.

(c) The nine items of a sample had the following values: 45, 47, 50, 52, 48, 47, 49, 53, and 51. Does the mean of 9 items differ significantly from assumed population mean of 47.5?

11.8 A certain drug was given to ten patients and the following increments in their blood pressure was observed 7, 3, –1, 4, –3, 5, 6, –4, 1, 2. Show that the data do not indicate that the drug was responsible for these increments. Use 5% level of significance.

11.9 Given a sample mean of 94.3 a sample standard deviation of 8.4, and a sample size of 6, test the hypothesis that the value of the population mean is 100 against the alternative hypothesis that it is less than 100. Use 5% level of significance.

11.10 (a) The two samples A and B detailed below were taken from a normal population of standard deviation 0.8. Test whether the difference of means is significant.

A	10.5	11.6	12.7	12.9	13.5	13.6	14.8	---
B	11.3	12.4	12.4	13.9	14. 2	14.7	14.9	15.6

(b) Two research laboratories have independently produced drugs that provide relief to arthritis suffers. The first drug was tested on group of 90 suffers and produced an average of 8.5 hours of relief and sample standard deviation of 1.8 hours. The second drug was tested on 80 suffers, producing an average of 7.9 hours of relief and standard deviation of 2.1 hours. At 0.05 level of significance, does the second drug provide a significantly shorter period of relief?

11.11 An examination was given to two classes of 40 and 50 students respectively. In the first class mean grade were 74 with a standard deviation of 8, while in the second class the mean grade was 78 with a standard deviation of 7. Is

there a significance difference between the mean grades of students at 1% level of significance?

11.12 (a) A simple sample of heights of 6400 Englishmen has a mean of 67.85 inches and a standard deviation of 2.56 inches, while a simple sample of heights of 1600 Australian has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that Australians are on the average taller than English men? Use $\alpha = 0.05$.

(b) Two independent samples of observations were collected. For the first sample of 60 elements, the mean was 86 and the standard deviation 6. The second sample of 75 elements had a mean of 82 and standard deviation of 9. Using $\alpha = 0.01$, test whether the two samples can reasonably be considered to have come from populations with the same mean.

11.3 For a random sample of 12 horses fed on diet A, the gains in weight were 25, 30, 28, 34, 24, 25, 33, 32, 24, 30, 31 and 35 pounds. For another random sample of 15 horses fed on diet B in the same period were 44, 34, 22, 28, 47, 31, 40, 30, 32, 35, 28, 21, 35, 29 and 22 pounds. Test whether there is a significant difference between the mean gains in the two samples. use 0.05 level of significance.

11.14 (a) A consumer research organization selects several car models each year and evaluates their fuel efficiency. In this year's study of two models, the average gas mileage for 12 cars of brand A was 27.2 miles per gallon, and the standard deviation was 3.8 mpg. The nine brands B cars were tested averaged 32.1 mpg and standard deviation was 4.3 mpg. At α - 0.01 should it conclude that brand A cars have lower average gas mileage than do brand B cars?

(b) The quality control manager selects 16 bearings from the supply of vendor-1 and found that mean and variance of the diameters were 385 mm and 2.5 mm respectively. Similarly, he selected 20 bearings from vendor-2 and found that mean and variance of the diameter were 40 mm and 4 mm respectively. Test at 0.05 significance level that there is significance difference between mean diameters of bearings supplied by two vendors.

11.15 A random sample of 6 cows of breed A had daily milk yields in lbs as 16, 15, 18, 17, 19 and 17 and another 8 cows of breed B had daily milk yields in lb, as 18, 22, 21, 23, 19, 20, 24 and 21. Test if breed B is better than breed A at $\alpha = 0.05$.

11.6 (a) Distinguish between situation requiring a two sample t-test and paired-sample t-test.

(b) The following data give paired yields of two varieties of wheat. Each pair was planted in a different locality. Test the hypothesis that the mean yield is equal at 5% level of significance.

A	45	32	58	57	60	38	47	51	42	38
B	47	34	60	59	63	44	49	53	46	41

11.17 It is claimed that a new diet will reduce a person weight by at least 10 pounds on the average in a period of 2 weeks. The weight of 7 women who followed this diet was recorded before and after a 2 week period.

Women	1	2	3	4	5	6	7
Weight Before	129	133	136	152	141	138	125
Weight After	130	121	128	137	129	132	120

Test the manufacturer's claim at a 5% level of significance for the mean difference in weights.

11.18 (a) A coin is tossed 50 times and it turns up head 20 times. Discuss whether the coin may be unbiased one.

(b) In a random sample of 1000 houses in a city, 600 own T.V. Sets. Is this sufficient evidence to conclude that 70% of the houses in this city have T.V sets?

(c) A basketball player has hit on 60% of his shots from the floor. If on the next 100 shots he makes 70 baskets, would you that his shooting has improved?

(d) The sex distribution of 98 births in a hospital was 52 boys and 46 girls. Is this consistent with an equal sex division in the population?

11.19 The sales manager of a pharmaceutical company claims that a new drug cures fever in 90% of the cases. When it is administered by a doctor to 80 patients, it cured only 68 patients. Check the claim of manager at 0.05 significance level.

11.20 (a) Two different areas of large city are being considered as sites for day care centre. Of 200 households surveyed in one section the proportion in which the mothers worked full time was 0.52. In another section 40 percent of 150 households surveyed had mother working at full time job. At 5% level of significance is there significant difference in the proportion of working mothers?

(b) In a population that have minor blood disorder, sample of 100 males and 100 females are taken. It is found that 31 males and 24 females have the blood disorder. Can we conclude at 1 % level of significance that proportion of men who have blood disorder is greater than of women?

(c) A machine puts out 16 imperfect items in a sample of 500. After the machine is overhauled, it puts out 3 imperfect items in a batch of 100. Has the machine been improved?

11.21 On Saturday, 11 stocks in random sample of 40 of the roughly, 2500 stocks traded on KSE advanced; that is, their price of their shares increased. In a sample of 60 KSE stocks taken on Friday, 24 advanced. At $\alpha = 0.05$, can you conclude that a smaller proportion of KSE stocks advanced on Saturday than did on Friday?

11.22 (a) Explain how you would test the hypothesis about variance of a normal population?

(b) The variability in the tensile strength of a type of steel wire must be controlled carefully. A sample of wire is subjected to test and it is found that sample variance is 31.5. the sample size was 16. Test the hypothesis that the population variance is 25 against the alternative that the variance is greater than 25.

11.23 (a) The manger of bottling plant is anxious to reduce the variability in net volume soft drink. Over a long period, the standard deviation has been 10.0 ml. A new machine is introduced and the net volume in ml in randomly selected bottles are 240, 255, 239, 259, 250, 245, 238 and 248. Would you report to the manger that the new machine has a better performance?

(b) Given that x are normally distributed and given the sample values $\bar{x} = 42$, $S = 5$ and $n = 20$. Test the hypothesis that $\sigma = 8$.

11.24 (a) A Professor gave an examination to his class of 31 students. The mean score was 72.7 and the standard deviation was 15.9. Past experience has led the professor to believe that a standard deviation of about 13 points on a 100 points. Does this exam meet his goodness criterion?

(b) A psychologist is aware of studies showing that the variability of attention spans of 5 years-old can be summarized by $\sigma^2 = 64$ minutes squared. She wonder whether the attention span of 6-years-old is different. A sample of 20, six years-old gives $S^2 = 28$ minutes squared. Test your hypothesis at 5% significance level.

11.25 (a) Given a sample variance of 127 from a set of 9 observations, construct a 95 percent confidence interval for the population variance.

(b) In checking its cars for adherence to emission standards set by the government, an automaker measured emissions of 30 cars. The sample variance was 50. Find 95 percent confidence interval for the variance in emission particles for these cars.

11.26 (a) Describe how you would test the equality of two variances?

(b) A quality control supervisor for an automobile manufactures is concerned with uniformity in the number of defects in cars coming off the assembly. The supervisor has collected the following data.

Number of Defects		
	Assembly Line A	Assembly Line B
Mean	10	11
Variance	9	25
Sample size	20	16

Does assembly line B have significantly more variability in the number of defects? Test at the 0.5 significance level.

11.27 (a) An insurance company is interested in the length of hospital stay for various illness. The company has randomly selected 20 patients from hospital A and 25 from B who were treated for the same disease. The amount of time spent in hospital. A had an average of 2.4 days with a standard deviation of 0.6 days. The treatment time in hospital B averaged 2.3 days with a standard deviation of 0.9 day. Do patients at hospital A have significantly less variability in their recovery time. Test at a 0.01 significance level.

(b) In an experiment on reaction times in seconds two individuals A and B, measured under identical condition, the following results were obtained

A	0.41, 0.38, 0.37, 0.42, 0.35, 0.38
B	0.32, 0.36, 0.38, 0.33, 0.38

Test the hypothesis at 5 percent level of significance that $Ho = \sigma_A^2 = \sigma_B^2$ against $H_1 = \sigma_A^2 \neq \sigma_B^2$

(c) Given two random samples of $n_1 = 12$ and $n_2 = 10$ for two independent normal populations, with $s_1 = 3.2$ and $s_2 = 1.90$. Find 90% confidence interval for σ_1^2/σ_2^2 and σ_1/σ_2.

11.28 Given two random samples of size $n_1 = 10$ and $n_2 = 12$ from two independent normal population with $s_1^2 = 13.33$ and $s_2^2 = 28.55$. Test whether the two populations have the same variance.

composed by ARFAN ALI MATHEMATICIAN 03148100197

11.29 SELECT THE CORRECT ANSWER:

(i) Statistical inference is divided into two major branches given below

(a) Point estimation and hypothesis testing

(b) Interval estimation and hypothesis testing

(c) Hypothesis testing and estimation

(d) None of these

(ii) The probability of rejecting a null hypothesis is called

(a) Level of significance (b) Level of confidence

(c) Type –I error (d) Type –II error

(iii) The probability of accepting the true null hypothesis is called

(a) Level of significance (b) Level of confidence

(c) Type –I error (d) Type –II error

(iv) If a false hypothesis is accepted, it is

(a) Level of significance (b) Level of confidence

(c) Type –I error (d) Type –II error

(v) The hypothesis which is tested for possible rejection is called

(a) Simple hypothesis (b) Composite hypothesis

(c) Null hypothesis (d) Alternative hypothesis

(vi) One sided and two sided tests are based on

(a) Null hypothesis (b) Alternative hypothesis

(c) Sample size (d) level of significance

(vii) Null hypothesis always contains some sign of

(a) Equality (b) Inequality

(c) Both (a) and (b) (d) None of these

(viii) If population standard deviation is not known and sample size is less than 30, use

(a) Z-test (b) t-test

(c) x^2 test (d) None of these

(ix) In t-test we use

(a) Population variance (b) Biased sample variance

(c) Unbiased sample variance (d) None of these

(x) **For testing of hypothesis about proportion we use:**

(a) Z-test (b) t-test

(c) x^2 test (d) F test

ANSWERS 11.29

(i)	(c)	(ii)	(a)	(iii)	(b)	(iv)	(d)	(v)	(c)
(vi)	(b)	(vii)	(a)	(viii)	(b)	(viii)	(b)	(ix)	(c)
(x)	(a)								

QUESTION ANSWERS

1. **Define statistical Hypothesis.**

Ans: Statement or an assumption which may or may not be true is called statistical hypothesis.

2. **What do you mean by Hypothesis testing?**

Ans: It is procedure which enables us to decide on the basis of sample information whether to accept or reject a statement or an assumption about the value of population parameter.

3. **Differentiate between Null and alternative Hypothesis.**

Ans: **Null Hypothesis:**

Any hypothesis which is to be tested for possible rejection under the assumption that is true. It is denoted by Ho.

Alternative Hypothesis:

Any hypothesis that is different from the null hypothesis is called an alternative hypothesis.

4. **What is the difference between simple and composite hypothesis?**

Ans: **Simple Hypothesis:**

If the hypothesis specifies the value of the parameter is called simple hypothesis.

Composite Hypothesis:

If the hypothesis does not specify the value of the parameter is called composite hypothesis.

5. **Define level of significance.**

Ans: It is the probability of rejecting the null hypothesis H_o, when H_o is assumed to be true. It is denoted by α.

6. What do you know about Test Statistic?

Ans: A statistic which provides a basis for testing the null hypothesis is called test statistic. The commonly used test statistic are z, t, χ^2 and F.

7. Distinguish between one tailed and two tailed test.

Ans: **One Tailed Test:**

A test for which the entire rejection region is located at one side, either in the right tail or in the left tail of the sampling distribution of the test statistic is called one tailed test.

Two Tailed Test:

A test for which the rejection region is divided equally between the two tails of the sampling distribution of the test statistic is called two tailed test.

8. Define Acceptance and Rejection Region.

Ans: The results obtained from the test statistic can be divided into two groups:

(i) The results on the basis of which, we reject the null hypothesis.

(ii) The results on the basis of which, we accept the null hypothesis.

The first group is called rejection region and the second group is called acceptance region.

9. Differentiate between type I and type II error.

Ans: **Type I Error:**

If we reject the null hypothesis Ho, when it is in fact true is called type I error. For example an innocent person may be hauled up by the police.

Type II Error:

If we accept the null hypothesis Ho, when it is in fact false is called an error of the second type. For example a weak student is passed by an examiner.

CHAPTER # 12

ASSOCIATION OF ATTRIBUTES AND CHI SQUARE

12.1 ATTRIBUTE

We have discussed so far quantitative data only. Quantitative data consists of measurable characteristics called variables. Height, weight, temperature, price are the examples of variables. A characteristic which varies only in quality from one individual to another and cannot be measured numerically is called qualitative data or attributes. Male or female, tall or short, marital status , honesty, beauty etc are the examples of an attribute.

12.1.1 Notations and Terminology

Capital letters A, B, C, ... are used to denote the presence of the attribute, while the Greek letter α, β, γ, ... are used to denote the absence of these attributes.

Class frequencies will be denoted by enclosing the class by the symbols in brackets. Thus (A) denotes the number of objects possessing the attribute A. ($A\beta$) denote the number of objects possessing the attributes A, but not attribute B. The attributes denoted by A, B, C, ... are called positive attributes while α, β, γ, ... are negative attributes.

12.2 INDEPENDENCE

Two attributes A and B are independent.

If $$(AB) = \frac{(A).(B)}{N}$$

Suppose the class frequencies may be grouped into a table with two rows and two columns as follows.

Attribute	B	β	Total
A	(AB)	(Aβ)	(A)
α	(αB)	(αβ)	(α)
Total	**(B)**	**(β)**	**N**

When the two attributed are independent this table must have the form.

Attributes	B	β	Total
A	$\frac{(A)(B)}{N}$	$\frac{(A)(\beta)}{N}$	(A)
α	$\frac{(\alpha)(B)}{N}$	$\frac{(\alpha)(\beta)}{N}$	(α)
Total	(B)	(β)	N

12.3 ASSOCIATION OF ATTRIBUTES

Two attributes A and B are associated

If $(AB) \neq \frac{(A)(B)}{N}$

Two attributes A and B are positively associated if

$(AB) > \frac{(A)(B)}{N}$

Two attributes A and B are negatively associated if

$(AB) < \frac{(A)(B)}{N}$

12.3.1 Coefficient of Association

The strength of association between two attributes A and B is measured by a coefficient called the co-efficient of association defined by

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Co-efficient of association lies between −1 and +1. If Q = 0, the attributes are independent, when there is complete association, Q = +1 and incase of complete disassociation, the value of Q will be −1.

EXAMPLE 12.1

Test the association between injection against typhoid and exemption from attack from the following contingency table.

Attributes	Attacked	Not attacked	Total
Inoculated	528	25	553
Not-inoculated	790	175	965
Total	**1318**	**200**	**1518**

Find the coefficient of association.

SOLUTION

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{528(175) - 25(790)}{528(175) + 25(790)}$$

$$= \frac{92400 - 19750}{92400 + 19750} = \frac{72650}{112150}$$

$Q = 0.647$ (There is positive association)

12.4 TESTING HYPOTHESIS OF INDEPENDENCE IN CONTINGENCY TABLES

The data presented in a contingency table can be used to test the hypothesis that the two variables of classification are independent. If this hypothesis is rejected, then we can say there is some association between the two variables of classification.

