CONTENTS

CHAPTER-I

DATA EXPLORATION AND REPRESENTATION

The word óStatisticsô is probably derived from the Latin word óstatusô (means a political state) or the Italian word óstatistaô or the German word óstatistikô, each of which means a ópolitical stateô. It is used in singular as well as in plural sense. In singular sense, statistics is used as a subject that deals with the principles and methods employed in collection, presentation, analysis and interpretation of data. In plural sense, statistics is considered as numerical description of quantitative information.

1.1 Statistics (Definition), Scope and Limitations:

Different persons defined Statistics in different ways. Some of the popular definitions of Statistics are given below.

According to Croxton and Cowden, ñ*Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data*ò.

Professor Horace Secrist defined Statistics as ñ*aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other*ò.

According to Sir R.A. Fisher ñ*The science of Statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data*ò.

Fisherôs definition is most exact in the sense that it covers all aspects and fields of Statistics. On the basis of above ideas, Statistics can be defined as a science which deals with collection, presentation, analysis of data and interpretation of results.

Statistics is often classified as Mathematical Statistics and Applied Statistics. The Mathematical Statistics deals with the theoretical developments, derivation of the formulae and statistical solutions of the problems. Applied Statistics on the other hand deals with the application of the statistical methods in different branches of science, art and industry.

The credit for applications of statistics in various diverse fields goes to Sir R.A. Fisher (1890-1962) known as ñFather of Statisticsò. Today, the statistics is not merely

confined to the affairs of the state for data collection but it is regarded as an important tool of all physical, social and biological sciences indispensable to research and intelligent judgement. The scope of statistics is given below:

Scope of Statistics:

i) Statistics has great significance in the field of physical and natural sciences. It is used in propounding and verifying scientific laws. Statistics is often used in agricultural and biological research for efficient planning of experiments and for interpreting experimental data.

ii) Statistics is of vital importance in economic planning. Priorities of planning are determined on the basis of the statistics related to the resource base of the country and the short-term and long-term needs of the country.

iii) Statistical techniques are used to study the various economic phenomena such as wages, price analysis, analysis of time series, demand analysis etc.

iv) Successful business executives make use of statistical techniques for studying the needs and future prospects of their products. The formulation of a production plan in advance is a must, which cannot be done in absence of the relevant details and their proper analysis, which in turn requires the services of a trained statistician.

v) In industry, the statistical tools are very helpful in the quality control and assessment. In particular, the inspection plans and control charts are of immense importance and are widely used for quality control purposes.

Limitations of Statistics:

i) Statistical methods are best applicable to quantitative data.

ii) Statistical decisions are subject to certain degree of error.

iii) Statistical laws do not deal with individual observations but with a group of observations.

iv) Statistical conclusions are true on an average.

v) Statistics is liable to be misused. The misuse of statistics may arise because of the use of statistical tools by inexperienced and untrained persons.

vi) Statistical results may lead to fallacious conclusions if quoted out of context or manipulated.

1.2 Some Basic Concepts:

Variable: A quantity that varies from individual to individual is called a variable. Height, weight, number of students in a college, number of petals in a flower, number of tillers in a plant etc. are a few examples of variables.

Discrete and Continuous Variables: A variable that takes only specific or distinguished values in a given range is known as discrete variable whereas a variable which can theoretically assume any value between two given values is called a continuous variable. For example, the number of students in a college, number of petals in a flower, number of tillers in a plant etc. are discrete variables. A continuous variable can take any value within a certain range, for example, yield of a crop, height of plants and birth rates etc. are continuous variables.

Qualitative and Quantitative Characters:

Those characteristics/attributes of individuals which cannot be measured numerically e.g. sex, blindness, honesty, colour etc. are called qualitative characters whereas the characteristics of the individuals which can be numerically measured, e.g. height, weight, age, yield etc. are called quantitative characters.

Raw Data and Array:

The collected data which have not been processed or organized numerically is known as raw data while, the raw data arranged in ascending or descending order of magnitude is called an array.

Primary and Secondary Data:

The data collected directly from the original source are called the primary data. Such data may be collected by sample surveys or through designed experiments. The data, which have already been collected by some agency and have been processed or used at least once are called secondary data. Secondary data may be collected from organizations or private agencies, government records, journals etc.

Classification of Data: It is process of arranging the data into a number of classes or groups on the basis of their resemblances and similarities. It is of four types:

i) **Geographical Classification:** The data may be classified according to geographical or locational differences such as regions, states, districts, cities etc.

for example, data on sale of automobiles in different states and districts in India in a particular year.

ii) **Chronological Classification:** Here the data are arranged according to time started from the some initial time period, For example, the data on sale of automobile in India over the last 10 years.

iii) **Qualitative Classification:** This type of classification is applicable for qualitative data and the data are classified according to some characteristic or attribute such as religion, sex, employment, national origin etc. The attributes cannot be measured but can be categorized e.g. the population of a town may be classified as follows:

Sex / **Employment Status**	**Male**	**Female**	**Total**
Employed	4600	940	**5540**
Unemployed	510	3950	**4460**
Total	**5110**	**4890**	**10,000**

iv) **Quantitative Classification:** It is for quantitative data like data on weight, height, income etc. of individuals. Here the construction of a frequency distribution is required.

Difference between Classification and Tabulation:

i) The classification is the basis of tabulation while tabulation is a mechanical function of classification.

ii) The classification divides the data into homogeneous groups and subgroups with regard to the similarity of the characteristics under study while tabulation arranges the classified data into rows and columns with regard to time, size, aim and importance of data.

iii) The classification is a technique of statistical analysis while tabulation is a technique of presenting the data.

iv) Classification facilitates the comparison between the two data sets while tabulation makes a comparison very easy through the use of ratios, percentages and coefficients etc.

Tabulation: It is the process in which the data are put in a table having different rows and columns. A table, which contains data related to one characteristic, is called a simple table. On the other hand, a table which contains data related to more than one characteristic, is called a complex table.

1.3 Frequency Distributions and Their Construction:

Frequency Distribution:

The number of observations lying in any class interval is known as the frequency of that class interval. Also the number of times an individual item is repeated in a series is called its frequency. The way in which the observations are classified and distributed in the proper class intervals is known as frequency distribution.

Relative Frequency:

It is the proportion of the number of observations belonging to a class and is obtained by dividing the frequency of that class by the total frequency.