The following steps are involved for testing hypothesis of independence.

(i) State the null and alternative hypothesis.

H_0 = The two criteria of classification are independent.

H_1 = The two criteria of classification are associated.

(ii) Choose the level of significance (α)

(iii) Test statistic to be used

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{(O_{ij} - e_{ij})^2}{e_{ij}}\right)$$

With $\upsilon = (r-1)(c-1)$ degree of freedom

r = Number of rows

c = Number of columns

(iv) Compute the expected frequencies and calculate the value of Chi square (χ^2).

(v) Determine the critical region.

$\chi^2 \geq \chi^2_{\upsilon(r-1)(c-1)}$

(vi) Draw the conclusion. Reject the hypothesis H_o if the calculated value of chi square falls in the critical region otherwise accept H_o.

EXAMPLE 12.2

The following table show the relation between the performance of students in Mathematics and Physics. Test the hypothesis that performance in Physics is independent of performance in mathematics, using 0.05 significance level.

		Mathematics			
		A grade	B grade	C grade	Σ
Physics	A grade	56	71	12	139
	B grade	47	163	38	248
	C grade	14	42	85	141
	Σ	117	276	135	528

SOLUTION

(i) H_o = The two criteria of classification are independent.

H_1 = The two criteria of classification are associated.

(ii) Level of significance

$\alpha = 0.05$

(iii) Test Statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

With $\upsilon = (r - 1)(c - 1)$ degree of freedom

i.e. $\upsilon = (3 - 1)(3 - 1) = 4$

(iv) Calculation

		Mathematics		
		A Grade	B Grade	C Grade
Physics	A Grade	$\frac{117 \times 139}{528} = 30.86$	$\frac{276 \times 139}{528} = 72.66$	$\frac{135 \times 139}{528} = 35.54$
	B Grade	$\frac{117 \times 248}{528} = 54.95$	$\frac{276 \times 248}{528} = 129.64$	$\frac{135 \times 248}{528} = 63.41$
	C Grade	$\frac{117 \times 141}{528} = 31.25$	$\frac{276 \times 141}{528} = 73.70$	$\frac{135 \times 141}{528} = 36.05$

Calculation of Chi Square

O_{ij}	e_{ij}	$O_{ij} - e_{ij}$	$(O_{ij} - e_{ij})^2$	$\frac{(O_{ij} - e_{ij})^2}{e_{ij}}$
56	30.80	25.20	635.0400	20.6182
47	54.95	−7.95	63.2025	1.1502
14	31.25	−17.25	297.5625	9.5220
71	72.66	−1.66	2.7556	0.0379
163	129.64	33.36	112.8896	8.5845
42	73.70	−31.70	1004.8900	13.6349
12	35.54	−23.54	554.1316	15.5918
38	63.41	−25.41	645.6681	10.1824
85	36.05	48.95	2396.1025	66.4661
Total	**528**	—	—	$\chi^2 = 145.7880$

(v) $\chi^2 \geq \chi^2_{0.05(4)}$

$\chi^2 \geq 9.49$

(vi) Conclusion

Since calculated value of χ^2 is greater than table value. So we reject H_0 and conclude that the two criteria of classification are associated.

EXAMPLE 12.3

Test the hypothesis that the two variables of classification are independent, using a 0.05 level of significance.

Classes	A_1	A_2	A_3	Total
B_1	337	291	302	930
B_2	225	207	238	670
Total	**562**	**498**	**540**	**1600**

SOLUTION

(i) H_o = The two criteria of classification are independents

H_1 = The two criteria of classification are not independent (associated)

(ii) Level of Significance

$\alpha = 0.05$

(iii). Test Statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

With $\upsilon = (r-1)(c-1)$ d.f

i.e. $\upsilon = (2-1)(3-1) = 2$ d.f

(iv) Calculation

	A_1	A_2	A_3
B_1	$\frac{562 \times 930}{1600} = 326.7$	$\frac{498 \times 930}{1600} = 289.5$	$\frac{540 \times 930}{1600} = 313.80$
B_2	$\frac{562 \times 670}{1600} = 235.3$	$\frac{498 \times 670}{1600} = 208.5$	$\frac{540 \times 670}{1600} = 226.2$

Calculation of χ^2

O_{ij}	e_{ij}	$O_{ij} - e_{ij}$	$(O_{ij} - e_{ij})^2$	$\frac{(O_{ij} - e_{ij})^2}{e_{ij}}$
337	326.7	10.3	106.09	0.32
225	235.3	−10.3	106.09	0.45
291	289.5	1.5	2.25	0.01
207	208.5	−1.5	2.25	0.01
302	313.8	11.8	139.24	0.44
238	226.2	−11.8	139.24	0.62
Total	**1600**	—	—	$\chi^2 = 1.85$

(v) Critical Region

$\chi^2 \geq \chi^2_{0.05\,(2)}$

$\chi^2 \geq 5.99$

(vi) Conclusion

Since calculated value of χ^2 is less than table value. So we accept H_0 and conclude that two criteria of classification are independent.

EXAMPLE 12.4

The following table shows the liking of three colours, Pink, White and blue in a sample of 200 male and female. Test whether there is any relation between sex and colour.

		Colours			
		Pink	**White**	**Blue**	**Total**
Sex	Male	20	40	60	120
	Female	40	20	20	80
	Total	60	60	80	**200**

SOLUTION

(i) H_o = The criteria of classification are independent.

H_1 = The criteria are associated.

(ii) $\alpha = 0.05$

(iii) $$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

(iv) Calculation

O_{ij}	e_{ij}	$(O_{ij} - e_{ij})$	$(O_{ij} - e_{ij})^2$	$\frac{(O_{ij} - e_{ij})^2}{e_{ij}}$
20	36	−16	−256	7.11
40	24	16	256	10.67
40	36	4	16	0.44
20	24	−4	16	0.67
60	48	12	144	3.00
20	32	−12	144	4.50
200	**200**			**26.39**

$\chi^2 = 26.39$

(v) Criteria Region

$\chi^2 \geq \chi^2_{0.05(2)}$

$\chi^2 \geq 5.99$

(vi) Conclusion

Since calculated value is greater than table value. So we reject H_o and conclude that two criteria *i.e.* sex and colours are associated.

EXERCISES

12.1 (a) What is difference between variable and attribute?

(b) State the important application of χ^2 distribution.

12.2 (a) Define association. What is the difference between correlation and association?

(b) What do you meant by independence of attributes?

12.3 The following table shows the results of an experiment to investigate the effect of vaccination of laboratory animals against particular disease. Calculate the coefficient of association.

	Got Disease	Did not got disease
Vaccinated	9	42
Not Vaccinated	17	28

12.4 100 students appeared in an examination, and 50 failed in mathematics 60 failed in statistics and 40 failed in both. Find if there is any association between the failing in mathematics and statistics.

12.5 The students of a college took three courses; arts, commerce and science. The students were classified according to the sex. The data on these students are given below.

Sex	Course of Study		
	Arts	Commerce	Science
Male	200	300	100
Female	100	200	100

Use Chi-square test whether there is any association between sex and choice of course of study. Use 5% level of significance.

12.6 In a locality 100 persons were randomly selected and asked about their educational achievement. The results are given below.

Sex	Education			
	Middle	High School	College	Total
Male	10	15	25	50
Female	25	10	15	50
Total	35	25	40	100

Use χ^2 test whether there is any association between sex and educational achievements.

12.7 A study is made to determine the possible relationship between religious affiliation and attitude towards a certain piece of legislation. The data are as under.

	For	Against	Indifferent
Catholic	60	30	20
Jewish	30	30	10
Protestant	50	60	20

Using the Chi-square, test the association when the tabulated value of chi-square for 4 degrees of freedom at 5% level of significance is 9.488.

12.8 The results of the use of two drugs in the treatment of a certain disease are as follows.

	Recovered	No change	Died
Drug A	40	18	12
Drug B	50	8	7

Test association using chi-square statistic tabulated value of χ^2 for 2 degree of freedom at 5% level of significance is 5.99.

12.9 Four hundred and ninety-two candidates for scientific posts gave particulars of their university degrees and their hobbies. The degrees were in either maths, chemistry or Physics and the hobbies could be classified roughly as music, craftswork, reading or drama. The data are classified in the following contingency table.

Categories	Maths	Chemistry	Physics	Total
Music	24	83	17	124
Craftswork	11	62	28	101
Reading	32	121	34	187
Drama	10	26	44	80
Total	**77**	**292**	**123**	**492**

Discuss the association between two criteria of classification *i.e.* the degrees and the hobbies.

12.10 Calculate chi-square from the following contingency table of attributes and test for independence at $\alpha = 0.01$.

Attributes	A_1	A_2	A_3
B_1	44	22	4
B_2	265	257	178
B_3	41	91	98

12.11 A random sample of personnel record fielded the following information about the employees tenure with company and his job classification. Test the hypothesis that the two classifications are statistically independent. Use a 0.05 level of significance.

No. of Years with Company	Job Classification			
	Prod.	Sales	Clerical	Adm.
Less than 5	25	15	20	13
5 to 10	10	8	7	10
Over 10	6	5	6	9

12.12 Given the following contingency table for hair colour and eye colour. Test the association between hair colour and eye colour at 5% level of significance.

Eye Colour	Hair Colour		
	Fair	Grey	Brown
Blue	69	49	28
Black	91	56	27
Dark Blue	57	34	33

12.13 The following table gives the condition at home and condition of the children.

Condition of Children	Condition at Home	
	Clean	Not Clean
Clean	175	143
Fairly Clean	136	116
Dirty	125	145

Test for the association between the conditions at home and condition of children.

12.14 Discuss the resemblances of the stature of parents and off-springs for the following data using chi-square at 5% level of significance.

	Tall	Medium	Short
Tall	14	125	85
Medium	3	140	165
Short	3	37	68

12.15 Discuss the association among 1000 school boys between the general ability and their mathematical ability from the following data using level of significance be 5%.

Mathematical Ability	General ability			
	Good	44	22	4
	Fair	265	257	178
	Poor	41	91	98

12.16 Test for association.

	A_1	A_2	A_3
B_1	20	15	30
B_2	30	18	35
B_3	35	20	40

Tabulated value of Chi-square for 4 degree of freedom at 5% level of significance = 9.488

12.17 The following data gives the distribution of 200 school children according to physical defect (P_1, P_2, P_3) and speech defect (S_1, S_2, S_3).

Speech defect	Physical defect			
		P_1	P_2	P_3
	S_1	24	22	19
	S_2	15	14	11
	S_3	21	28	26

Do the data suggest association between Physical defect and speech defect? Use 5% level of significance.

12.18 SELECT THE CORRECT ANSWER:

(i) **A characteristic which cannot be measured numerically is called:**

(a) Quantitative variable (b) Qualitative variable

(c) Attribute (d) Both (b) and (c)

(ii) **Which of the following is not the example of attribute?**

(a) Beauty (b) Religion

(c) Weight (d) Color

(iii) **The degree of relationship between two attributes is called:**

(a) Correlation (b) Association

(c) Regression (d) None of these

(iv) **In contingency table with "r" rows and "c" columns, the degree of freedom is**

(a) $(r-1)(c-1)$ (b) $(r-1)(c+1)$

(c) $(r+1)(c+1)$ (d) $(r+1)(c-1)$

(v) **The chi-square distribution is:**

(a) Negatively skewed (b) Positively skewed

(c) Symmetrical (d) None of these

(vi) **In 4 × 4 contingency table, the degree of freedom is**

(a) 16 (b) 12

(c) 9 (d) 4

(vii) **If A and B are independent attributes, then coefficient of association is:**

(a) 1 (b) -1

(c) -1 and 0 (d) 0

(viii) **If $(AB) = \frac{(A)(B)}{N}$, then attributes are said to be**

(a) Associated (b) Positively associated

(c) Independent (d) None of these

ANSWERS 12.18

(i)	(d)	(ii)	(c)	(iii)	(b)	(iv)	(a)	(v)	(b)
(vi)	(c)	(vii)	(d)	(viii)	(c)				

SHORT QUESTION ANSWERS

1. Define an attribute.

Ans: A characteristics which varies only in quality from one individual to another and cannot be measured numerically is called qualitative data or an attribute e.g. Honesty, beauty, religion, colour etc.

2. When two attributes are independent?

Ans: Two attributes A and B are said to independent if:

$$(AB) = \frac{(A)\,(B)}{N}$$

3. When two attributes are associated?

Ans: Two attributes A and B are associates if:

$$(AB) \neq \frac{(A)\,(B)}{N}$$

4. When two attributes are positively associated?

Ans: Two attributes A and B are positively associated if:

$$(AB) > \frac{(A)\,(B)}{N}$$

5. When two attributes are negatively associated?

Ans: Two attributes A and B are negatively associated if:

$$(AB) < \frac{(A)\,(B)}{N}$$

6. What is the coefficient of association?

Ans: The strength of association between two attributes A and B is measured by a coefficient called the coefficient of association defined by:

$$Q = \frac{(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta) + (A\beta)\,(\alpha B)}$$

7. What is the difference between correlation and association?

Ans: The relationship between two variables is called correlation, while the relationship between two attributes is called association.

CHAPTER # 13

ANALYSIS OF VARIANCE

13.1 INTRODUCTION

Earlier, we compared two population means by using a two sample t-test. In those tests, we assumed that the two population from which the samples were drawn had the same variance. In many situations we are required to compare more than two population means. In such cases we may apply the two-samples t-test to all possible pairwise comparisons of means. For example if we wish to test 5 population means, there will be $\binom{5}{2} = 10$ pairs and to test the null hypothesis that all five population means are equal, would require 10 two sample t-test. This procedure is tedious and time consuming. Hence we need another method for simultaneously comparing many means to determine whether they differ significantly. The commonly used procedure for this purpose is the analysis of variance technique.

Analysis of variance is a procedure by which the total variation in the sample data is split up into components that measure different source of variation. This technique is based on the fact that more the sample means differ, the larger the variance becomes and secondly the separate components provide independent and unbiased estimate of the common population variance. This procedure compares two different estimates of variance by using F-distribution to determine whether the population means are equal.

When each observation is classified into single sample is called a one way classification. On the other hand if each observation is classified on the basis of two criteria is called two-way classification.

13.2 ONE-WAY ANALYSIS OF VARIANCE

The one way analysis of variance is also called single-factor experiments. The data are classified into K classes on the basis of single criterion. We use the word treatment for the sample and each treatment has n replications.

Suppose we have k samples of equal size r selected randomly and independently with means $\mu_1, \mu_2, \ldots\ldots, \mu_k$ and common variance σ^2 and we wish to test the null hypothesis that all the k-population means are equal i.e

$$H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_k$$

against the alternative hypothesis

H_1: Not all means are equal

Let X_{ij} denote the ith observation of the jth sample (or treatment). Then the data can be arranged as in table below:

Observation	Samples (or Treatments)						Total
	1	2	...	j	...	k	
1	X_{11}	X_{12}		X_{1j}		X_{1k}	
2	X_{21}	X_{22}		X_{2j}		X_{2k}	
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	
i	X_{i1}	X_{i2}		X_{ij}		X_{ik}	
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	
r	X_{r1}	X_{r2}		X_{rj}		X_{rk}	
Totals	$T_{.1}$	$T_{.2}$	...	$T_{.J}$	...	$T_{.k}$	$T_{..}$
Means	$\bar{X}_{.1}$	$\bar{X}_{.2}$	...	$\bar{X}_{.j}$	...	$\bar{X}_{.k}$	$\bar{X}_{..}$

Here $\bar{X}_{.j}$, $\bar{X}_{..}$, $T_{.J}$ and $T_{..}$ represents the mean of the sample, overall or grand mean, total of observations in the jth sample and the total of all rk = n observations respectively.