Cumulative Frequency:

The cumulative frequency corresponding to any value or class is the number of observations less than or equal to that value or upper limit of that class. It may also be defined as the total of all frequencies up to the value or the class.

Cumulative Frequency Distribution:

It is an arrangement of data in class intervals together with their cumulative frequencies. In less than cumulative frequency distribution, the frequency of each class is added successively from the top to bottom but in more than type, the frequencies of each class are added successively from bottom to top.

Rules for Construction of Frequency Distribution:

i) The number of classes should preferably be between 5 and 15, however, there is no rigidity about it and it depends upon total frequency and the range of the data. The following formula suggested by H.A. Sturges may be used for finding approximate number of class intervals and their width.

$$h = \frac{L - S}{K}$$

where N = total frequency ; $K = 1 + 3.322 \log_{10} N$ is the number of classes. L and S are the largest and smallest observation in the data and h = class width.

ii) The class limits should be well defined so that one can place an observation in a class without any confusion.

iii) As far as possible, the class intervals should be of equal size.

iv) Counting of number of observations belonging to a class gives the frequency of the class.

Discrete (ungrouped) Frequency Distribution:

When the number of observations in the data are small, then the listing of the frequency of occurrence against the values of variable is called as discrete frequency distribution.

Example-1: The number of seeds per pod in 50 pods of a crop variety is given below: Prepare a discrete frequency distribution.

9, 2, 3, 1, 4, 5, 2, 6, 2, 3, 8, 9, 7, 6, 5, 4, 1,
3, 2, 7, 5, 4, 5, 4, 3, 8, 7, 5, 4, 3, 6, 5, 3, 4, 8, 6,
8, 5, 4, 7, 3, 4, 5, 6, 9, 8, 5, 1, 4, 5

Solution: The number of seeds can be considered as a variable (X) and the number of pods as the frequency (f).

No. of Seeds (X)	Tally Marks	No. of Pods (f)
1	III	3
2	IIII	4
3	~~IIII~~ II	7
4	~~IIII~~ IIII	9
5	~~IIII IIII~~	10
6	~~IIII~~	5
7	IIII	4
8	~~IIII~~	5
9	III	3
Total:	**N**	**50**

Grouped (Continuous) Frequency Distribution:

When the data set is very large it becomes necessary to condense the data into a suitable number of class intervals of the variable alongwith the corresponding frequencies. The following two methods of classification are used

Exclusive Method: When the data are classified in such a way that the upper limit of a class interval is the lower limit of the next class interval, then it said to be the exclusive method of classification i.e. upper limits are not included in the class interval.

Inclusive Method: When the data are classified in such a way that both lower and upper limits are included in the class interval, then it said to be inclusive method of classification.

Remarks:

1. An exclusive method should be used for the continuous data and inclusive method may be used for discrete data.
2. If the continuous data are classified according the inclusive method, there is need to find class boundaries to maintain continuity let d be the difference between the upper limit of a class and lower limit of the next class, then

   lower class boundary = lower limit - 0.5d.

   and upper class boundary = upper limit + 0.5d.

   i.e. to obtain the class boundaries, take the difference (d) between 20 and 19 which is one. Thus d/2 = 0.5 Now deduct 0.5 from the lower limits and 0.5 to the upper limits.

Table: Class intervals showing Dividend declared by 45 companies

Exclusive method	Inclusive Method	Class Boundaries	Frequency (Number of Companies)
10-20 (10 but less than 10)	10-19	9.5-19.5	7
20-30 (20 but less than 30)	20-29	19.5-29.5	13
30-40 (30 but less than 40)	30-39	29.5-39.5	15
40-50 (40 but less than 50)	40-49	39.5-49.5	10

Example-2: The marks obtained by 30 students are given below. Classify the data using Sturges rule

30, 32, 45, 52, 47, 52, 58, 63, 59, 75, 49, 55, 77,

28, 26, 33, 47, 45, 59, 73, 75, 65, 55, 68, 67, 79, 35, 39, 68, 75

Let us find the suitable number of class intervals with the help of Sturges rule

No. of classes (K) = $1 + 3.322 \log_{10} N = 1 + 3.322 \log_{10} 30 = 1 + (3.322 \times 1.477)$

$= 1 + 4.91 = 5.91 \simeq 6$

$$\text{Class width (h)} = \frac{L - S}{K} = \frac{79 - 26}{6} = \frac{53}{6} = 8.8 \simeq 9$$

Thus number of classes will be 5.91 (=6) and size of class interval will be 9

We take 10 as the size of the class interval. Since the minimum value is 26, therefore, the first class interval is taken as 25-35.

Marks Obtained (Class Interval)	Tally Marks	No. of Students (f)
25-35	~~HH~~ I	5
35-45	II	2
45-55	~~HH~~ II	7
55-65	~~HH~~ I	6
65-75	~~HH~~	5
75-85	~~HH~~	5
	N	**30**

1.4 Measures of Central Tendency:

Different observations have a tendency to concentrate around a central point in a data series which is often called central tendency. A central tendency or an average can also be defined as a single value within the range of the data that represents all the values in the data series. Since an average is somewhere within the range of the data, it is sometimes called a measure of central value or location of a distribution.

Arithmetic mean, geometric mean, harmonic mean, median and mode are the popular measures of central tendency. The arithmetic mean, geometric mean and harmonic mean are known as mathematical averages while median and mode are positional averages.

Properties of a Good Measure of Central Tendency:

i) It should be rigidly defined, based on all the observations and easy to calculate and understand.

ii) It should be suitable for further mathematical treatment.

iii) It should be least affected by extreme values and fluctuations in sampling.

Arithmetic Mean:

Arithmetic mean of a group of n observations $x_1, x_2, ..., x_n$ is defined as

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum x_i$$

If the observations $x_1, x_2, ..., x_n$ form a discrete frequency distribution with respective frequencies $f_1, f_2, \cdots\cdots, f_n$ then arithmetic mean is given by:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + ... + f_nx_n}{N} = \frac{1}{N}\sum f_i x_i \text{ where } N = \Sigma f_i$$

For grouped frequency distribution, mid values are determined for the various class intervals and then arithmetic mean is computed as in case of a discrete frequency distribution.

Properties of Arithmetic Mean:

i) Arithmetic mean is rigidly defined and based on all the observations.

ii) It is suitable for further mathematical treatment and is least affected by fluctuations in samplings.

iii) Sum of deviations of the given values from their arithmetic mean is always zero.

iv) Sum of the squares of deviations of the given values from their arithmetic mean is always minimum.

Demerits of Arithmetic Mean:

i) Arithmetic mean cannot be used as a suitable measure of central tendency if we are dealing with qualitative characteristics or in case of extremely asymmetrical distributions.

ii) Arithmetic mean is likely to be affected by extreme values.

iii) Arithmetic mean cannot be calculated in case of open-ended classes.

Pooled or Combined Arithmetic Mean: If $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ are the arithmetic means of k groups, based on $n_1, n_2, \cdots\cdots, n_k$ observations respectively then combined mean of all ($n_1 + n_2 + ..$é n_k) observations of the k groups taken together is:

$$\bar{x}_p = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + ... + n_k\bar{x}_k}{n_1 + n_2 + ... + n_k}$$

Weighted Mean: When different values have unequal weightage or importance or contributions then instead of simple mean the weighted mean is used. Let $x_1, x_2, \cdots\cdots, x_k$ be k values with weights $w_1, w_2, \cdots\cdots, w_k$ respectively then weighted mean is

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + ... + w_kx_k}{\sum w_i} = \frac{\sum w_i x_i}{\sum w_i}$$

Example-3: Find over all grade point average (OGPA) in a semester from the following:

Course	**Cr. Hrs. (w_i)**	**Grade point (x_i)**	**$w_i x_i$**
Stat-101	2+1	6.3	18.9
Comp-101	1+1	7.0	14.0
Math-101	3+1	7.5	30.0
Total:	**6+3**		**62.9**

OGPA is $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{62.9}{9} = 6.99$

Geometric Mean: Geometric mean is more appropriate if the observations are measured as ratios, proportions, growth rates or percentages. When the growth rates or increase in production etc. are given for a number of years or periods then geometric mean should be used as a measure of central tendency. If $x_1, x_2, \cdots, x_n$ be the n positive observations then $G = (x_1 x_2 \ldots é \ x_n)^{1/n}$ or $\log G = \frac{1}{n}\sum \log x_i$ For a frequency distribution, the geometric mean is defined as

$$G = (x_1^{f1}.x_2^{f2} é \ .. \ x_n^{fn})^{1/N} \quad \text{or} \quad \log G = \frac{1}{N}\sum f_i \log x_i \text{ ,where } N = \Sigma f_i$$

Properties of Geometric Mean:

i) It is rigidly defined and is based on all the observations.

ii) It is suitable for further mathematical treatment. If n_1 and n_2 are the sizes of two data series with G_1 and G_2 as their geometric means respectively, then geometric mean of combined series G is given by:

$$\text{Log } G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

iii) It is not affected much by fluctuations in sampling.

Demerits:

i) Geometric mean is a bit difficult to understand and to calculate for a non-mathematician.

ii) It cannot be calculated when any value is zero or negative and gives an absurd value if computed in case of even number of negative observations

iii) Like arithmetic mean it is also affected by the extreme values but to a lesser extent.

Harmonic Mean: The reciprocal of the arithmetic mean of reciprocals of non-zero values of a variable is called harmonic mean (H). Harmonic mean is a suitable average when observations are given in terms of speed rates and time.

If $x_1, x_2, \cdots\cdots, x_n$ are n observed values, the harmonic mean is given by:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}}$$

When observations are given in the form of a frequency distribution, then harmonic mean is given by:

$$H = \frac{N}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots\ldots + \frac{f_n}{x_n}} = \frac{N}{\sum \frac{f_i}{x_i}} \text{ where } N = \sum f_i$$

Harmonic mean for a grouped frequency distribution can be obtained by first calculating the class marks.

Properties of Harmonic Mean:

i) Harmonic mean is rigidly defined and is based on all observations.

ii) It is suitable for further mathematical treatment.

iii) Like geometric mean it is not affected much by fluctuation of sampling.

Demerits:

i) Harmonic mean is not easily understood and is difficult to compute.

ii) It gives greater importance to small items.

iii) It cannot be calculated when any observation is zero.

Relation among Arithmetic Mean (A), Geometric Mean (G) and Harmonic Mean (H):

i) $G = \sqrt{A \times H}$

ii) $A \geq G \geq H$, equality holds when all observations have same magnitude.

Median: The median of distribution is that value which divides it into two equal parts. Median is a positional average and is the suitable measure of central tendency when individuals are ranked on the basis of some qualitative characteristics such as intelligence, poverty etc., which cannot be measured numerically. It is also used when

open-ended classes are given at one or both the ends or data arise from some skewed distribution such as income distribution.

Median is computed by arranging the observations in ascending or descending order of magnitude. If n, the number of observations is odd, then the size of $(n+1)/2^{th}$ observation will be the median and if n is even then average of two middle observations is the median.

Median for discrete frequency distribution is the observation corresponding to the cumulative frequency (less than type) just greater than N/2.

For a grouped frequency distribution the class corresponding to the cumulative frequency just greater than N/2 is called the median class and median is determined by the formula:

$$M_d = L + \frac{h}{f}(N/2 - C)$$

where L: is the lower limit of the median class

f: is the frequency of the median class

h: is the size of the median class

C: is the cumulative frequency of the class preceding the median class

N $= \Sigma f$

Merits: It is rigidly defined, easy to calculate and easy to understand. It is a positional average and hence not affected by extreme values. It can be calculated for open-ended distribution.

Demerits: Median is not based on all the observations. It is not suitable for further treatment and is more sensitive to the fluctuations of sampling.

Mode: Mode is defined as the value, which occurs most frequently in a set of observations and around which other observations of the set cluster densely. Mode of a distribution is not unique. If two observations have maximum frequency then the distribution is bimodal. A distribution is called multi-modal if there are several values that occurs maximum number of times. Mode is often used where we need the most typical value e.g. in business forecasting the average required by the manufacturers of the sizes of readymade garments, shoes etc.

Mode of a discrete frequency distribution can be located by inspection as the variable value corresponding to maximum frequency. For a continuous frequency distribution we first calculate the modal class and then mode is determined by the following formula:

$$\text{Mode} = L + h \times \frac{f_m - f_1}{2f_m - f_1 - f_2}$$

Here L, h, f_m, f_1 and f_2 are respectively the lower limit of the modal class, the size of modal class, frequency of the modal class, frequency preceding the modal class and frequency following the modal class.