The linear model on which the method of one way analysis of variance is based on

$$X_{ij} = \mu + \alpha_j + e_{ij},\ i = 1,2,\ldots\ldots n$$
$$j = 1,2,\ldots\ldots k$$

Where X_{ij} is the ith observation in the jth treatment, μ is the overall mean for all treatments, α_j is the effect of jth treatment and e_{ij} is the random error. The model is based on following assumptions.

(i) The model assumes that each observation X_{ij} is the sum of three linear components i.e μ, α_j and e_{ij}

(ii) Each of the k samples is selected randomly and independently from a normal population.

(iii) The random error e_{ij} is normally distributed random variable with $E(e_{ij}) = 0$ and $var(e_{ij}) = \sigma^2$.

Partitioning the total variability is presented as follows

Total sum of squares is $\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{..})^2$

To partition the total variation. Let us consider the following identity:

$$X_{ij} - \bar{X}_{..} = (X_{ij} - \bar{X}_{.j}) + (\bar{X}_{.j} - \bar{X}_{..})$$

Squaring both sides and summing, we get

$$\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{..})^2 = \sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{j})^2 + \sum_{j=1}^{k} \sum_{i=1}^{r} (X_{.j} - \bar{X}_{..})^2 + 2 \sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{j})(X_{.j} - \bar{X}_{..})$$

The last term on right-hand side of the equation is equal to 0 because

$$\sum_{i=1}^{r} (X_{ij} - \bar{X}_{j}) = \sum_{i=1}^{r} X_{ij} - r\bar{X}_{j} = \sum_{i=1}^{r} X_{ij} - r\left(\frac{\sum_{i=1}^{r} X_{ij}}{r}\right) = 0$$

Therefore, the total sum of squares is reduced to

$$\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{..})^2 = \sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{j})^2 + r\sum_{j=1}^{k} (X_{.j} - \bar{X}_{..})^2$$

Now we get

Total sum of square = (Within/Error *SS*) + (Between/Treatment *SS*)

i.e Total *SS* = Within SS + Between *SS*

or Total *SS* = Error *SS* + Treatment *SS*

$$\text{Mean sum of square of treatment} = \frac{\text{Treatment } SS}{\text{Degree of freedom}} = \frac{SSB}{k-1}$$

$$\text{Mean sum of square of Error} = \frac{\text{Sum of Sequares of Error}}{\text{Degree of freedom}} = \frac{SSE}{n-k}$$

Therefore F ratio becomes

$$F = \frac{\text{estimated variance from between SS}}{\text{estimated variance from within SS}}$$

which, if H_o is true, has an F-distribution with $\upsilon_1 = k\text{-}1$ and $\upsilon_2 = n\text{-}k$ degrees of freedom. It has already been stated that, when H_o is *not* true, *MSB* will be larger than σ^2, we will therefore reject H_o at the α level of significance, if $F \geq F_\alpha(\upsilon_1, \upsilon_2)$.

13.2.1 Short Cut Formulae for Computation of Sums of Squares

$$\text{Total } SS = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_i \sum_j X^2_{ij} - \frac{(T_{..})^2}{n}$$

Where $\frac{(T_{..})^2}{n}$ is called correction factor (C.F).

$$\text{Treatment SS} = r \sum_{j=1}^{k} (X_j - \bar{X}_{..})^2 = \frac{\sum_j T^2_j}{r} - \frac{(T_{..})^2}{n}$$

Analysis of Variance Table

Source of Variation	d.f	Sum of squares (SS)	Mean square (MS)	F-Ratio
Between Treatment	k-1	$r \sum_{j=1}^{k} (\bar{X}_j - \bar{X}_{..})^2 = SSB$	$s^2_b = \frac{SSB}{k\text{-}1}$	$F = \frac{s^2_b}{s^2_w}$
Within Treatment	n-k	$\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_j)^2 = SSE$	$s^2_w = \frac{SSE}{n\text{-}k}$	-
Total	n-1	$\sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_{..})^2 = SST$	-	-

EXAMPLE 13.1

Given the data below, test the hypothesis that the means of three populations are equal.

Saleman		
A	B	C
23	18	16
26	28	25
20	17	12
17	21	14

SOLUTION

(i) H_o: $\mu_1 = \mu_2 = \mu_3$

H_1: Not all means are equal

(ii) Level of significance

$\alpha = 0.05$

(iii) Test Statistic

$$F = \frac{s_b^2}{s_w^2}$$

Which if H_o is true, has F distribution with $\upsilon_1 = k\text{-}1$ and $\upsilon_2 = n\text{-}k$ degree of freedom

(iv) The necessary calculations are given below

	Saleman			
	A	B	C	Total
	23 26 20 17	18 28 17 21	16 25 12 14	
T.J	86	84	67	237 (T..)

$$\text{Total SS} = \sum_i \sum_j X_{ij}^2 - \frac{(T..)^2}{n}$$

$$\text{Correction faction (C.F)} = \frac{(T..)^2}{n} = \frac{(237)^2}{12} = 4680.75$$

$$\text{Total SS} = (23)^2 + (26)^2 + (20)^2 + \ldots\ldots + (14)^2 - 4680.75$$

$$= 4953 - 4680.75 = 272.25$$

$$\text{Between SS} = \sum_j \frac{T_{.j}^2}{r} - \text{C.F}$$

$$= \left[\frac{(86)^2 + (84)^2 + (67)^2}{4}\right] - 4680.75$$

$$= \frac{18941}{4} - 4680.75 = 54.50$$

$$\text{Within SS} = \text{Total SS} - \text{Between SS}$$

$$= 272.25 - 54.50$$

$$= 217.75$$

The analysis of variance table is

Source of Variation	d.f	Sum of squares	Mean square	F-Ratio
Between Saleman	2	54.50	27.25	1.13
Within Saleman	9	217.75	24.19	—
Total	11	272.25	—	—

(v) Critical Region

$$F \geq F_{0.05(2,9)}$$

$$F \geq 4.26$$

(vi) Conclusion

Since calculated value of F = 1.13 does not fall in the rejection region, so we accept the null hypothesis and may conclude that there is no difference among the performance of three saleman.

13.2.2 One Way Analysis of Variance (Unequal Sample Sizes)

The short cut formulae for computing Total SS and Between SS in case of unequal sample sizes are given below

$$\text{Total SS} = \sum_{j=1}^{k} \sum_{i=1}^{r_j} (X^2_{ij} - \frac{(T..)^2}{n}$$

$$\text{Between SS} = \sum_{j=1}^{k} \frac{T^2_{.j}}{r_j} - \frac{(T..)^2}{n}$$

$$\text{Within SS} = \text{TSS} - \text{BSS}$$

The rest of the procedure will be the same.

EXAMPLE 13.2

A research company has designed three different system to clean up oil spills. The following table contains the result, measured by how much surface area is cleared in one hour. Are the three system equally effective? Use the 0.05 level of significance.

System A:	55	60	63	56	59	55
System B:	57	53	64	49	62	
System C:	66	52	61	57		

SOLUTION

(i) H_0: $\mu_1 = \mu_2 = \mu_3$ and

H_1: Not all means are equal

(ii) $\alpha = 0.5$

$F = \frac{s_b^2}{s_w^2}$, which, if H_0 is true, has F distribution with $\upsilon_1 = 2$ and $\upsilon_2 = 12$ degrees of freedom

(iv)

	A	B	C	Total
	55 60 63 56 59 55	57 53 64 49 62	66 52 61 57	
$T_{.j}$	348	285	236	869

$$C.F = \frac{T^2_{..}}{n} = \frac{(869)^2}{15} = 50344.07$$

$$\text{Total SS} = \sum_i \sum_j X^2_{ij} - CF$$

$$= [(55)^2 + (60)^2 + \ldots\ldots, + (57)^2] - 50344.07$$

$$= 50665 - 50344.07 = 320.93$$

$$\text{Between SS} = \sum_j \frac{T^2_{.j}}{n_j} - C.F$$

$$= \left[\frac{(348)^2}{6} + \frac{(285)^2}{5} + \frac{(236)^2}{4}\right] - 50344.07$$

$$= 50353 - 50344.07 = 8.93$$

$$\text{Within SS} = TSS - BSS$$

$$= 320.93 - 8.93 = 312$$

ANOVA TABLE

Source of Variation	d.f	Sum of squares	Mean square	F
Between System	2	8.93	4.465	0.17
Within System	12	312	26	—
Total	14	320.93	—	—

(v) Critical Region

$F \geq F_{0.05(2,12)}$

$F \geq 3.89$

(vi) Conclusion

Since calculated value of F does not fall in the critical region. Therefore the system do not have significantly different effectiveness.

EXAMPLE 13.3

Given the following information

	Samples			
	1	2	3	4
No. of values (n)	4	6	7	3
Sample means ($\bar{x}$)	58	57	43	42
Variance (s^2)	10	30.4	5.67	9

Construct an analysis of variance table and test the hypothesis that the population means are equal at $\alpha = 0.05$

SOLUTION

(i) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: Not all means are equal

(ii) $\alpha = 0.5$

(iii) $F = \frac{S_b^2}{S_w^2}$ with $\upsilon_1 = k\text{-}1$ and $\upsilon_2 = n\text{-}k$ degrees of freedom

(iv) Calculation

$$\bar{x}_{..} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + n_4\bar{x}_4}{n_1 + n_2 + n_3 + n_4}$$

$$= \frac{4(58) + 6(57) + 7(43) + 3(42)}{4 + 6 + 7 + 3}$$

$$\bar{x}_{..} = \frac{1001}{20} = 50.05$$

Now Between SS $= \sum_j n_j (\bar{x}_j - \bar{x}_{..})^2$

$$= [4(58 - 50.05)^2 + 6(57 - 50.05)^2 + 7(43 - 50.05)^2 + 3(42 - 50.05)^2]$$

$$= 1084.95$$

Within SS $= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2$

$$= 3(10) + 5(30.4) + 6(5.67) + 2(9) = 234.02$$

ANOVA TABLE

Source of Variation	d.f	Sum of squares	Mean square	F
Between System	3	1084.95	361.65	24.72
Within System	16	234.02	14.63	-
Total	19	1318.97	-	-

(v) Critical Region:

$$F \geq F_{0.05(3,16)} = 3.24$$

(vi) Conclusion:

Since calculated value of F = 24.72 falls in the critical region, we therefore reject H_0.

13.3 TWO WAY ANALYSIS OF VARIANCE

When each observation is classified according to two criteria of classification simultaneously, we use two way analysis of variance technique. Let X_{ij} denote an observation in the ith row and j^{th} column in the table, having r rows and c columns and containing sample data from normal populations with means μ_{ij} and common variance σ^2.

	Columns					
Rows	1	2	...j...	C	Total	Means
1	X_{11}	X_{12}	...X_{ij}...	X_{1C}	$T_{1.}$	$\bar{X}_{1.}$
2	X_{21}	X_{22}	...X_{2j}...	X_{2C}	$T_{2.}$	$\bar{X}_{2.}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	X_{i1}	X_{i2}	X_{ij}...	X_{ic}	$T_{i.}$	$\bar{X}_{i.}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
r	X_{ri}	X_{r2}	X_{rj}	... X_{rc}	$T_{r.}$	$\bar{X}_{r.}$
Total	$T_{.1}$	$T_{.2}$	...$T_{.j}$	$T_{.c}$	$T_{..}$	
Means	$\bar{X}_{.1}$	$\bar{X}_{.2}$	... $\bar{X}_{.j}$	$\bar{X}_{.c}$	-	$\bar{X}_{..}$

There are now two null hypothesis i.e r-row means are equal and other-column means are equal. The two null hypothesis are

H'_o: $\mu_{1.} = \mu_{2.} = ... = ur.$

H''_o: $\mu_{.1} = \mu_{.2} = ...= u.c$

and the alternative hypothesis are

H_1': Not-all $\mu i.$ are equal

H_1'': Not all $\mu.j$ are equal

13.3.1 Partitioning Sum of Square

The estimates of variance can be obtained by partitioning the total sum of squares into three components according to three possible source of variation i.e Between rows, Between Columns and Error or within samples. For this purpose, we construct the following identity

$$Xij - \bar{X}.. = (\bar{X}i. - \bar{X}..) + (\bar{X}.j - \bar{X}..) + (Xij - \bar{X}i. - \bar{X}.j + \bar{X}..)$$

Squaring both sides and summing, we get

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\bar{X}_{..})^2=\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{X}_{i.}-\bar{X}_{..})^2+\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{X}_{.j}-\bar{X}_{..})^2+\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..})^2$$ + three cross product terms which becomes zero

The above equation becomes

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\bar{X}_{..})^2=c\sum_{i=1}^{r}(\bar{X}_{i.}-\bar{X}_{..})^2+r\sum_{j=1}^{c}(\bar{X}_{.j}-\bar{X}_{..})^2+\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..})^2$$

Total SS = SSR + SSC + SSE

The corresponding partition of the degrees of freedom is

Total d.f = Rows d.f + column d.f + Error d.f

$rc-1=(r-1)+(c-1)+(r-1)(c-1)$

Here we get four estimates of the population variance σ^2. The first estimate is based on r-1 d.f and is given by

$$s^2_1=\frac{SSR}{r-1}$$

Similarly second estimate is based on c-1 d.f

$$s^2_2=\frac{SSC}{c-1}$$

and the third estimate is based on (r-1) (c-1) d.f. and is given by

$$s^2_3=\frac{SSE}{(r-1)(c-1)}$$

To test the hypothesis that the row-means are equal, we use statistic

$$F_1=\frac{s_1^2}{s_3^2},$$

which, when the hypothesis is true, has an F-distribution with $\upsilon_1 = r-1$, $\upsilon_2 = (r-1)(c-1)$ d.f. We will reject the hypothesis at the α level of significance, when $F_1 \geq F_\alpha; [(r-1), (r-1)(c-1)]$.

Similarly, to test the null hypothesis that the column-means are equal, we compute the ration

$$F_2=\frac{s_2^2}{s_3^2},$$

which has an F-distribution with $\upsilon_1 = c-1$, $\upsilon_2 = (r-1)(c-1)$ d.f., if the null hypothesis is true. Hence we reject the null hypothesis at the α level of significance, when

$$F_2 \geq F_\alpha; [(c-1), (r-1)(c-1)].$$

Source of Variation	d.f	Sum of Squares	Mean Square	F
Between Rows	r-1	$SSR = \sum_{i}^{r} \frac{T^2_{i.}}{c} - \frac{T^2_{..}}{rc}$	$s_1^2 = \frac{SSR}{r-1}$	$F_1 = \frac{s_1^2}{s^2_3}$
Between Columns	c-1	$SSC = \sum_{i}^{c} \frac{T^2_{.j}}{r} - C.F$	$s_2^2 = \frac{SSC}{c-1}$	$F_2 = \frac{s_2^2}{s^2_3}$
Error	(r-1)(c-1)	SSE = by subtraction	$s_3^2 = \frac{SSR}{(r-1)(c-1)}$	—
Total	rc-1	$SST = \sum_{i}\sum_{j} X^2_{ij} - CF$	—	—

In the above table shortcut methods for the calculations of Total SS, RSS and CSS are used.

EXAMPLE 13.4

A certain company had 4 salesman A,B,C and D, each of whom was sent for a week into four different cities C_1, C_2, C_3 and C_4. The sales in dollars per week are given below

	Salesman			
	A	B	C	D
C1	23	30	25	28
C2	34	20	29	34
C3	45	30	45	40
C4	35	20	35	40

Carryout analysis of variance to test the hypothesis that

(a) Salesman have no influence on sales

(b) Cities have no influence on sales.