If there are irregularities in the distribution or the maximum frequency is repeated or the maximum frequency occurs in the very beginning or at the end of the distribution then the mode is determined by grouping method.

Merits: Mode is not affected by extreme observations and can be calculated for open ended distributions.

Demerits: It is ill defined, not based on all the observations, not unique and often does not exit.

Relationships among Mean, Mode and Median:

i) For a symmetrical distribution Mean = Median = Mode

ii) For a skewed distribution Mean - Mode = 3(Mean ï Median) or
Mode = 3 Median ï 2 Mean

iii) For a positively skewed distribution Mean > Median >Mode

iv) For a negatively skewed distribution Mean < Median <Mode

Partition Values or Quantiles: Some times we not only need the mean or middle value but also need values which divide the data in four or ten or hundred equal parts.

Quartiles: Are the three values which divide data into four equal parts. These are denoted by Q_1, Q_2, Q_3

On the lines of median formula, quartiles are given by

$$Q_i = L + \frac{iN/4 - C}{f} \times h \;\; ; \;\; i = 1, 2, 3$$

where L, C and f are lower limit, C.F. of preceding class and frequency of the class containing that quartile. Q_1 and Q_3 are called lower and upper quartiles respectively and Q_2 is equal to the median.

Deciles: Are the nine values which divide the data into 10 equal parts and deciles are denoted by D_1, D_2, ·····, D_9; where

$$D_i = L + \frac{iN/10 - C}{f} xh \;\; ; \;\; i = 1, 2, ..., 9$$

Percentiles: Are the ninety nine values which divide the data into 100 equal parts and are denoted by P_1, P_2, ·····, P_{99}; where

$$P_i = L + \frac{iN/100 - C}{f} xh \;\; ; \;\; i = 1, 2, ..., 99$$

Percentiles help to find the cut off values when the data be divided into a number of categories and per cent of observations in various categories are given.

1.5 Measures of Dispersion:

For comparing two sets of data or distributions, comparison of average values may not give complete picture as it may be possible that two sets of data or distributions may have the same mean but they may differ in dispersion or scatter or spread. Hence we must compare scatter/dispersion in addition to comparison of locations or means.

The degree to which numerical data tend to scatter or spread around the central value is called dispersion or variation. Further, any quantity that measures the degree of spread or scatter around the central value is called measure of dispersion. The various measures of dispersion are:

i) Range

ii) Quartile Deviation

iii) Mean Deviation

iv) Variance

v) Standard Deviation

vi) Quartile Coefficient of Dispersion

vii) Coefficient of Mean Deviation

viii) Coefficient of Variation

Absolute and Relative Measures of Dispersion:

A measure of dispersion indicates the degree to which the numerical data tend to spread or scatter around an average value. The measures expressed in terms of the units of the data are called absolute measures. Range, quartile deviation, mean deviation and

standard deviation are the common examples of absolute measures. The measures, which are independent of the units of measurement are called the relative measures. These are pure numbers and often expressed as percentages. Quartile coefficient of dispersion, coefficient of mean deviation and coefficient of variation are a few examples of relative measures.

Range: It is defined as the difference of the two extreme observations of a data set. Range is used when we need a rough comparison of two or more sets of data or when the observations are too scattered to justify the computation of a more precise measure of dispersion.

Merits and Demerits of Range: Range is rigidly defined and is the simplest measure of dispersion. It is also easy to interpret and calculate. Range is a crude and unreliable measure of dispersion and it being based only on two extreme observations and has greater chances of being affected by fluctuations in sampling.

Quartile Deviation or Semi-Quartile Range (QD): It is mathematically defined as

$$QD = (Q_3 - Q_1)/2$$

Quartile deviation is preferred when distribution is skewed or open-ended. It is also used when error caused by extreme values is to be minimized.

Merits and Demerits of QD: It is rigidly defined and easy to calculate and understand. It is not affected by the extreme values. The main demerits of quartile deviation are that it is neither based on all the observations nor suitable for further mathematical treatment. It is also sensitive to the fluctuations in sampling.

Mean Deviation (MD): The arithmetic mean of the absolute deviations about any point A is called the mean deviation or mean absolute deviation about the point A. The point óAô may be taken as mean, median or mode of the distribution. It is a useful measure of dispersion in business and economics when extreme observations influence the standard deviation unduly.

The MD of n observations $x_1, x_2, x_3, \cdots\cdots, x_n$ about any point A is given by

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$$

If a frequency distribution is given then,

$$MD = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - A| \qquad \text{where } N = \sum f_i$$

Merits: Mean deviation is rigidly defined, based on all the observations, easy to calculate and relatively easy to interpret. It is not affected much by the extreme values.

Demerits: Mean deviation is not suitable for further mathematical treatment. It also ignores the signs of deviations and hence creates some artificiality in the result.

Variance: The arithmetic mean of the squared deviations taken about mean of a series is called variance ($\mathring{u}^2$). Mathematically, variance is defined as:

$$\mathring{u}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^{\,2} - \bar{x}^2$$

For a frequency distribution, the variance is given by:

$$\mathring{u}^2 = \frac{1}{N}\sum f_i(x_i - \bar{x})^2 = \frac{1}{N}\sum f_i x_i^{\,2} - \bar{x}^2 \qquad \text{where } N = \sum f_i \text{ and } \bar{x} = \frac{\sum f_i x_i}{N}$$

Standard Deviation: The positive square root of the arithmetic mean of the squared deviations of observations in a data series about its arithmetic mean is called standard deviation ($\mathring{u}$) i.e. it is the square root of variance i.e.

$$\mathring{u} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum x_i^{\,2} - \bar{x}^2} \text{ and for a frequency distribution}$$

$$\mathring{u} = \sqrt{\frac{1}{N}\sum f_i(x_i - \bar{x})^2} = \sqrt{\frac{1}{N}\sum f_i x_i^{\,2} - \bar{x}^2}$$

Standard deviation is a stable measure and is regarded as the best and most powerful measure of dispersion.

Merits and Demerits of Standard Deviation:

Standard deviation is rigidly defined and based on all the observations. It is relatively less sensitive to the sampling fluctuations and also suitable for further mathematical treatment. Standard deviation is independent of change of origin but not of scale. The main demerit of standard deviation is that it gives greater weightage to extreme values and cannot be calculated in case of open-ended classes.

Shortcut method for computing AM and variance

AM and Variance are the most frequently used measures of central tendency and dispersion respectively and they are also used in finding the C.V. which is a relative measure of dispersion. Shortcut method is given in the following steps:

i) Find the mid values of class intervals if not given. Let $x_1, x_2, \cdots, x_n$ be the mid values.

ii) For every class interval find $u_i = (x_i$ ï $A)/h$ where A = assumed mean and h = width of the class interval.

iii) Find $\Sigma f_i u_i$ and $\Sigma f_i u_i^2$ and

$$\text{AM}(\bar{x}) = A + h\,\bar{u} = A + \frac{\sum f_i u_i}{N} \text{ x } h$$

$$\text{Variance } (\sigma_x^2) = h^2\left[\sum f_i u_i^2 / N - (\sum f_i u_i / N)^2\right]$$

Relative Measures of Dispersion:-

i) Quartile Coefficient of Dispersion (QCD) = $\frac{Q_3 - Q_1}{Q_3 + Q_1} x100$

ii) Coefficient of Mean Deviation = $\frac{\text{MD about A}}{A} x100$, where A = mean or median or mode

iii) Coefficient of Variation (CV) = $\frac{SD}{\text{Mean}} x\,100$

Relative measures of dispersion are used while comparing the variability of series having different or same units of measurement. They can also be used for comparing two or more series for consistency. A series with smaller coefficient of dispersion is said to be less dispersed or more consistent (or homogeneous).

Combined Standard Deviation: Combined standard deviation of two groups is denoted by σ_p and is computed follows:

$$\mathring{u}_p = \sqrt{\frac{n_1\mathring{u}_1^2 + n_2\mathring{u}_2^2 + n_1 d_1^2 + n_1 d_2^2}{n_1 + n_2}}$$

where σ_1 = SD of first group

σ_2 = SD of second group

$d_1 = (\overline{X}_1 - \overline{X}_p)$; $d_2 = (\overline{X}_2 - \overline{X}_p)$

$\overline{X}_1$, $\overline{X}_2$, $\overline{X}_p$ are the group means and pooled mean respectively.

The above formula can be extended to find the combined SD of three or more groups.

1.6 Skewness and Kurtosis:

Skewness: Literally means lack of symmetry, skewness gives an idea about the shape of the curve that can be drawn with the help of the given data. The frequency curve of a skewed distribution is not symmetrical but stretched more to one side than to the other.

Skewness is often measured by the Karl-Pearson Coefficient of skewness defined as:

$$S_k = \frac{3(\text{Mean - Median})}{\text{Standard Deviation}} = \frac{\text{Mean - Mode}}{\text{Standard Deviation}}$$

Since (M ï M_d)/ů lies between ± 1, hence S_k lies between ± 3.

While in terms of moments $ƃ_1 = \frac{ε_3^2}{ε_2^3}$ and $ɔ_1 = \sqrt{ƃ_1}$

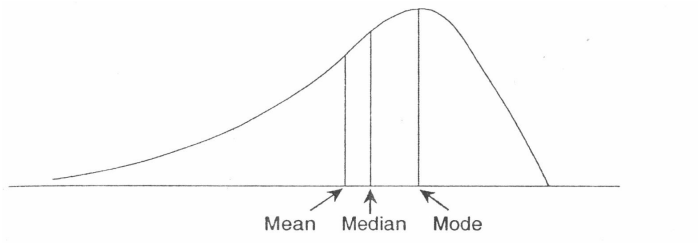
where $\mu_r = (1/N)\sum f_i (x_i - \bar{x})^r$ is r^{th} order central moment of the variable X and sign of $ɔ_1$ is same that of μ_3.

Graphically it can be shown as under:

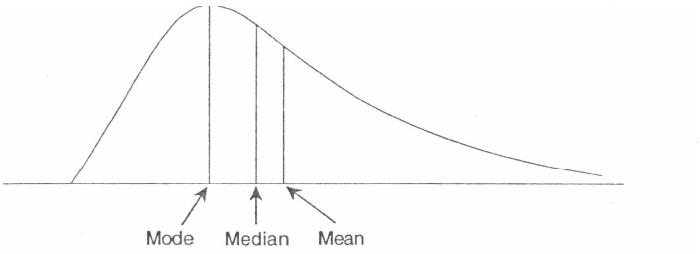
If $ɔ_1 = 0 \Rightarrow$ curve is normal (symmetrical)

$ɔ_1 < 0 \Rightarrow$ curve is skewed to the left (negative skewness)

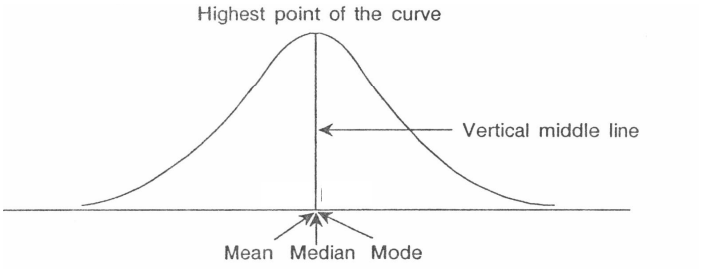
$ɔ_1 > 0 \Rightarrow$ curve is skewed to the right (positive skewness)



Graph showing negative skewness (the curve stretched more to the left)



Graph showing positive skewness (the curve stretched more to the right)



Graph showing normal (symmetrical) curve

Remarks:

i) For a symmetrical distribution mean, mode and median are equal

ii) Skewness is a pure number having no units of measurement and thus can be used to compare the skewness in sets of data with same or different units of measurements.

iii) For a positively skewed distribution AM > Median > Mode and for a negatively skewed distribution AM < Median < Mode.