SOLUTION

(i) Hypothesis with respect to salesman

H'_o: $\mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$

H'_1: Not all columns means are equal

Hypothesis with respect to cities

H''_o: $\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$

H''_1: Not all rows columns means are equal

(ii) Level of significance: $\alpha = 0.05$

(iii) Test Statistic

$$F_1 = \frac{s_1^2}{s_3^2} \text{ and } F_2 = \frac{s_2^2}{s_3^2}$$

Which have F-distribution with $\upsilon_1 = 3$, $\upsilon_2 = 9$ and $\upsilon_1 = 3$, $\upsilon_2 = 9$ d.f respectively

(iv) Computation

The necessary calculations are given below

		Salesman				
		A	B	C	D	$T_{i.}$
Cities	C1	23	30	25	28	106
	C2	34	20	29	34	117
	C3	45	30	45	40	160
	C4	34	20	35	40	129
	T_j	136	100	134	142	512

$$\text{Total SS} = \sum_{i=1}^{4}\sum_{j=1}^{4} X^2_{ij} - \frac{(T_{..})^2}{rc}$$

$$= \left[(23)^2 + (30)^2 + (25)^2 + ... + (40)^2\right] - \frac{(512)^2}{16}$$

$$= 938$$

$$\text{Between Salesman SS} = \sum_{j=1}^{4} \frac{T_{.j}^2}{r} - \frac{(T_{..})^2}{rc}$$

$$= \left[\frac{(136)^2 + (100)^2 + (134)^2 + (142)^2}{4}\right] - \frac{(512)^2}{16}$$

$$= 270$$

$$\text{Between Cities SS} = \sum_{i=1}^{4} \frac{T^2_{i.}}{c} - \frac{(T_{..})^2}{rc}$$

$$= \left[\frac{(106)^2 + (117)^2 + (160)^2 + (129)^2}{4}\right] - \frac{(512)^2}{16}$$

$$= 407.5$$

$$\text{Error SS} = 938 - 407.5 - 270$$

$$= 260.5$$

ANOVA TABLE

Source of Variation	d.f	Sum of Square	Mean Square	F-ratio
Between Salesman	3	270.0	90.0	$F_1 = 3.11$
Between Cities	3	407.5	135.83	$F_2 = 4.69$
Error	9	260.5	28.94	
Total	15	938.0	-	-

(v) The critical region are

(a) $F_1 \geq F_{0.05}(3,9) = 3.86$

(b) $F_2 \geq F_{0.05}(3,9) = 3.86$

(vi) Conclusion:

Since the computed value of $F_1 = 3.11$ does not fall in the critical region, but the computed value of $F_2 = 4.69$ falls in the critical region so we accept the hypothesis relating to salesman and we reject the hypothesis corresponding to the cities. Therefore we conclude that there is no significant difference in terms of weekly sales made between the different salesman, but different cities have influenced significantly difference among weekly sales.

13.4 THE LEAST SIGNIFICANT DIFFERENCE (LSD) TEST

When the null hypothesis of equal means is rejected by F-test after the ANOVA, we conclude that the treatment means are not all equal. We usually interested in finding out which means differ from each other. When all samples are of the same size, a very simple method used for locating significant differences is called the least significant difference test, (LSD). the LSD is given by

$$LSD = t_{\alpha/2}\sqrt{\frac{2(MSE)}{r}}$$

Where MSE is the error means square, r is the size of sample and $t_{\alpha/2}$ is the value of t at $\alpha/2$ level taken against error d.f. Two means are considered to be significant, when the absolute value of their difference exceeds the LSD.

It is important to note that LSD test is only applied after the null hypothesis is rejected and the number of observations are the same for each treatment.

EXAMPLE 13.5

The following data represent the grades obtained by 4 students in Math, Stat and Computer. Test the hypothesis that

(a) the students have equal ability

(b) the courses are of equal difficulty

(c) if null hypothesis is rejected, determine which means are significantly different.

Student	Subjects		
	Math	Stat	Computer
1	84	92	94
2	75	77	67
3	79	86	78
4	78	77	73

SOLUTION

(i) (a) H'_o: The students have equal ability

H'_1: At least 2 students do not have equal ability

(b) H''_o: The courses are of equal difficulty

H''_1: At least 2 courses are not of equal difficulty

(ii) $\alpha = 0.05$

(iii) $F_1 = \frac{s_1^2}{s_3^2}$ and $F_2 = \frac{s_2^2}{s_3^2}$

With $\upsilon_1 = 3$, $\upsilon_2 = 6$ and $\upsilon_1 = 2$, $\upsilon_2 = 6$ d.f

(iv) Calculation

To simplify the arithmetic we subtract a suitable constant, say 80 from each observations.

Student	Math	Stat	Computer	$T_{i.}$	$\bar{X}_{i.}$
1	4	12	14	30	90
2	-5	-3	-13	-21	73
3	-1	6	-2	3	81
4	-2	-3	-7	-12	76
T.J	-4	12	-8	0	80

$$C.F = \frac{(T_{..})^2}{rc} = \frac{(0)^2}{12} = 0$$

$$\text{Total SS} = \sum_i \sum_j X^2_{ij} - CF$$

$$= \left[(4)^2 + (12)^2 + \ldots\ldots + (-7)^2\right] - 0 = 622$$

$$\text{Between Students} = \frac{\Sigma T_i^2}{c} - CF$$

$$= \left[\frac{(30)^2 + (-21)^2 + (3)^2\ (-12)^2}{3}\right] - 0$$

$$= 498$$

$$\text{Between Courses} = \frac{\Sigma T_j^2}{r} - CF$$

$$= \left[\frac{(-4)^2 + (12)^2 + (-8)^2}{3}\right] - 0 = 56$$

Error SS = TSS – Between students – Between courses

$= 622 - 498 - 56 = 108$

ANOVA TABLE

Source of Variation	d.f	Sum of Squares	Mean Square	F-ratio
Between student	3	498	166	$F_1 = 166/18 = 9.22$
Between courses	2	56	28	$F_2 = 28/18 = 1.56$
Error	6	108	18	-
Total	11	622	-	-

(v) Critical region

(a) $F_1 \geq F_{0.05(3,6)} = 4.76$

(b) $F_2 \geq F_{0.05(2,6)} = 5.14$

(vi) Conclusion:

(a) Since calculated value of $F_1 = 9.22$ falls in the critical region, we therefore, conclude that the students do not have equal ability

(b) Since calculated value of $F_2 = 1.56$ does not fall in the critical region, we do not reject Ho. We conclude that the courses are of equal difficulty.

(c) We apply the LSD test to determine which means grades of students are significantly different.

$$LSD = t_{\alpha/2}\sqrt{\frac{2(MSE)}{r}}$$

Here MSE = 18, $t_{0.025}(6) = 2.447$ and r = 3 substituting these values, we have

$$LSD = (2.447)\sqrt{\frac{2(18)}{3}} = 8.48$$

The last column of the table gives the mean grades of students

Arranging the mean grades in ascending order

$\bar{X}_{2.}$	$\bar{X}_{4.}$	$\bar{X}_{3.}$	$\bar{X}_{1.}$
73	76	81	90

The sample means which are not significantly different are underlined. Therefore significantly different means are

$(\bar{X}_{2.}, \bar{X}_{1.})$, $(\bar{X}_{4.}, \bar{X}_{1.})$ and $\bar{X}_{3.}, \bar{X}_{1.})$

SUMMARY

The brief summary of the chapter is given below

One way Analysis of Variance

Total sum of square = Within SS + Between SS

Where

$$TSS = \sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_j)^2 = \sum_j \sum_i (X^2_{ij}) - \frac{(T..)^2}{n}$$

$$\text{Within SS} = \sum_{j=1}^{k} \sum_{i=1}^{r} (X_{ij} - \bar{X}_j)^2 \text{ and}$$

$$BSS = r \sum_{j=1}^{k} (\bar{X}_j - \bar{X}..)^2 = \sum \frac{T.j^2}{r} - \frac{(T..)^2}{n}$$

$$F = \frac{\text{Means SS (Between)}}{\text{Mean SS (Within)}} = F = \frac{s^2_b}{s^2_w} \text{ with } \upsilon_1 = k\text{-}1,\ \upsilon_2 = n\text{-}k \text{ d.f}$$

Two Way Analysis of Variance

Total SS = Rows SS + Columns SS + Error SS

$$\text{Total SS} = \sum_i \sum_j X^2_{ij} - C.F, \text{ where } C.F = \frac{(T..)^2}{rc}$$

$$\text{Row SS} = \sum_{i=1}^{r} \frac{T^2_{i.}}{c} - CF$$

$$CSS = \sum_{j=1}^{k} \frac{T.j^2}{r} - CF$$

ESS = By Subtraction

$$F_1 = \frac{\text{Mean SS (Rows)}}{\text{MSS (Error)}} = \frac{s^2_1}{s^2_3} \text{ with } \upsilon_1 = (r\text{-}1), \text{ and } \upsilon_2 = (r\text{-}1)\ (c\text{-}1) \text{ df}$$

$$F_2 = \frac{\text{Mean SS (Columns)}}{\text{MSS Error}} = \frac{s^2_2}{s^2_3} \text{ with } \upsilon_1 = (c\text{-}1) \text{ and } \upsilon_2 = (r\text{-}1)\ (c\text{-}1) \text{ df}$$

LSD

$$LSD = t_{\alpha/2}\sqrt{\frac{2(MSE)}{r}}$$

Where MSE = mean square error, r is the size of sample and $t_{\alpha/2,\upsilon}$ is the value of t

EXERCISES

13.1 What is meant by analysis of variance? Also write down the assumptions underlying a one way analysis of variance.

13.2 A study compared the effects of four 1 month point of purchase promotions on sales. The unit sales of five stores using all four promotions in different months follow

Free Sample	One Pack Gift	Cent off	Refund by Mail
78	94	73	79
87	91	78	83
81	87	69	78
89	90	83	69
85	88	76	81

Test the hypothesis that there is no difference between promotions on sales.

13.3 Three training methods were compared to see whether they led to greater productivity after training. The following are productivity measures for individuals trained by each method

Method 1	45	40	50	39
Method 2	59	43	47	51
Method 3	41	37	43	40

At the 0.05 level of significance, do the three training methods lead to different levels of productivity?

13.4 Following data is about damage plants to different crops in 3 different sides. test the hypothesis that damage in all three areas are same.

Left Side	Right Side	Middle
129	301	292
14	26	19
43	166	222
62	232	263

13.5 The following data show the number of claims processed per day for a group of 4 insurance company employees observed for a number of days. Test the hypothesis that the employees mean claim per day are all the same. Use the 0.05 level of significance.

Employee 1	15	17	14	12	
Employee 2	12	10	13	17	
Employee 3	11	14	13	15	10
Employee 4	13	12	12	14	9

13.6 Given the following information

Format	n	$\bar{x}$	s^2
1	10	90	4.37
2	10	86	3.76
3	11	83	4.21

Perform the analysis of variance to determine whether all three means are equal at 0.05 level of significance.

13.7 (a) What is meant by two way analysis of variance? Sketch the table of analysis of variance and define the statistic used to test the significance of differences.

(b) Describe the Least Significance Difference (LSD) test in analysis of variance.

13.8 Four breeds of Cattle B_1, B_2, B_3 and B_4 were fed on three difference rations R_1, R_2 and R_3. Gain in weight in pounds over a given period were recorded as given below.

	B_1	B_2	B_3	B_4
R_1	46.5	62.0	41.0	45.0
R_2	47.5	41.5	22.0	31.5
R_3	50.0	40.0	25.5	28.5

At $\alpha = 0.05$, test the null hypothesis that

(a) There is no differences in mean gains in weight for the different breeds of cattle.

(b) There is no differences between rations

13.9 Four varieties of wheat are planted in a field. Each variety is treated with four different fertilizers. Yield in Kg per plot in shown, in the table given below Test the hypothesis that

(a) There is no difference in the mean yield of four varieties

(b) There is no difference among fertilizers.

	Varieties			
Fertilizer	V_1	V_2	V_3	V_4
1	2	5	4	1
2	2	3	3	1
3	4	6	6	2
4	1	4	2	3

13.10 A time and motion study in made to determine the best method of assembling a complex electronic mechanism. Three methods are under consideration. Four workers are selected at random. The out put is given below.

	Methods of Assembly		
Workers	1	2	3
A	4	8	10
B	6	8	12
C	10	14	16
D	14	18	22

(i) Test the hypothesis that average output for three methods are equal.

(ii) If null hypothesis is rejected, determine which means are significantly different.

13.11 The ANOVA from two way experiment output is shown below

Source of Variation	d.f	Sum of square	Mean square	F
Treatment	4	1020.56	?	30.14
Block	?	?	64.765	?
Error	20	169.33	?	
Total	29	1513.71	-	-

(a) Fill the missing values

(b) How many blocks are used in this experiment?

(c) What conclusion can you draw?

13.12 Construct the analysis of variance table for the following data.

	Factor A			
Factor B	1	2	3	4
1	8	5	16	10
2	13	7	15	7
3	9	6	11	4

Test the hypothesis that

(a) Factor A has no influence on yield, and

(b) Factor B has no influence on yield

(c) Apply LSD if applicable

13.13 TICK THE CORRECT ANSWER

(i) **The F-ratio contains:**

(a) Two estimates of the variance

(b) Two estimates of the mean

(c) One estimate of mean and other is variance

(d) None of these

(ii) **Which of these distributions has a pair of degree of freedom?**

(a) Poisson (b) Normal

(c) Chi-Square (d) F

(iii) **When each observation is classified into single sample, it is called:**

(a) One way ANOVA (b) Two way analysis of variance

(c) Three way (d) None of above

(iv) **In one way analysis of variance, mean sum of squares of error is equal to**

(a) $\frac{SST}{k-1}$ (b) $\frac{TSS}{n-1}$

(c) $\frac{SSE}{n-k}$ (d) $\frac{SSE}{k-1}$

(v) **In one way analysis of variance total sum of square is partitioned into:**

(a) 2 components (b) 3 components

(c) 4 components (d) 5 components

(vi) **In two way analysis of variance, the total sum of square is partitioned into:**

(a) 2 components (b) 3 components

(c) 4 components (d) None of these

(vii) The ANOVA is used for testing that:

(a) Means of 2 samples are equal

(b) Means of more than two samples are equal

(c) Proportion of two populations are equal

(d) Independence of attributes

(viii) Analysis of variance technique was introduced by:

(a) Karl Pearson (b) Galton

(c) Laplace (d) R.A. Fisher

(ix) LSD test is used only when

(a) We reject null hypothesis of equal means

(b) We accept null hypothesis of equal means

(c) We reject null hypothesis of equal variances

(d) Both (a) and (c)

Answers 13.13

(i)	(a)	**(ii)**	(d)	**(iii)**	(a)	**(iv)**	(c)	**(v)**	(a)
(vi)	(b)	**(vii)**	(b)	**(viii)**	(d)	**(ix)**	(a)		

SHORT QUESTION ANSWERS

1. What is analysis of variance?

Ans. It is a procedure by which the total variation in the sample data is split up into its components that measure the different source of variations.

2. Define one way analysis of variance.

Ans. When each observation is classified into single sample is called one way analysis of variance.

3. What do you meant by two way ANOVA?

Ans. When each observation is classified into two criteria of classification is called two way analysis of variance.

4. In what situations LSD test is used?

Ans. When the null hypothesis of equal means is rejected by F-test after ANOVA. We conclude that the treatment means are not equal. We usually interested in finding out which means differ significantly from each other. For this purpose LSD test is used. The LSD test is given by

$$LSD = t_{\alpha/2(v)}\sqrt{\frac{2MSE}{r}}$$

CHAPTER # 14

EXPERIMENTAL DESIGN

14.1 INTRODUCTION

In an experiment, we start with collection of experimental units and number of treatments. The treatments are applied on the experimental units, one treatment to a unit. The effect is measured by recording certain characteristics of the experimental unit, which is called yield of the unit. The basic purpose of the experiment is to compare the effect of different treatments.

The experimental design is an efficient procedure to conduct a planned experiment to draw valid conclusion from the given data. Experimental design are widely use in agriculture, biology, chemistry, medicine and engineering.

14.2 BASIC PRINCIPLES OF EXPERIMENTAL DESIGN

The basic principles of experimental designs are randomization, replication and local control.

1. **Randomization**

   It is a random process of assigning treatments to the experimental unit. The random process means that every possible allotment of treatments has equal probability. The purpose of randomization is to remove bias, due to personal preferences of the researcher.

2. **Replication**

   Replication is the repetition of the basic experiment i.e. we perform the experiment more than once because experimental units such as plots of land can not be physically identical. It helps in increasing the precision.

3. **Local Control**

   We need to choose a design in which all extraneous sources of variations are brought under control. For this purpose, we make use of local control. This term is used to the amount of balancing, blocking and grouping of experimental unit. Balancing means that the treatments should be assigned to experimental units in such a way that the result is balanced arrangement of treatments. A block is also a replicate. It means that like experimental units should be collected together to form a relatively homogeneous group.

14.3 BASIC EXPERIMENTAL DESIGNS

The basic experimental designs are given below

(i) Completely Randomized Design (CRD)

(ii) Randomized Complete Block Design (RCBD)

(iii) Latin Square Design (LS Design)

14.4 COMPLETELY RANDOMIZED DESIGN

A completely randomized design is the simplest experimental design. In this design the treatments are assigned to experimental units completely at random. The design is completely flexible i.e any number if treatments and any number of units per treatment may be used. This design is considered to be most useful when the experimental units are homogeneous and the experiments are small such as laboratory experiments.

14.4.1 Experimental Layout

As an example of CR design, we consider 4 fertilizers (treatments) A, B, C and D, each applied to 3 plots. The layout of this design is given below

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

14.4.2 Statistical Model and Analysis

Let Y_{ij} denote the yield of ith observation on j^{th} treatment. Then Y_{ij} may be represented by linear model.

$$Y_{ij} = \mu + \tau_j + e_{ij} \quad ; \quad i = 1, 2, 3, \ldots r$$
$$j = 1, 2, 3, \ldots k$$

Where μ represents the true mean effect, τ_j represents the effect of treatment j and e_{ij} is the random error. ($e_{ij} \sim N(0, \sigma^2)$)

The null hypothesis is

Ho: $\tau_j = 0$ against H_1: $\tau_j \neq 0$ for all $j = 1,2,3 \ldots k$

These hypothesis may be written as

Ho: $\mu_1 = \mu_2 = \ldots \mu k$, and

H1: Not all μ are equal

The data structure of yield is shown in table given below

Observation	Treatments				
	1	2	...j...	k	
1	Y_{11}	Y_{12}	...Y_{1j}...	Y_{1K}	
2	Y_{21}	Y_{22}	...Y_{2j}...	Y_{2K}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
i	Y_{i1}	X_{i2}	Y_{ij}	Y_{ik}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
r_j	Y_{r1}	Y_{r2}	Y_{rj}	$Y_{rk}k$	
Totals	T.1	T.2	T_j	$T_{.k}$	T..

This design is analysed by the same method described in the previous chapter for one way ANOVA

Total SS = Treatment SS + Error SS

$$\text{Total SS} = \sum_{j=1}^{k} \sum_{i=1}^{rj} y^2_{ij} - CF$$

$$\text{Where } CF = \frac{(T..)^2}{n}$$

$$\text{Treatment SS} = \sum_{j=1}^{k} \frac{T^2_{.j}}{r} - CF$$

Error SS = TSS – Treatment SS

The analysis of variance table for CR design is

Source of Variation	d.f	Sum of squares	Mean Square	F-ratio
Treatments	k-1	$\sum_j \frac{T^2_{.j}}{r} - CF$	$s_t^2 = \frac{SS}{k-1}$	$F = \frac{s_t^2}{s_e^2}$
Error	n-k	By subtraction	$s_e^2 = \frac{SS}{n-k}$	
Total	n-1	$\Sigma\Sigma y^2_{ij} - CF$		

$$F = \frac{\text{MS for treatment}}{\text{MS for error}} = \frac{s_t^2}{s_e^2}$$

has an F distribution with $\upsilon_1 = k-1$ and $\upsilon_2 = n-k$ degrees of freedom. The hypothesis is rejected at the α level of significance, when $F \geq F_\alpha(\upsilon_1, \upsilon_2)$. If the hypothesis is rejected then we use LSD test to pick up the best treatment.

14.4.3 Advantages and Disadvantages of CR Design

Advantages of CR design are

(i) The design is very simple

(ii) It is very easy to layout.

(iii) The analysis is very simple

(iv) The design is flexible i.e any number of treatments and replication may be used.

(v) It provides the maximum number of degrees of freedom for error sum of squares.

Some disadvantages of CR design are

(i) It can be used only when the number of treatment is small

(ii) Since the randomization is not restricted in any direction, the whole of the variation among the units enters into the experimental error.

EXAMPLE 14.1

Three fertilizers A, B, C each applied to 4 plots. the yields were as follows.

Fertilizer		
A	B	C
29	50	41
22	44	36
17	45	28
15	30	27

Test the hypothesis that the three fertilizers are not different in the yielding capabilities. Apply L.S.D if applicable?

SOLUTION

(i) Ho: $\tau_j = 0$, Hi: $\tau_j \neq 0$, OR

Ho: $T_1 = T_2 = T_3$

Hi: Not all treatments are equal

(ii) $\alpha = 0.05$

(iii) $F = \dfrac{\text{MS for Treatment}}{\text{MS for Error}}$

has an F-distribution with $\upsilon_1 = k\text{-}1$ and $\upsilon_2 = n\text{-}k$ degrees of freedoms

(iv) The necessary calculations are given below

	Fertilizers			
	A	B	C	
	29 22 17 15	50 44 45 30	41 36 28 27	
T_j	83	169	132	384
Mean	20.75	42.25	33.0	—

Total SS = $\Sigma\Sigma\, Y^2_{ij} - C.F$

$$C.F = \frac{(T..)^2}{n} = \frac{(384)^2}{12} = 12288$$

$$TSS = \left[(29)^2 + (22)^2 + ... + (27)^2\right] - 122288$$

$$= 13690 - 12288 = 1402$$

$$\text{Treatment SS} = \Sigma \frac{T^2_j}{r} - CF$$

$$= \left[\frac{(83)^2 + (169)^2 + (132)^2}{4}\right] - 12288$$

$$= 13218.5 - 12288 = 930.5$$

Error SS = TSS – Treatment SS

$$= 1402 - 930.5 = 471.5$$

Source of Variation	d.f	Sum of Squares	Mean Squares	F-ratio
Treatment	2	930.5	465.25	8.89
Error	9	471.5	52.39	-
Total	11	1402.0	-	-

(v) Critical Region:

$F \geq F_{0.05(2,9)} = 4.26$

(vi) Conclusion:

Since calculated value of F = 8.89 is greater than critical value. So we reject Ho and conclude that the three fertilizers are significantly different.

Since the treatment effects prove to be significant we therefore apply LSD test to find out which pairs of treatments means are significantly different

$$LSD = t_{\alpha/2\,(Error\ df)}\sqrt{\frac{2(MSE)}{r}}$$

$$= t_{0.05(9)}\sqrt{\frac{2 \times 52.39}{4}}$$

$$2.262 \times 5.118$$

$$= 11.58$$

Arranging the means of three fertilizers in an ascending order

A	C	B
20.75	33.0	42.25

The pair of treatments/fertilizers (A,C) and (A, B) are significantly different. How ever pair of treatment (B,C) are approximately equal.

14.5 RANDOMIZED COMPLETE BLOCK DESIGN

In this design, the experimental material is divided into groups or blocks in such a way that experimental units within each blocks are homogeneous. Every block contains a complete set of treatments. Furthermore the treatments are assigned at random to the experimental units within each block. It is most widely used experimental design. The experiments which involved two factors we use randomized complete block (RCB) design.

14.5.1 Experimental Layout

An experimental layout of such a design, using 5 treatments A, B, C, D and E in 4 blocks might be as follows.

Block I	A	C	D	B	E
Block II	B	A	C	E	D
Block III	D	C	B	A	E
Block IV	C	A	D	B	E

14.5.2 Statistical Model

Let Y_{ij} represent the ith observation on j^{th} treatment, then the linear model for RCB design would be

$$Y_{ij} = \mu + B_i + \tau_j + e_{ij}$$

Where μ = overall mean

B_i = effect of ith block

τ_j = effect of jth treatment

e_{ij} = error term and $e_{ij} \sim N(0, \sigma^2)$

14.5.3 Statistical Analysis

The analysis of RCB design consists of two way analysis of variance test where the hypothesis are

(i) Ho: $\tau_j = 0$ i.e all treatment effects are zero

and (ii) H'o: $B_i = 0$ i.e all block effects are zero

Data structure of a RCB design may be arranged in a table given below

	Columns					
Rows	1	2	...j...	C	Total	Means
1	Y_{11}	Y_{12}	Y_{1j}	X_{1k}	$B_{1.}$	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	Y_{2j}	Y_{2k}	$B_{2.}$	$\bar{Y}_{2.}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	Y_{i1}	Y_{i2}	Y_{ij}...	Y_{ik}	$B_{i.}$	$\bar{Y}_{i.}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
r	Y_{ri}	Y_{r2}	Y_{rj}	... Y_{rk}	$B_{r.}$	$\bar{Y}_{r.}$
Total	T.1	T.2	T_j	$T_{.k}$	T..	-
Means	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.j}$	$\bar{Y}_{.k}$	-	$\bar{Y}_{..}$

Total sum of squares may be partitioned into its components for treatment, blocks and error would be.

Total SS = Treatment SS + Block SS + Error SS

Total SS = $\Sigma\Sigma\, y^2_{ij}$ –CF

Where CF (Correction factor) = $\frac{(T..)^2}{n}$

Treatment SS = $\sum_j \frac{T^2_{.j}}{r}$ – CF

$$\text{Block SS} = \sum_i \frac{B^2_i}{c} - CF$$

Error SS = Total SS – Treatment SS – Block SS

ANOVA TABLE

Source of Variation	d.f	Sum of Squares	Mean Squares	F-ratio
Treatment	c-1	$\sum_j \frac{T^2_j}{r} - CF$	$\frac{SST}{c-1} = s_t^2$	$F_1 = \frac{s_t^2}{s_e^2}$
Block	r-1	$\sum_j \frac{B^2_i}{c} - CF$	$\frac{SSB}{r-1} = s_b^2$	$F_2 = \frac{s_b^2}{s_e^2}$
Error	(r-1) (c-1)	By subtraction	SSE/(r-1)(c-1)= s_e^2	
Total	rc-1	$\Sigma\Sigma Y^2_{ij} - CF$	-	-

$$F1 = \frac{\text{MS for treatment}}{\text{MS for Error}} = \frac{s_t^2}{s_e^2}$$

With

$\upsilon_1 = (c-1)$ and $\upsilon_2 = (r-1)(c-1)$ d.f

Critical Region for hypothesis are

(i) Ho: $\tau_j = 0$ is $F \geq F_{\alpha(\upsilon 1, \upsilon 2)}$ and

(ii) Ho: $B_i = 0$ is $F \geq F_{\alpha(\upsilon 1, \upsilon 2)}$

14.5.4 Advantages and Disadvantages of RCB Design

Some important advantages of RCB design are as follows

(i) It is most frequently used design.

(ii) The statistical analysis is simple.

(iii) The design is flexible i.e any number of treatments and replicates may be used.

(iv) By means of blocking / grouping, more accurate results are usually obtained than CR design.

(i) It controls variability only in one direction.

(ii) It is not suitable for large number of treatments

(iii) It is not suitable when the blocks are not homogeneous.

EXAMPLE 14.2

Three varieties A, B and C of a crop are tested in RCB design with four replicates, the layout is given below. The plot yield in pounds are also indicated therein. Analysis the experimental yields and state your conclusions.

Replications	1	A	32.1	C	34.2	B	31.7
	2	C	30.7	A	17.0	B.	32.7
	3	A	40.8	B	25.3	C	48.2
	4	B	47.9	C	59.6	A	26.8

SOLUTION

(i) Ho: $\mu_{.1} = \mu_{.2} = \mu_{.3}$

Hi: Not all means are equal

H'o: $\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$

H'i: Not all replications are equal

(ii) $\alpha = 0.05$

(iii) Test statistic to be used is

$$F = \frac{\text{MS for Treatment}}{\text{MS for Errors}} = \frac{s_t^2}{s_e^2}$$

has an F-distribution with $\upsilon_1 = (c-1)$ and $\upsilon_2 = (r-1)(c-1)$ d.f

(iv) The necessary calculations are given below

Replication	Varieties			$B_{i.}$
	A	B	C	
1	31.1	31.7	34.2	98.0
2	17.0	32.7	30.7	80.4
3	40.8	25.3	48.2	114.3
4	26.8	47.9	59.6	134.3
$T_{.j}$	116.7	137.6	172.7	427.0

$$\text{Total SS} = \sum_i \sum_j y^2_{ij} - CF$$

$$CF = \frac{(427.0)^2}{12} = 15194.08$$

$$TSS = \left[(32.1)^2 + (17)^2 + \ldots.. + (59.6)^2\right] - 15194.0$$

$$= 16698.50 - 15194.08 = 1504.42$$

$$\text{Treatment SS} = \sum_j \frac{T^2_j}{r} - CF$$

$$= \left[\frac{(116.7)^2 + (137.6)^2 + (172.7)^2}{4}\right] - 15194.08$$

$$= 15594.48 - 15194.08 = 400.40$$

$$\text{Replicate SS} = \sum_i \frac{B^2_i}{c} - CF$$

$$= \left[\frac{(98)^2 + (80.4)^2 + (114.3)^2 + (134.3)^2}{3}\right] - 15194.08$$

$$= 15728.05 - 15194.08 = 528.97$$

These results are displayed in the following analysis of variance table:

Source of Variation	d.f	SS	MS	F
Treatment	2	400.40	200.20	$F_1 = 2.09$
Replicates	3	528.97	176.32	$F_2 = 1.87$
Error	6	575.05	-	-
Total	11	1504.42	-	-

(v) The critical region are

$F_1 \geq F_{0.05(2,6)} = 5.14$ and

$F_2 \geq F_{0.05(3,6)} = 4.76$

(vi) Conclusion:

The calculated values of F_1 and F_2 do not fall in the critical region. Therefore we do not reject Ho and H'o and conclude that all treatment means are equal.

14.6 LATIN SQUARE DESIGN

The RCB design removes one source of variation when the varitation is found in two directions, it becomes necessary to remove these two sources of variation simultaneously. In that situation the experimental units are grouped in blocks in two different ways that is, by rows and columns. The design that is particularly useful in controlling two source of variations is called Latin Square Design. If there are k treatments, the experimental area will be divided into k-rows and k-columns. This design is in a square pattern.

14.6.1 Experimental Layout

The experimental layout for a Latin square design for 5 treatments A, B, C, D and E might be as follows

	Columns				
Rows	1	2	3	4	5
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C
5	E	A	B	C	D

In this design, there is a complete set of treatments in each rows and columns.

14.6.2 Statistical Model and Analysis

Let $y_{ij(h)}$ denotes the observation in i^{th} row j^{th} column and h^{th} treatment, the statistical model might be

$$y_{ij(h)} = \mu + R_i + C_j + \tau_h + e_{ij(h)}$$

$$i,j,h = 1,2,3,.....K$$

Where μ = overall mean

R_i = effect of ith row

C_j = effect of jth column

τ_h = effect of hth treatment and

$e_{ij(h)}$ = error term and $e_{ij(h)} \sim N(0, \sigma^2)$

Partitioning the sum of squares into its components in Latin Square Design will be

Total SS = Rows SS + Column SS + Treatment SS + Error SS

$$\text{Total SS} = \Sigma\Sigma\, y^2_{ij(h)} - CF$$

Where CF = (correction factor) $= \dfrac{G^2}{K^2}$

$$\text{Rows SS} = \sum_i \frac{R^2_i}{k} - CF$$

$$\text{Columns SS} = \sum_j \frac{C^2_j}{k} - CF$$

$$\text{Treatment SS} = \sum_h \frac{T_h^2}{k} - CF$$

Error SS = TSS – CSS – RSS – SST

To test Ho = $\tau_h = 0$, against Hi: $\tau_h \neq 0$

The F-ratio is

$$F = \frac{\text{MS for Treatment}}{\text{MS for Error}} = \frac{s_t^2}{s_e^2}$$

has an F-distribution with $\upsilon_1 = k\text{-}1$ and $\upsilon_2 = (k\text{-}1)(k\text{-}2)$ degrees of freedom. The resulting analysis of variance is shown in a table.