Kurtosis: The flatness or peakedness of top of the curve is called kurtosis, which is also an important characteristic of a distribution. Kurtosis is measured in terms of moments and is given by:

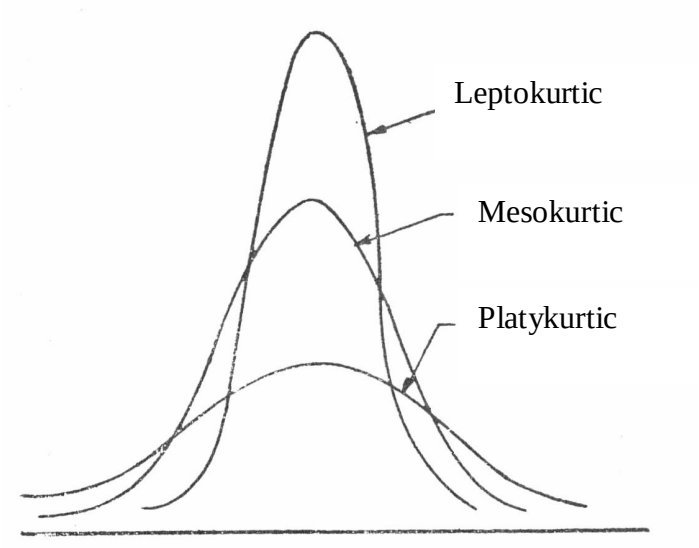
$\beta_2 = \mu_4 / \mu_2^2$ and $\gamma_2 = \beta_2 - 3$

For normal distribution curve $\beta_2 = 3$ (or $\gamma_2 = 0$) and flatness or peakedness of the curve of a distribution is compared in relation to normal curve and the distributions have been classified as follows:

If $\beta_2 = 3 \Rightarrow \gamma_2 = 0$, then the curve is properly peaked, i.e. **Mesokurtic** curve.

$\beta_2 > 3 \Rightarrow \gamma_2 > 0$, then the curve is more peaked than normal, i.e. **Leptokurtic** curve

$\beta_2 < 3 \Rightarrow \gamma_2 < 0$, then the curve is less peaked than normal, i.e. **Platykurtic** curve



Graph showing Mesokurtic, Leptokurtic and Platykurtic Curves

Coefficient of Kurtosis is also a pure number having no units of measurements and thus can be used to compare the flatness of the top of curves for frequency distributions with same or different units.

Example-4: Calculate the AM, median, Q_1, D_4, P_{55}, mode, G.M. and H.M. for the data on salaries of 1000 employees in a company.

Salaries (000Rs)	**No. of employees (f)**	**Class mark (x)**	**fx**	**c.f.**	**log x**	**f log x**	**f/x**
1-3	50	2	100	50	0.3010	15.50	25
3-5	110	4	440	160	0.6021	66.23	27.5
5-7	162	6	972	322	0.7782	126.07	27.00
7-9	200	8	1600	522	0.9031	180.62	25.00
9-11	183	10	1830	705	1.0000	183.00	18.3
11-13	145	12	1740	850	1.0792	156.48	12.1
13-15	125	14	1750	975	1.1461	143.33	8.9
15-17	15	16	240	990	1.2041	18.06	0.9
17-19	8	18	144	998	1.2553	10.04	0.4
19-21	2	20	40	1000	1.3010	2.60	0.1
Total	**N = 1000**		**8856**			**901.93**	**145.2**

Solution:

(i) AM ($\overline{X}$) = $\sum fx/N = 8856/1000 = 8.856$ = Rs. 8856

(ii) Median = $L + \frac{N/2 - C}{f} \times h = 7 + \frac{500 - 322}{200} \times 2 = 8.78$ = Rs.8780

(iii) Lower Quartile (Q_1) = $L + \frac{N/4 - C}{f} \times h = 5 + \frac{250 - 160}{162} \times 2 = 6.111$ = Rs.6111

(iv) 4th decile (D_4) = $L + \frac{4N/10 - C}{f} \times h = 7 + \frac{400 - 322}{200} \times 2 = 7.78$ = Rs.7780

(v) 55^{th} percentile (P_{55}) = $L + \frac{55N/100 - C}{f} \times h = 9 + \frac{550 - 522}{183} \times 2 = 9.306$ = Rs. 9306

(vi) Mode = $L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h = 7 + \frac{200 - 162}{2(200) - 162 - 183} \times 2 = 8.382$ = Rs.8382

(vii) GM = Antilog ($\sum f \log x/N$) = Antilog (901.93/1000)

= Antilog (0.901) = 7.979 = Rs. 7979

(viii) HM = Reciprocal of ($1/N \sum f/x$) = Rec. of (145.2/1000) =1000/145.2

= 6.887 = Rs. 6887

Example-5: Find the average rate of increase in population which in the first decade has increased by 20%, in the second decade by 30% and in the third decade by 40%. Find the average rate of increase in the population.

Solution: To find the average rate of growth, geometric mean is the appropriate average.

Decade	% Rise	Population at the end of the decade (x)	log x
1^{st}	20	120	2.0792
2^{nd}	30	130	2.1139
3^{rd}	40	140	2.1461
			Ælog x = 6.3392

GM = Antilog (1/n Ælog x) = Antilog (6.3392/3) = Antilog (2.1130) = 129.7

Thus average rate of increase in the population is (129.7 ï 100) = 29.7 per cent per decade.

Example-6: A taxi driver travels from plain to hill station 100 km distance at an average speed of 20 km per hour. He then makes the return trip at average speed of 30 km per hour. What is his average speed over the entire distance (200 km)?

Solution: To find the average speed, harmonic mean is an appropriate average. Harmonic mean of 20 and 30 is:

$$\text{Harmonic mean} = \frac{2}{\frac{1}{20}+\frac{1}{30}} = \frac{2}{\frac{5}{60}} = \frac{2 \text{ x } 60}{5} = 24 \text{ km per hour}$$

Example-7: Find the mean deviation about AM and hence find the coefficient of mean deviation.

Daily wages (Rs.)	0-20	20-40	40-60	60-80	80-100	Total
No. of wage earners (f)	2	5	10	10	5	**N=32**
Mid value (x)	10	30	50	70	90	
fx	20	150	500	700	450	**1820**
$\lvert x - \bar{x} \rvert$	46.9	26.9	6.9	13.1	33.1	
$f\lvert x - \bar{x} \rvert$	93.8	134.5	69	131	165.5	**593.8**

AM ($\bar{x}$) = $\sum fx/N$ = 1820/32 = 56.9

MD = $\sum f\lvert x - \bar{x} \rvert /N$ = 593.8./32 = 18.55

Coefficient of $\text{MD} = \frac{\text{MD}}{\text{AM}} \times 100 = \frac{18.55}{56.9} \times 100 = 32.6\%$

Example-8: The runs scored by two batsmen X and Y in 10 innings are given below. Find out which is better runner and who is more consistent player?

											Total
X	90	110	5	10	125	15	35	16	134	10	**550**
Y	65	68	52	47	63	25	25	60	55	60	**520**
$x - \bar{x}$	35	55	-50	-45	70	-40	-20	-39	79	-45	
$(x - \bar{x})^2$	1225	3025	2500	2025	4900	1600	400	1521	6241	2025	**25462**
$y - \bar{y}$	13	16	0	-5	11	-27	-27	8	3	8	
$(y - \bar{y})^2$	169	256	0	25	121	729	729	64	9	64	**2166**

Series X: $\bar{X} = \sum x/n = 550/10 = 55$

$$\mathring{u}_x^2 = \sqrt{\Sigma(x-\bar{x})^2/n} = \sqrt{25462/10} = \sqrt{2546.2} = 50.46$$

$$CV = \frac{\mathring{u}_x}{\bar{X}} \times 100 = \frac{50.46}{55} \times 100 = 91.74\%$$

Series Y: $\bar{Y} = \sum y/n = 520/10 = 52$

$$\mathring{u}_y^2 = \sqrt{\Sigma(y-\bar{y})^2/n} = \sqrt{2166/10} = \sqrt{216.6} = 14.71$$

$$CV = \frac{\mathring{u}_y}{\bar{Y}} \times 100 = \frac{14.71}{52} \times 100 = 28.28\%$$

Conclusion: X is better runner since $\bar{X}$ = 55 is more than $\bar{Y}$ = 52 and Y is more consistent player since CV in Y series is less than CV in X series.

Example 9: The profits (in million rupees) earned by 70 companies during 2008-09 are given below. Compute the AM, SD and CV

Profit	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**Total**
No. of Companies (f)	6	20	24	15	5	**70**
Class mark (x)	15	25	35	45	55	
$(x - \bar{x})$	-19	-9	1	11	21	
$(x - \bar{x})^2$	361	81	1	121	441	
$f(x - \bar{x})^2$	2166	1620	24	1815	2205	**7830**
fx	90	500	840	675	275	**2380**
fx^2	1350	12500	29400	30375	15125	**88750**
u=(x-35)/10	-2	-1	0	1	2	
fu	-12	-20	0	15	10	**-7**
fu^2	24	20	0	15	20	**79**

AM ($\bar{x}$) = $\sum fx/N$ = 2380/70 = 34 million rupees

Method-I (when AM is a fraction)

$$\text{Variance } (\sigma_x^2) = \frac{1}{N}\Sigma fx^2 - \left(\frac{1}{N}\Sigma fx\right)^2 = 88750/70 - 34^2 = 111.86$$

$$\text{SD } (\sigma_x) = \sqrt{\text{Variance}} = \sqrt{111.86} = 10.58 \text{ million rupees}$$

Method-II (when AM is an integer)

$$\text{Variance } (\sigma_x^2) = \frac{1}{N}\Sigma f(x - \bar{x})^2 = 7830/70 = 111.86$$

$$\text{SD } (\sigma_x) = 10.58$$

Method –III (Shortcut method)

Consider u = (x-A)/h where A (assumed mean) = 35 and h (width of class interval) = 10

$$AM = A + \frac{\Sigma fu}{N} \times h = 35 + \left(\frac{-7}{70}\right) \times 70 = 35 - 1 = 34$$

Variance (σ_x^2) = $h^2 [\Sigma fu^2/N - (\Sigma fu/N)^2] = 100[79/70 - (-7/70)^2]$

= $100 [1.1286 - 0.01] = 100(1.1186) = 111.86$

SD $(\sigma_x) = \sqrt{111.86}$ = 10.58

Example-10: The data regarding daily income of families in two villages are given below:

	Village A	**Village B**
Number of families	600	500
Average income (Rs.)	175	186
Variance (Rs^2)	100	81

i) In which village there is more variation in income.

ii) What is the combined standard deviation of income of two villages?

Solution:

i) Village A CV = $\frac{SD}{AM} \times 100 = \frac{10}{175} \times 100 = 5.71\%$

Village B CV = $\frac{9}{186} \times 100 = 4.84\%$

Since CV in village A is more than village B, therefore, there is more variation of income in village A.

ii) The Combined SD of income in both the villages A and B is given by

$$\mathring{u}_p = \sqrt{\frac{n_1\mathring{u}_1^2 + n_2\mathring{u}_2^2 + n_1d_1^2 + n_2\mathring{u}_2^2}{n_1 + n_2}}$$

Here $d_1 = \overline{X}_1 - \overline{X}_p$; $d_2 = \overline{X}_2 - \overline{X}_p$

$$\overline{X}_p \text{ (pooled mean)} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2} = \frac{600x175 + 500x186}{600 + 500}$$

$$= \frac{105000 + 93000}{1100} = \text{Rs.}180$$

$d_1 = |175 \text{ï } 180| = 5$; $d_2 = |186 \text{ï } 180| = 6$

$$\mathring{u}_p = \sqrt{\frac{600x100 + 500x81 + 600(25) + 500(36)}{1100}} = \sqrt{\frac{133500}{1100}}$$

$$= \sqrt{121.36} = \text{Rs.}11.02$$

1.7 Diagrammatic and Graphical Representation of Data:

Numerical data may be represented in a simple and attractive manner in the form of diagrams and graphs. Diagrammatic representation is used when data relating to different times and places are given and are independent of one another. Graphical representation is used when we have to represent the data of frequency distribution or of a time series. Diagrammatic representation includes bar diagram, rectangular diagram and pie diagram, whereas a graphical representation includes frequency graphs such as histogram, frequency polygon, and ogive.

Simple Bar Diagram:

It is diagram of rectangles where each rectangle has some value and with following features:

i) Length or height of the bar varies with value of the variable under study.

ii) Bars are of the same width, equidistant from each other and are based on the same line.

Rectangular Diagram: In a rectangular diagram both length as well as width of the rectangles, are taken into consideration. The rectangular diagram is also called two-dimensional diagram. This diagram is used when two sets of data with different subdivisions are to be compared.

Pie Diagram: A pie diagram is circle divided into sectors indicating the percentages of various components, such that the areas of these sectors are proportional to the shares of components to be compared. Pie diagram is useful when percentage distribution is presented diagrammatically.