ANOVA-Table for K ×K Latin Square Design

Source of Variation	d.f	Sum of Squares	Mean Squares	F-ratio
Rows	k-1	$\sum_i \frac{R_i^2}{k} - CF$	$\frac{RSS}{K\text{-}1} = s_r^2$	
Columns	k-1	$\sum_j \frac{C_j^2}{k} - CF$	$\frac{CSS}{K\text{-}1} = s_c^2$	
Treatments	k-1	$\sum_h \frac{T_h^2}{k} - CF$	$\frac{SST}{K\text{-}1} = s_t^2$	$F_1 = \frac{s_t^2}{s_e^2}$
Error	(k-1)(k-2)	By subtraction	$ESS/K\text{-}1 = s_e^2$	
Total	$K^2\text{-}1$	$\sum_i \sum_j y^2_{ijh} - CF$	—	—

Critical Region and Conclusion

Critical region for hypothesis in case of treatment

Ho: $\tau_h = 0$ is $F \geq F_{\alpha(k\text{-}1,\ (k\text{-}1)(k\text{-}2))}$

In a similar way we can proceed to test the hypothesis that all row-effect or column-effect are zero.

If we reject the null hypothesis after ANOVA, then we use LSD test, given below

$$LSD = t_{\alpha/2(\omega)}\sqrt{\frac{2(MSE)}{K}}$$

Where K is the number of rows or columns

14.6.3 Advantages and Disadvantages of Latin Square Design

The important advantages of Latin Square Design are as follows.

(i) Its analysis is simple and straight forward

(ii) It is generally more efficient than RCB design

(iii) It reduces the error by controlling two source of variation.

Its disadvantages are

(i) It is less flexible than CR and RCB design

(ii) Replication in Latin Square Design is costly

(iii) When the treatments are small, it does not provide sufficient number of degrees of freedom to give a reliable error.

EXAMPLE 14.3

The research and development manager of auto mobile company wishes to study of the effect of tyre brand A, B, C and D on the tread loss (in mm) of tyres. Four different tyres of four different brands are filled at four different positions of four different cars, because there might be variability from one car to another car as well as from one position to another position, which is in a Latin Square Design. The data as per this design are presented as

Position	Car							
	1		2		3		4	
1	A	8	B	7	C	1	D	7
2	B	11	C	2	D	7	A	8
3	C	6	D	8	A	6	B	6
4	D	4	A	2	B	3	C	8

Write the corresponding model. Also check whether each component of the model has any effect on the tread loss of tyres at a significance level of 5%.

SOLUTION

Linear model of Latin Square Design is

$$y_{ij(h)} = \mu + P_i + C_j + \tau_h + e_{ij(h)}$$

The null and alternative hypothesis are:

(i) (a) H_0: $C_1 = C_2 = C_3 = C_4$

H_1: At least 2 columns means (cars) are not equal.

(b) H'_0: $P_1 = P_2 = P_3 = P_4$

H'_1: At least 2 rows means (positions) are not equal.

(c) H''_0: $T_1 = T_2 = T_3 = T_4$

H''_1: At least 2 treatment means (tyres) are not equal

(ii) $\alpha = 0.05$

(iii) $F_1 = \dfrac{s_c^2}{s_e^2}$, $F_2 = \dfrac{s_p^2}{s_e^2}$ and $F_3 = \dfrac{s_t^2}{s_e^2}$

Which has F distribution with υ_1, υ_2 d.f.

(iv) The necessary calculations are given below

Positions	Cars								R_i
	1		2		3		4		
1	A	8	B	7	C	1	D	7	23
2	B	11	C	2	D	7	A	8	28
3	C	6	D	8	A	6	B	6	26
4	D	4	A	2	B	3	C	8	17
C_j	29		19		17		29		94

Tyres					
	A	B	C	D	
	8	11	6	4	
	2	7	2	8	
	6	3	1	7	
	8	6	8	7	
T_h	24	27	17	26	94

$$\text{Total SS} = \sum_i \sum_j y^2_{ijh} - CF$$

$$CF = \frac{G^2}{K^2} = \frac{(94)^2}{16} = 552.25$$

$$TSS = \left[(8)^2 + (11)^2 + ... + (8)^2\right] - 552.25$$

$$= 666 - 552.25 = 113.75$$

$$\text{Columns/Car SS} = \sum_j \frac{C^2_j}{k} - CF$$

$$= \left[\frac{(29)^2 + (19)^2 + (17)^2 + (29)^2}{4}\right] - 552.25$$

$$= 583 - 552.25 = 30.75$$

$$\text{Rows/Position SS} = \sum_i \frac{R^2_i}{k} - CF$$

$$= \left[\frac{(23)^2 + (28)^2 + (26)^2 + (17)^2}{4}\right] - 552.25$$

$$= 569.5 - 552.25 = 17.25$$

$$\text{Treatments / Tyres SS} = \sum_h \frac{T^2_h}{k} - CF$$

$$= \left[\frac{(24)^2 + (27)^2 + (17)^2 + (26)^2}{4}\right] - 552.25$$

$$= 567.5 - 552.25 = 15.25$$

$$\text{Error SS} = TSS - CSS - RSS - SSt$$

$$= 113.75 - 30.75 - 17.25 - 15.25$$

$$= 50.50$$

The results are arranged in the following ANOVA table

Source of Variation	d.f	Sum of Squares	Mean Squares	F-ratio
Column/Cars	3	30.75	10.25	$F_1 = \frac{10.25}{8.42} = 1.22$
Rows/Positions	3	17.25	5.75	$F_2 = \frac{5.75}{8.42} = 0.68$
Treatments/Tyre	3	15.25	5.08	$F_3 = \frac{5.08}{8.42} = 0.60$
Error	6	50.50	8.42	-
Total	15	-	-	-

(v) Critical Regions for hypothesis are

$F_1 \geq F_{0.05(3,6)} = 4.76$

$F_2 \geq F_{0.05(3,6)} = 4.76$

$F_3 \geq F_{0.05(3,6)} = 4.76$

(vi) Conclusion

(a) Calculated value F_1 does not fall in the critical region. Hence null hypothesis Ho should be accepted. Therefore we conclude that there is no significant differences in terms of tread loss between different cars.

(b) Calculated value of F_2 does not fall in the critical region therefore accept H_o and conclude that there is no significant difference in terms of tread loss between different positions.

(c) In the third case calculated value also does not fall in the critical region, so we accept H_o and conclude that there is no significant difference in terms of treads loss between different brand of tyres.

SUMMARY

The brief summary of the chapter is given below

1. **Completely Randomized Design (CR Design)**

Total sum of squares = Treatment SS + Error SS

Where

$$\text{Total SS} = \sum_i \sum_j y^2_{ij} - CF; \qquad CF = \frac{(T..)^2}{n}$$

$$\text{Treatment SS} = \sum_j \frac{T^2_j}{r} - CF$$

Error SS = Total SS – Treatment SS

$$\text{Mean square treatment} = \frac{\text{Treatment SS}}{k\text{-}1} = s_t^2$$

$$\text{Mean square Error} = \frac{\text{Error SS}}{n\text{-}k} = s_e^2$$

$$F = \frac{s_t^2}{s_e^2}, \text{ With } \upsilon 1 = k\text{-}1 \text{ and } \upsilon 2 = n\text{-}k \text{ d.f}$$

Statistical Model for CR design is $y_{ij} = \mu + \tau_j + e_{ij}$

2. **Randomized Complete Block Design (RCB Design)**

Total SS = Block SS + Treatment SS + Error SS

Where

$$\text{Total SS} = \Sigma\Sigma\, y^2_{ij} - CF; \; CF = \frac{(T..)^2}{n}$$

$$\text{Block SS} = \sum_i \frac{B^2_i}{c} - CF$$

$$\text{Treatment SS} = \sum_j \frac{T^2_j}{r} - CF$$

Error SS = TSS – Block SS – Treatment SS

$$\text{Mean square Treatment} = s_t^2 = \frac{\text{Treatment SS}}{c\text{-}1}$$

$$\text{Mean square Block} = s_b^2 = \frac{\text{Block SS}}{r\text{-}1}$$

Mean square Error $= s_e^2 = \dfrac{\text{Error SS}}{(r-1)\,(c-1)}$

$F_1 = \dfrac{s_t^2}{s_e^2}$, with $\upsilon_1 = (c-1)$ and $\upsilon_2 = (r-1)(c-1)$ d.f

$F_2 = \dfrac{s_b^2}{s_e^2}$ with $\upsilon_1 = (r-1)$ and $\upsilon_2 = (r-1)(c-1)$ d.f

Statistical Model for RCB design is $y_{ij} = \mu + B_i + \tau_j + e_{ij}$

3. **Latin Square Design (LS Design)**

Statistical Model for Latin square design is

$$y_{ij(h)} = \mu + R_i + C_j + \tau_h + e_{ij(h)}$$

Partitioning of total sum of squares in LS design would be

Total SS = Rows SS + Columns SS + Treatment SS + Error SS

Where

Total SS $= \Sigma\Sigma\, y^2_{ij(h)} - CF$, $CF = \dfrac{(G)^2}{K^2}$

Rows SS $= \sum\limits_i \dfrac{R_i^2}{K} - CF$

Column SS $= \sum\limits_j \dfrac{C_j^2}{K} - CF$

Treatment SS $= \sum\limits_h \dfrac{T_h^2}{k} - CF$ and

Error SS = TSS – CSS – RSS – SST

MS column $= \dfrac{\text{Column SS}}{k-1}$, MS Rows $= \dfrac{\text{Row SS}}{k-1}$

MS treatment $= \dfrac{\text{Treatment SS}}{k-1}$ and MS Error $= \dfrac{\text{Error SS}}{(k-1)(k-2)}$

$F_1 = \dfrac{s_c^2}{s_e^2}$ with $\upsilon_1 = k-1$ and $\upsilon_2 = (k-1)\,(k-2)$ d.f

$$F_2 = \frac{s_r^2}{s_e^2} \qquad \text{with} \quad \upsilon_1 = (k\text{-}1) \text{ and } \upsilon_2 = (k\text{-}1)\,(k\text{-}2) \text{ df}$$

Similarly

$$F_3 = \frac{s_t^2}{s_e^2} \qquad \text{with } \upsilon_1\ (k\text{-}1) \text{ and } \upsilon_2 = (k\text{-}1)(k\text{-}2) \text{ df}$$

Null hypothesis of equal means will be rejected in all designs if

$$F \geq F_{\alpha;\,(\upsilon 1,\,\upsilon 2)}$$

EXERCISES

14.1 What is meant by an experimental design? Explain the basic principles of design.

14.2 (a) Describe completely randomized design, its model and analysis.

(b) What are the advantages and disadvantages of CR design?

14.3 The manager of an automobile company wishes to study the effect of tyre brand on the tread loss (in mm) of tyres. Four tyres of brands A, B, C and D are fitted to four different cars using completely randomized design. The data as per this design are as follows

Tyre Brand			
A	B	C	C
6	3	8	4
7	6	6	2
10	2	7	1
9	3	2	4

(a) Check whether the tyre brand has an effect on the tread loss of tyres.

(b) If the tyre brand has an effect on tread loss then perform LSD test to compare the significant difference between brand means and comment on the results.

14.4 A stock market analyst wants to study the impact of the type of company on the quarterly average of earning per share (EPS). So he collected four quarterly averages of data of four different companies during the least year.

Company			
A	B	C	D
12	16	25	13
8	18	15	8
16	10	22	20
19	11	9	5

Check whether there is a significant difference between different companies in terms of EPS.

14.5 Perform the complete analysis of the following data obtained by CR design.

Factor			
A	B	C	D
3	4	10	9
5	3	9	2
7	1	11	5
6	5	7	1
9	2	8	3

14.6 The Govt. Accounting Office is interested in seeing whether similar size offices spend similar amounts on persomel and equipment. Monthly expenses for three offices have been examined. At 5% level of significance, are there differences in expenses for different offices?

Monthly Expenses in Million Rs.						
Agriculture	10	8	11	9	12	
State	15	9	8	10	13	13
Interior	8	16	12			

14.7 What do you mean by randomized complete block design? What are its advantages and disadvantages?

14.8 (a) Four varieties of wheat were tried in a RCB design in four replications. Yield in Kg per plot is shown in the table given below. Test the hypothesis that there is no difference in the mean of four varieties.

	Varieties			
Replicates	V_1	V_2	V_3	V_4
I	20	50	40	10
II	20	30	30	10
III	40	60	60	20
IV	10	40	20	30

(b) What would have been done if no blocking had been done?

14.9 (a) Compare randomized complete block design with completely randomized design.

(b) Complete the following analysis of variance table and test whether the treatments differ significantly at 5% level of significance.

Source	d.f	S.S	M.S	F
Blocks	3	270	–	
Treatments	–	–	–	–
Error	–	–	28.94	
Total	15	938	–	

14.10 An agriculture officer wants to study the effect of four different fertilizers on yield (in tons) of a crop. Because of there might be variability from one plot to another plot, he decides to use the randomized complete block design. The corresponding data are presented as follows.

	Fertilizers			
Plot	A	B	C	D
1	100	150	120	70
2	80	70	110	100
3	68	90	85	78
4	125	138	60	124

Check whether different fertilizers have any effect on the yield of crop.

14.11 There are four different technological alternatives to manufacturer a product. The research and development manager feels that the type of technology may have some impact on hourly output (in units). Because of variability from one plant to another plant, he decides to use RCB design. Check whether each component of the model has any effect on the output of the product.

	Technology			
Plant	T_1	T_2	T_3	T_4
P_1	73	68	74	71
P_2	73	57	75	52
P_3	45	38	68	40
P_4	73	41	75	75

14.12 For the data given below. Test for the significance between treatment means.

	Treatments				
BLOCK	1	2	3	4	5
I	40	30	55	25	35
II	45	60	10	22	33
III	38	55	40	55	28
IV	30	27	32	56	17
V	45	34	20	34	37

14.13 **(a)** Define the following terms

(i) Replication (ii) Randomization

(iii) Local Control

(b) What do you know about LSD test? Explain briefly.

14.14 Given the following analysis of variance for RCB design:

Source of Variation	d.f	Sum of Square	Mean Square
Block	9	0.4074	
Treatments	3	1.1986	
Error	27	0.6249	

(a) Complete the analysis

(b) Compute the standard error for a treatment mean and for the difference between 2 treatment means.

(c) The treatment means are 1.465, 1.195, 1.325 and 1.662. What mean or means do you suspect might represent different population?

14.15 Define a Latin Square Design. Explain the difference between a Latin Square Design and a randomized complete Block design.

14.16 The production manager of a firm is keen in analyzing the hourely production. He feels that operators (A, B, C, D and E), machines (M_1, M_2, M_3, M_4 and M_5) suppliers of raw material (S_1, S_2, S_3, S_4 and S_5) will affect the production volume of the product. Hence he designed a Latin Square Design as shown below.

	Suppliers									
Machines	S_1		S_2		S_3		S_4		S_5	
M_1	A	12	B	8	C	9	D	15	E	18
M_2	B	14	C	7	D	13	E	10	A	12
M_3	C	7	D	15	E	12	A	8	B	6
M_4	D	16	E	17	A	13	B	11	C	12
M_5	E	9	A	14	B	8	C	10	D	8

Test whether there is significant difference between operators in terms of production of volume of the product.

14.17 Perform the complete analysis of variance of the following data by Latin square design.

	Columns							
Rows	1		2		3		4	
1	A	7	B	3	C	5	D	9
2	B	4	C	1	D	7	A	8
3	C	6	D	0	A	8	B	10
4	D	3	A	8	B	4	C	13

14.18 Five varieties of wheat A, B, C, D and E, were tested by arranging plants in a Latin Square design in a field. The rows and columns in the table are rows and columns in the field. The yield in Kg per plot are given below.

	Columns									
Rows	1		2		3		4		5	
1	49	B	64	D	33	E	95	A	118	C
2	93	C	40	A	62	B	51	E	54	D
3	76	D	154	C	65	A	60	B	46	E
4	53	E	76	B	132	C	86	D	49	A
5	93	A	63	E	118	D	159	C	76	B

Analysis the data that mean yields are not equal for five varieties of wheat. Also pickup the best variety of wheat.

14.19 TICK THE CORRECT ANSWER

(i) **The basic principles of experimental design are:**

(a) 2 (b) 3

(c) 4 (d) 5

(ii) **A design in which treatments are assigned to experimental units completely at random is called:**

(a) CR design (b) RCB design

(c) Latin Square Design (d) None of these

(iii) **A design which controls variability in one direction is called:**

(a) CR design (b) RCB design

(c) Latin Square Design (d) Both (b) and (c)

(iv) **A design in which experimental material is divided into groups or blocks is called:**

(a) RCB design (b) CR design

(c) Latin Square (d) None of these

(v) **A design which controls variability in two directions is called:**

(a) CR design (b) RCB design

(c) Latin Square Design (d) None of these

(vi) **The most frequently used experimental design is:**

(a) CR design (b) RCB design

(c) Latin Square Design (d) Factorial design

(vii) **Five varieties of wheat were tried in RCB design in four replicates the error degrees of freedom will be:**

(a) 3 (b) 4

(c) 5 (d) 12

(viii) **The mean square is equal to sum of squares divided by:**

(a) Total number of observations

(b) Its corresponding degrees of freedom-1

(c) Its corresponding degrees of freedom

(d) None of these

(ix) **$Y_{ij} = \mu + B_i + \tau_j + e_{ij}$ is the statistical model of:**

(a) CR design (b) Latin Square Design

(c) RCB design (d) Graeco LS design

Answers 14.19

(i)	(b)	**(ii)**	(a)	**(iii)**	(b)	**(iv)**	(a)	**(v)**	(c)
(vi)	(b)	**(vii)**	(d)	**(viii)**	(c)	**(ix)**	(c)		

SHORT QUESTION ANSWERS

1. **What is experimental design?**

Ans. An efficient procedure to conduct a planned experimental to draw valid conclusion from the given data is called experimental design.

2. **What are the basic principles of design?**

Ans. The basic principles of the experimental design are

(i) Randomization (ii) Replication

(iii) Local control

3. Define Randomization.

Ans. It is a random process of assigning treatments to the experimental unit. It is used to remove bias due to personal preferences of the researcher.

4. What do you mean by Replication?

Ans. Replication is the repetition of the basic experiment. It helps in increasing the precision.

5. Define Local Control.

Ans. This term is used to the amount of balancing, blocking and grouping of experimental units.

6. Write down the basic experimental designs.

Ans. The basic experimental designs are

(i) Completely Randomized Design (CR Design)

(ii) Randomized Complete Block Design (RCB Design)

(iii) Latin Square Design (LS Design)

7. What do you mean by CR design?

Ans. In CR design the treatments are assigned to experimental units completely at random. The design is completely flexible. It is the simplest type of basic design. It can be used only when number of treatments is small.

8. Define RCB design.

Ans. In this design, the experimental material is divided into groups or blocks in such a way that experimental units within each blocks are homogenous. Every block contains a complete set of treatments. It is most widely used design. This design is also flexible. It controls variability only on one direction.

9. Describe Latin Square Design.

Ans. The RCB design removes one source of variation. For controlling two source of variations we use Latin Square Design. In this design the experimental units are grouped in blocks into two different ways that is by rows and columns. If there are k treatments, then experimental area will be divided into k rows and k columns. This design is in a square pattern.

ANSWERS

CHAPTER – 1

1.10	(i)	Quantitative	(ii)	Qualitative	(iii)	Qualitative
	(iv)	Quantitative	(v)	Qualitative	(vi)	Quantitative
	(vii)	Quantitative	(viii)	Qualitative	(ix)	Quantitative
	(x)	Qualitative	(xi)	Quantitative	(xii)	Qualitative
1.11	(i)	Continuous	(ii)	Discrete	(iii)	Discrete
	(iv)	Continuous	(v)	Continuous	(vi)	Discrete
	(vii)	Continuous	(viii)	Discrete	(ix)	Discrete
	(x)	Continuous	(xi)	Continuous	(xii)	Continuous
	(xiii)	Discrete	(xiv)	Discrete	(xv)	Continuous
	(xvi)	Discrete	(xvii)	Discrete		

CHAPTER – 2

2.9 Classes are 70 – 79, 80 – 89, 90 – 99, ...

CHAPTER – 3

3.2 (b) (i) 3 (ii) 6 (iii) 15.75 (iv) Mean = 44.5

(c) Mean = 50.36

3.3 (a) Mean = 2 (b) Mean = 33.24 ≈ 33

3.4 (a) Mean = 48 (b) Mean = 49.5 (c) 35.25

3.5 Mean = 67.45 inches

3.6 (a) Mean = 26.46 marks ≈ 26 (b) Mean = 128.8

3.7 (b) 203.4 (c) Weighted mean = 70 (d) Mean = 34.75

3.8 (a) Combined Mean = 21.67 (b) 67.46

3.9 (a) Mean = 12.38 (b) Mean = 5.896

3.10 (b) (i) Median = 55 (ii) 9 (iii) Median = 90

3.11 (a) Median = 4 (b) 5

3.12 Median = 40.79

3.13 Median = 109.67

3.14 (b) Median = 16.11

3.15 (b) (i) Mode = 7 (ii) Two Mode, (60, 70) (iii) No Mode

3.16 (a) Mode = 8 (b) Mode = 4 children per house

(c) Median = 3, Mode = 3

3.17 (a) Mode = 37.31 years

(b) Mean = 11.09, Median = 11.07, Mode = 11.06

3.18 (a) Mean = 146.98, Median = 146.75, Mode = 147.2

(b) Mean = 1.32, Median = 1.27, Mode = 1.22

3.19 (a) Mode = 337.50 rupees (b) Missing value = 12

3.20 (b) Mode = 14 (c) Mode = 33.5 marks

3.21 (i) Mean (ii) Mode (iii) Median

(iv) Mode (vi) Mean (vii) Mode

(vii) Mean (viii) Mean

3.22 Mean = 61.318, Median = 61.805

3.23 155.33, 153.1

3.24 (a) Q1 = 23.26, Q3 = 45.5, Med = 34.25

D7 = 42.83, P20 = 21.07, P60 = 38.25

(b) Q1 = 27, Q3 = 40, D5 = 35, P80 = 41.2

(c) Median = 65, $Q_1 = 56$

$Q_3 = 74$, $D_8 = 76$, $P_{65} = 70$

3.26 (a) (i) 6.43 (ii) 19.368 (iii) 69.11 (iv) 117.2967 (v) 39.68

(b) 3.4087

3.27 (a) 117.7 grams

3.28 329.5

3.29 23.17%

3.30 **(b)** (i) 2.1899 (ii) 21.87 (iii) 63.90 **(c)** 14.26

3.31 H.M. = 35.48, Mean = 42.12

3.32 **(b)** H = 63.062

3.33 **(i)** 69.918 **(ii)** 66.740 **(iii)** 69.667

3.34 **(i)** Mean = 19.25 **(ii)** 18. 423 **(iii)** 16.84

CHAPTER – 4

4.3 **(i)** 14 , 0.39 **(ii)** 13 , 0.06

4.4 **(b)** Range = 180, Coefficient = 0.43, **(a)** = 20

4.5 **(i)** 3.375 , 0.1047 **(ii)** 3 , 0.028

4.6 5.425

4.7 9.78, 0.068

4.8 **(b)** 4.8 **(c)** 5.4

4.9 6.9 , 0.14

4.10 5.167

4.11 2.1962 , 0.3629

4.12 **(i)** 2.0, 1.4142 **(ii)** 63.84, 7.99

(iii) 6.85, 2.62

4.13 **(a)** 2.94 , 1.715 , 2.57% **(b)** Variance = 12, S = 3.46

4.14 90.55, 9.516

4.15 47.06, 6.86, 17.59%

4.16 220.36, 24.62%

4.18 $\bar{x}$ = Rs. 90.15, S = 15.99, 65%, 95%, 100%

4.19 $\bar{x}$ = Rs. 12.006, S = Rs. 2.626

4.20 64.44, 17.07

(i) 74.44, 17.07 **(ii)** 70.89, 18.78

(b) **(i)** 10, 100 **(ii)** 5, 25 **(iii)** 10, 100

4.21 $\bar{x}_A = 44.47$, C.VA = 15.38, $\bar{x}B = 47.56$, C.VB = 11.73

4.22 Town A, S = 11.88, C.V. 21.15%

Town B, S = 12.95, C.V. = 23.59%

4.23 A is good student and A is also consistent

4.24 (a) $\bar{x}_C = 53.80$, Sc = 8.74

(b) 20, 20

4.25 47, 6.86

4.26 (b) (i) Left skewed (ii) Symmetrical (iii) Right skewed

4.27 (b) Skewness = 0.034

4.28 0.0628

4.29 –0.020 , – 0.071

4.30 0.299

4.31 (i) 0.957 (ii) 0.917 (iii) 0.066

4.32 (b) 1,15 (c) 1.80, 1.59

4.33 (a) 0,1,1,7 (b) Symmetrical (c) 0.122

4.34 (a) 0,1.7 (b) 0, 12, 0, 324

4.35 0, 2.49, 0.7, 18.33

Skewness = 0.178 Kurtosis = 2.96

CHAPTER – 5

5.2 $y = 0.751 + 0.00435x$, $\hat{y} = 3.361$

5.3 (ii) $y = 11.8 + 1.142x$, $x = 4.405 + 0.605y$

5.4 $r = 0.995$, $y = 9.2 + 0.56x$, $x = -16.057 + 1.7675y$

5.5 $r = -0.96$, $y = 301.67 - 2.00x$

5.6 (a) $y = 18.88 - 0.609x$

5.7 (b) – 0.415

5.8 $y = 35.9383 + 0.4835x$, $\hat{y} = 69.75$

5.9 (a) $Y = 3.63 + 0.46x$ (b) 0.674

5.10 0.834

5.11 – 0.397

5.12 (a) – 0.96 (b) $b_{yx} - 2$, $b_{xy} = -0.46$

5.13 $b_{yx} = 2.064$, $b_{yx} = 0.474$

5.14 0.649

5.15 r = 0.954, y = 124. 84

5.16 $\hat{Y} = 9.2 + 0.56\,X$, y = 21.52

5.17 $\hat{X} = -64.248 + 2.275\,y$

5.18 (b) 0.835 (c) 0.775

5.19 (a) 0.977 (b) 0.98

5.20 (i) –0.98 (ii) –0.98

5.21 0.876

5.22 (a) 0.6133 (b) 0.5

5.23 0.5333

5.24 0.667

5.25 0.857

5.26 0.834

5.27 (c) 0.947, 4.048

5.28 (b) 0.8667

5.29 (a) 0.60 (b) 0.8857

CHAPTER – 6

6.2 (a) $\hat{y} = 96.4581 + 136.4847\,x_1 - 2.4035\,x_2$

(b) 365 pounds

6.3 (a) $\hat{y} = 3.88 + 2.09\,x_1 + 2.65\,x_2$

(b) 1.4335 (c) $R^2_{y.12} = 0.945$, $R_{y.12} = 0.972$

6.4 $\hat{y} = 9.6721 + 0.0774\,x_1 + 0.0651\,x_2$

6.5 (a) $\hat{x}_3 = 61.40 - 3.65\,x_1 + 2.54\,x_2$

(b) 40

6.6 (a) $\hat{x}_3 = 16.07 + 4.36\,x_1 + 4.04\,x_2$

(b) 0.757

6.7 $r_{12.3} = 0.759$, $r_{23.1} = -0.436$, $r_{13.2} = 0.097$

$\hat{x}_1 = 9.22 + 3.37\, x_2 + 0.0038\, x_3$

6.8 (b) $r_{12.3} = -0.86$, $r_{23.1} = 0.92$, $r_{31.2} = 0.98$

(c) $r_{12.3} = 0.546$, $r_{13.2} = 0.0765$, $r_{23.1} = -0.7935$, $R_{1.23} = 0.80$

6.9 (a) $R_{2.13} = 0.83$

(b) $r_{23.1} = 0.6520$, $R_{2.13} = 0.67$

6.10 (a) $\hat{y} = -1.33 + 0.38\, x_1 + 1.62\, x_2$

(b) $S_{y.12} = 9.27$, $R^2_{y.12} = 0.55$, $R_{y.12} = 0.74$

CHAPTER – 7

7.2 (i) {2} (ii) {2, 3, 4, 6, 8} (iii) ϕ
(iv) {(1, 3), (1, 4), (2, 3), (2, 4), (3, 3), (3, 4)}

7.3 (b) (i) {2, 4, 6, 8, 10} (ii) {1, 3, 5, 7, 9}
(iii) {1, 2, 3, ..., 10} (iv) ϕ

7.4 (b) (i) 72 (ii) 6840
(iii) 45 (iv) 2598960

7.6 (b) (i) $\frac{1}{2}$ (ii) $\frac{1}{4}$ (iii) $\frac{1}{6}$ (iv) $\frac{1}{2}$

(c) (i) $\frac{1}{2}$ (ii) $\frac{1}{4}$ (iii) $\frac{3}{4}$ (iv) 1

7.7 (a) $\frac{1}{8}$ (b) (i) $\frac{3}{8}$ (ii) $\frac{7}{8}$ (iii) $\frac{1}{2}$

7.8 (a) (i) $\frac{1}{2}$ (ii) $\frac{1}{6}$

(b) (i) $\frac{1}{2}$ (ii) $\frac{5}{18}$ (iii) $\frac{1}{6}$ (iv) $\frac{1}{36}$

(c) $\frac{15}{36}$

(d) (i) $\frac{1}{6}$ (ii) $\frac{5}{18}$ (iii) $\frac{3}{4}$

7.9 (a) (i) $\frac{1}{5}$ (ii) $\frac{1}{2}$

(b) (i) $\frac{4}{7}$ (ii) $\frac{1}{7}$ (iii) $\frac{5}{7}$

7.10 (a) (i) $\frac{3}{11}$ (ii) $\frac{3}{22}$ (iii) $\frac{1}{220}$

(b) (i) $\frac{3}{11}$ (ii) $\frac{3}{44}$

7.11 (i) $\frac{1}{2}$ (ii) $\frac{1}{52}$ (iii) $\frac{1}{13}$ (iv) $\frac{1}{4}$

7.12 (a) (i) $\frac{325}{1326}$ (ii) $\frac{169}{1326}$ (iii) $\frac{650}{1326}$

(b) (i) $\frac{8}{15}$ (ii) $\frac{7}{15}$ (iii) $\frac{4}{15}$

(iv) $\frac{1}{5}$ (v) $\frac{1}{5}$

7.13 (b) $\frac{2}{3}$

7.14 (a) $\frac{4}{25}$ (b) $\frac{5}{6}$

7.15 (a) $\frac{4}{13}$ (b) 0.88

7.16 (b) $\frac{2}{13}$

7.17 (a) $\frac{1}{2}$ (b) $\frac{5}{18}$

7.18 (b) $\frac{2}{3}$

7.19 (a) (i) $\frac{1}{3}$ (ii) $\frac{2}{7}$ (iii) 0

(b) $\frac{3}{7}$

7.20 (c) (i) $\frac{1}{169}$ (ii) $\frac{1}{221}$

7.21 (a) 0.15

(b) (i) 0.60 (ii) 0.40 (iii) 0.76

7.22 (i) $\frac{1}{12}$ (ii) $\frac{1}{4}$ (iii) $\frac{1}{6}$

(iv) $\frac{1}{2}$ (v) $\frac{1}{2}$

7.24 (i) $\frac{5}{18}$ (ii) $\frac{1}{6}$

7.25 (i) $\frac{1}{2}$ (ii) $\frac{3}{4}$

CHAPTER – 8

8.1 (b) 1, 2, 3, 4, 5, 6

x	1	2	3	4	5	6
f(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

8.2 (b)

x	0	1	2	3
f(x)	$\frac{5}{30}$	$\frac{15}{30}$	$\frac{9}{30}$	$\frac{1}{30}$

8.3

x	0	1	2	3
f(x)	$\frac{1}{56}$	$\frac{15}{56}$	$\frac{30}{56}$	$\frac{10}{56}$

8.4

x	0	1	2	3	4
f(x)	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Mean = 2, variance = 1

8.5 Mean = 1.5, variance = 0.75

8.6

x	2	3	4	5	6	7
f(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$
x	8	9	10	11	12	
f(x)	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

Mean = 7, variance = 5.83

8.7

x	0	1	2	3	4	5
f(x)	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Mean = 1.94

8.8 Mean = 5.27, variance = 8.94

8.9 Mean = 2, variance = 1

8.10

x	0	1	2	3
f(x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

8.11 **(b)** Mean = 0.55, variance = 1.35

(ii) 2.1, 5.39

8.12 Rs. 36, Rs. 30

8.13 Rs. 76, Rs. 57, Rs. 42

8.14 Rs. 61.25

8.15

x	1	2	3	4	5	6
f(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

E(x) = 3.5

8.16 **(b)** **(ii)** $\frac{16}{27}$ **(iii)** $\frac{1}{3}$

8.17 $A = \frac{3}{8}$, E(x) = 1, var(x) = $\frac{1}{5}$

8.18 E(x) = $\frac{1}{2}$, S.D.(x) = 0.387

CHAPTER – 9

9.1 **(b)** $\frac{1}{16}, \frac{4}{16}, \frac{6}{16}, \frac{4}{16}, \frac{1}{16}$

9.2 **(a)** 32/81, 0, $\frac{8}{81}$, 0, $\frac{8}{9}$ **(b)** $\frac{1}{1024}, \frac{120}{1024}, \frac{7}{128}, \frac{121}{128}$

9.3 **(a)** 0.9011, 0.0169 **(b)** 0.2269

9.4 **(a)** 0.169, 0.297, 0.534 **(b)** 0, 0.0078, 0.00168

9.5 **(a)** 12, 4.8 **(b)** n = 224, P = 0.17

9.6 3, 15, 30, 30, 15, 3

9.7 **(b)** 1/56 , 15/56, 30/56, 10/56

9.8 **(a)** 0.6364, 0.3636 **(b) (i)** 0.1143 **(ii)** 0.3714

9.9 $\frac{15}{210}$, 80/210, 90/210, 24/210, 1/210

9.10 0, 4/20, 12/20, 4/20

9.11 **(b) (i)** 0.2103 **(ii)** 0.4100 **(iii)** 0.0360

9.12 **(a) (i)** 0.1431 **(ii)** 0.1353 **(b)** 0.0183, 0.1464, 0.2379, 0.7621

9.13 **(i)** 0.368 **(ii)** 0.061 **(iii)** 0.920 **(iv)** 0.08

9.14 **(i)** 0.9513 **(ii)** 0.9989

9.15 **(i)** 0.1008 **(ii)** 0.5976 **(iii)** 0.4233

9.16 **(b)** 0.0335

9.17 **(a)** 0.1367 **(b)** 0.018

9.18 **(b)** 0.067 **(c)** 1/4

9.19 **(b) (i)** 0.2119 **(ii)** 0.9192 **(iii)** 0.0228 **(iv)** 0.1587

9.20 **(a)** 7.95% **(b) (i)** 0.9980 **(ii)** 0.3983

9.21 **(i)** 0.3336 **(ii)** 0.1496 **(iii)** 0.0526

9.22 **(i)** 0.5 **(ii)** 0.4332 **(iii)** 0.0228 **(iv)** 0.7881

9.23 **(i)** 0.3446 **(ii)** 0.8413

9.24 **(i)** 0.7454 **(ii)** 0.8554 **(iii)** 0.4382

9.25 **(a)** 20.2, 39.8 **(b)** $P_{95} = 664.5$ **(c)** 168cm

9.26 **(i)** 49.6 **(ii)** 65.8

CHAPTER – 10

10.5 **(b)** $\mu_{\bar{x}} = 6$, $\sigma^2_{\bar{x}} = 5.83$, $\mu = 6$, $\sigma^2 = 11.67$

10.6 $\mu_{\bar{x}} = 6$, $\sigma_{\bar{x}} = 1.732$, $\mu = 6$, $\sigma = 4.243$

10.7 $\mu_{\bar{x}} = 3$, $\sigma^2_{\bar{x}} = 2$, $\mu = 3$, $\sigma^2 = 6$

10.8 $\mu_{\bar{x}} = 6$, $\sigma_{\bar{x}} = 1.826$, $\mu = 6$, $\sigma = 2.582$

10.9 $\mu_{\bar{x}} = 6$, $\sigma^2_{\bar{x}} = 3$, $\mu = 6$, $\sigma^2 = 18$

10.10 $\mu_{\bar{x}} = 10.5$, $\sigma_{\bar{x}} = 2.29$, $\mu = 10.5$, $\sigma^2 = 5.12$

10.11 **(i)** $\mu = 13.5$ **(ii)** $\sigma = 5.12$

(iii) $\mu_{\bar{x}} = 13.5$ **(iv)** $\sigma_{\bar{x}} = 2.29$

10.12 **(i)** $\mu_{\bar{x}} = 6$ **(ii)** $\sigma^2_{\bar{x}} = 4$

(iii) $\mu = 6$ **(iv)** $\sigma^2 = 8$

10.13 $\mu_{\bar{x}} = 5$, $\sigma^2_{\bar{x}} = 4.67$, $\mu = 5$, $\sigma^2 = 11.67$

10.14 $\mu_{\bar{x}} = 4$, $\sigma^2_{\bar{x}} = 0.67$, $\mu = 4$, $\sigma^2 = 2.67$

10.15 $\mu_{\bar{x}} = 7$, $\sigma^2_{\bar{x}} = 0.1458$, $\mu = 7$, $\sigma^2 = 5.83$

10.16 $\mu_{\bar{x}} = 2.429$, $\sigma^2_{\bar{x}} = 0.1813$, $\mu = 2.429$, $\sigma^2 = 0.8147$

10.17 **(a)** $\sigma_{\bar{x}} = 3$ **(b)** $n = 256$

10.18 **(a)** 0.0401 **(b)** **(i)** 0.0032 **(ii)** 0.8621 **(iii)** 0.2407

10.19 **(b)** $\mu_{\hat{p}} = 0.40$, $\text{Var}(\hat{p}) = 0.04$

10.20 $\mu_{\hat{p}} = \frac{1}{2}$, $\text{Var}(\hat{p}) = \frac{1}{8}$

10.21 $\mu_{\hat{p}} = \frac{1}{2}$, $\text{Var}(\hat{p}) = \frac{1}{20}$

10.22 **(a)** 0.0026 **(b)** **(i)** 0.7291 **(ii)** 0.8720

10.23 **(i)** 0.0043 **(ii)** 0.093

10.24 **(b)** 46.92, 53.08mm

10.25 **(a)** 23.48, 24.52 **(b)** 5.87, 6.53

10.26 **(a)** 55.64, 68.36 **(b)** 36.24, 42.16 km

10.27 **(a)** 0.316, 0.524 **(b)** 0.549, 0.771

10.28 0.8814, 0.9186

10.29 **(b)** 1970

10.30 **(a)** 404 **(b)** 35

CHAPTER – 11

11.3 **(b)** z = 2.0, reject Ho

11.4 **(a)** z = –3.57, reject Ho **(b)** z = 1.40, accept Ho

11.5 **(a)** z = –4.0, reject Ho **(b)** z = 16.29, reject Ho

11.6 **(a)** z = – 5.0, reject Ho **(b)** Z = -2.88, reject Ho

11.7 **(b)** t = 1.89, accept Ho **(c)** t =1.845, No

11.8 **(b)** t = 1.69, accept Ho

11.9 t = –1.66, reject Ho

11.10 **(a)** z = –2.11, reject Ho **(b)** z = 1.99 reject Ho

11.11 z = –2.49, accept Ho

11.12 **(a)** z = 9.92, reject Ho

11.12 **(b)** z = 3.09, reject Ho

11.13 t = –1.07, accept Ho

11.14 **(a)** t = –2.766, reject Ho **(b)** t = –2.45, reject Ho

11.15 t = –4.162, reject Ho

11.16 **(b)** t = –6.72, reject Ho

11.17 t = –0.91, accept Ho

11.18 **(a)** z = 1.41, accept Ho **(b)** z = –6.90, reject Ho

(c) z = 1.94, reject Ho **(d)** z = 0.51, accept Ho

11.19 z = –1.49,m accept Ho

11.20 **(a)** z = 2.23, reject Ho **(b)** z = 1.11 **(c)** z = 0,10

11.21 z = –1.28, accept Ho

11.22 $\chi^2 = 20.16$, accept Ho

11.23 **(a)** $\chi^2 = 4.16$, accept Ho **(b)** $\chi^2 = 7.81$, reject Ho

11.24 **(a)** $\chi^2 = 46.37$, accept Ho **(b)** $\chi^2 = 8.75$, reject Ho

11.25 **(a)** (65.16, 524.31) **(b)** (32.81, 93.46)

11.26 **(b)** F = 2.78, reject Ho

11.27 **(a)** F = 2.25, accept Ho **(b)** F = 1.16, accept Ho

(c) (0.9, 8.23) and (0.96, 2.87)

11.28 F = 2.14, accept Ho

CHAPTER – 12

12.3 –0.478

12.4 Q = 0.71

12.5 χ^2 = 13.89, reject H_o

12.6 χ^2 = 9.929, reject H_o

12.7 χ^2 = 9.89, reject H_o

12.8 χ^2 = 6.11, reject H_o

12.9 χ^2 = 54.06, reject H_o

12.10 χ^2 = 69.15, reject H_o

12.11 χ^2 = 4.204, accept H_o

12.12 χ^2 = 6.11 , accept H_o

12.13 χ^2 = 5.027 accept H_o

12.14 χ^2 = 29.68, reject H_o

12.15 χ^2 = = 69.15, reject H_o

12.16 χ^2 = = 0.71 , accept H_o

12.17 χ^2 = = 1.77, accept H_o

CHAPTER – 13

13.2 F = 9.69, reject Ho

13.3 F = 3.76, accept Ho

13.4 F = 2.12, accept Ho

13.5 F = 0.98, accept Ho

13.6 F = 31.24, reject Ho

13.8 Breeds, F_1 = 6.81, Rations F_2 = 5.7, Both reject Ho

13.9 Varieties F_1 = 6.50, Fertilizer, F = 4.02, reject Ho

13.10 (i) F = 75.60, reject Ho (ii) LSD = 1.30

13.11 (a) d.F = 5, SS = 323.82, MS = 255.14, 8.47, F = 7.65

(b) 6 blocks (c) Reject Ho

13.12 Factor I,F = 8.16, reject Ho; Factor II, F = 2.05 accept Ho, L.SD = 4.36

CHAPTER – 14

14.3 **(a)** F = 5.98, reject Ho **(b)** LSD = 3.03

14.4 F = 0.82, accept Ho

14.5 F = 7.0, reject Ho

14.6 F = 0.60, accept Ho

14.8 **(a)** F = 6.51, reject Ho **(b)** F = 3.71, reject Ho

14.9 **(b)** d.F = 3.9, SS = 407.54, 260.46, MS = 90, 135.85 and F = 4.69, reject Ho

14.10 F = 0.43 accept Ho

14.11 Technology F_1 = 4.11, Plant F_2 = 4.89, Both reject Ho

14.12 F = 0.636, accept Ho

14.14 **(a)** MS ; Block = 0.0453, Treatment = 0.3995

Error = 0.0231, F = 17.29

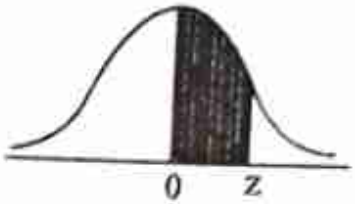
(b) 0.048, 0.068

14.16 F = 2.20, accept Ho

14.17 Treatment; F_1 = 0.98, Rows F_2 = 0.37, Columns; F_3 = 4.84

14.18 F = 21.0, reject Ho; LSD = 21.0

AREAS UNDER THE STANDARD NORMAL CURVE
From 0 to z



	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.01-20	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4551	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4062	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Critical Values for Student's t Distribution

Significance level = α

Degrees of Freedom	.005 (1-tail) .01 (2-tails)	.01 (1-tail) .02 (2-tails)	.025 (1-tail) .05 (2-tails)	.05 (1-tail) .10 (2-tails)	.10 (1-tail) .20 (2-tails)	.25 (1-tail) .50 (2-tails)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718
7	3.500	2.998	2.365	1.895	1.415	.711
8	3.355	2.896	2.306	1.860	1.397	.706
9	3.250	2.821	2.262	1.833	1.383	.703
10	3.169	2.764	2.228	1.812	1.372	.700
11	3.106	2.718	2.201	1.796	1.363	.697
12	3.054	2.681	2.179	1.782	1.356	.696
13	3.012	2.650	2.160	1.771	1.350	.694
14	2.977	2.625	2.145	1.761	1.345	.692
15	2.947	2.602	2.132	1.753	1.341	.691
16	2.921	2.584	2.120	1.746	1.337	.690
17	2.898	2.567	2.110	1.740	1.333	.689
18	2.878	2.552	2.101	1.734	1.330	.688
19	2.861	2.540	2.093	1.729	1.328	.688
20	2.845	2.528	2.086	1.725	1.325	.687
21	2.831	2.518	2.080	1.721	1.323	.686
22	2.819	2.508	2.074	1.717	1.321	.686
23	2.807	2.500	2.069	1.714	1.320	.685
24	2.797	2.492	2.064	1.711	1.318	.685
25	2.878	2.485	2.060	1.708	1.316	.684
26	2.779	2.479	2.056	1.706	1.315	.684
27	2.771	2.473	2.052	1.703	1.314	.684
28	2.763	2.467	2.048	1.701	1.313	.683
29	2.756	2.462	2.045	1.699	1.311	.683
Large	2.575	2.327	1.960	1.645	1.282	.675

Critical Values for the Chi-square (χ^2) Distribution

Degree of freedom	Significance Level (α)						
	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3,	6.251	7.8 5	9.348	9.837	11.345	12.838	16.268
4	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	13.362	15.507	17.535	18.168	29.090	21.955	26.125
9	14.684	16.919	19.023	19 579	21.o66	23.589	27.877
10	15.987	18.307	20.483	21.161	23.209	25.188	29.588
1 1	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	21.064	23.685.	26.119	26.873	29.141	31.319	36.123
15	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	36.74lt>	40.113	43.194	44.140	46.963	49.645	55.476
28	37.916	41.337	44.461	45.419	48.278	50.993	56.893
29	39.087	42.557	45.722	46.693	49.588	52.336	58.302
30	40.256	43.773	46.979	47.962	50.892	53.672	59.703

composed by ARFAN ALI MATHEMATICIAN 03148100197

Annexure I: F Tables

Percentage Points of F-Distribution ($\alpha = 0.10$)

Denominator degree of freedom (n_2)	Numerator degrees of freedom (n_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.79	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.33	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Annexure II: F Tables ($\alpha = 0.05$)

Denominator degree of freedom (n_2)	Numerator degrees of freedom (n_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

composed by ARFAN ALI MATHEMATICIAN 03148100197

Annexure III: F Tables

Percentage Points of F-Distribution ($\alpha = 0.01$)

Denominator degree of freedom (n_2)	Numerator degrees of freedom (n_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	23.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.25	13.27	12.06	11.39	10.97	10.67	10.46	10,29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
5	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.83
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.25	8.65	7.59	7.01	6.63	6.37	6.13	6.03	5.91	5.81	5.67	5.52	5.36	5.23	5.20	5.12	5.03	4.95	4.86
9	10.55	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.55	655	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	5.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3,78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	4.35	6.51	5.56	5.04	4.69	4.46	4.2S	4.14	4.03	3.94	3,80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.63	6.35	5.42	4.S9	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
15	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.40	6.11	5.13	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
13	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.OS	3.00	2.92	2.84	2.75	2.66	2.57
19	8.13	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.54	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	6.02	5.78	4.37	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7,SS	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
25	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3,29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.33
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00