Histogram: A histogram is a two dimensional diagram used to represent a continuous frequency distribution. For drawing a histogram we first mark along the x-axis all the class intervals and frequencies along the y-axis according to a suitable scale. With class intervals as bases we draw rectangles whose areas are proportional to the frequencies of the class intervals. If the class intervals are equal then the length of rectangles will be proportional to the corresponding class frequencies.

Frequency Polygon and Frequency Curve: A frequency polygon is obtained if we plot mid values of the class intervals against frequencies and the points are joined by means of straight lines. If we join one mid value before the first group and one mid value after the last group then areas under the frequency polygon and the corresponding histogram are equal. A frequency polygon can also be obtained by joining the mid points of the upper sides of rectangles in the histogram along with one mid value before the first class interval and one mid value after the last class interval.

A frequency curve is obtained if we join the mid points by means of a free hand curve. Frequency polygon is observations used to represent a discrete frequency distribution.

Ogive: An ogive is a cumulative frequency curve in which cumulative frequencies are plotted against class limits. There are two types of ogives.

Less Than Ogive: First we form a less than type cumulative frequency distribution. We then plot the upper limits of the classes along x-axis and the less than cumulative frequencies along y-axis. These points are joined by a free hand curve. This curve is called less than ogive or less than cumulative frequency curve.

More Than Ogive: First we form a more than type cumulative frequency distribution. Then we plot the lower limits of the classes along x-axis and more than cumulative frequency along the y-axis. The points are joined by free hand curve. This curve is known as ôMore than cumulative frequency curve or more than ogive.

Graphical Representation: Graphs are used to represent a frequency distribution which makes the data understandable and attractive. It also facilitates the comparison of two or more frequency distributions. In addition to it, some of the statistical measures like mode, median and partition values can be located through graphs. The following types of graphs are generally used.

Limitations of Graphical Representation: The graphs cannot show all those facts, which are available in the tables. They also take more time to be drawn than the tables.

Graphical Location of Mode:

Mode in a frequency distribution is located graphically by drawing the histogram of the data. The steps are:

i) Draw a histogram of the given data

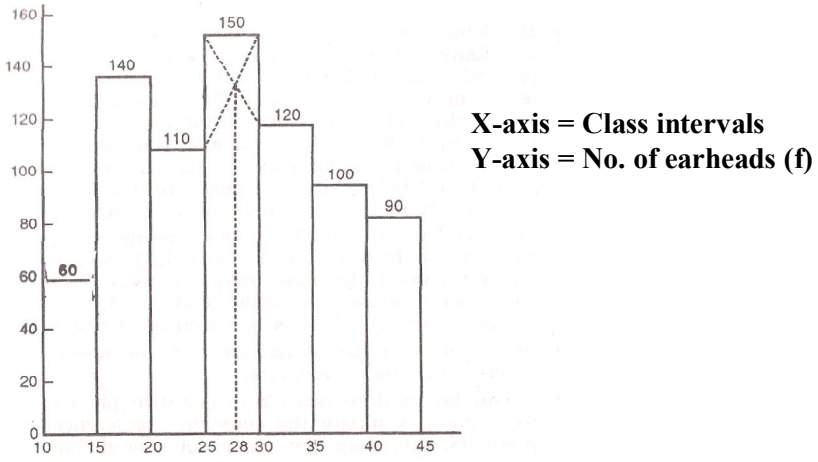
ii) Draw two lines diagonally in the inside of the modal class starting from each upper corner of the bar to the upper corner of the adjacent bar

iii) Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis which gives the modal value.

Example-11: Draw a histogram for the following distribution of earhead weights and locate the modal weight of the earheads from histogram and check by direct calculation.

Weight (in gms)	**10-15**	**15-20**	**20-25**	**25-30**	**30-35**	**35-40**	**40-45**
No. of earheads (f)	60	140	110	150	130	100	90

Solution: The histogram of this data is given below



Thus modal weight of earheads = 28 gms

Direct Calculation: Mode lies in the class 25-30

Modal Weight $= L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h = 25 + \frac{150 - 110}{2(150) - 110 - 120} \times 5$

$$= 25 + \frac{40}{70} \times 5 = 25 + 2.86 = 27.86 \simeq 28 \text{ gms}$$

Graphical Location of Median and other Partition Values: Median is graphically located from the cumulative frequency curve i.e. ogive. The various steps are:

Step-I: Draw less than or more than cumulative frequency curve i.e. Ogive.

Step-II: Mark a point corresponding to N/2 on the frequency axis (i.e. y-axis) and from this point draw a line parallel to x- axis which meets the Ogive at the point A (say).

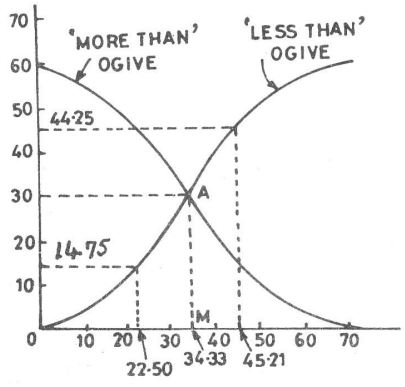
Step-III: From A, draw a line parallel to y-axis, which meets the x-axis at the point M (say). Then distance of the point M from the origin on the x- axis is the median.

Note: The other partition values viz. quartiles, deciles and percentiles can also be located graphically as described above.

Example-12: The bonus (%) announced by 59 steel industries in India is given below. Draw the less than and more than Ogives and locate the median, Q_1 and Q_3.

Bonus (%)	No. of industries	Less than c.f.	More than c.f.
0-10	4	4	59
10-20	8	12	55
20-30	11	23	47
30-40	15	38	36
40-50	12	50	21
50-60	6	56	9
60-70	3	59	3

Solution: The less than and more than Ogives are drawn below



X-axis: Bonus (%)
Y-axis: No. of industries

To locate the median, mark a point corresponding to N/2 along the Y-axis. At this point, draw a line parallel to X-axis meeting the Ogive at the point A. From A draw perpendicular to the X-axis meeting at M. The abscissa of M gives the median.

For quartiles Q_1 and Q_3, mark the points along the Y-axis corresponding to N/4 and 3N/4 and proceed as above. From the graph we see that

Median = 34.33, Q_1 = 22.50 and Q_3 = 45.21. Other partition values viz. deciles and percentiles can be similarly located.

Remark: Median can also be located by drawing a perpendicular from the point of intersection of the two Ogives as shown.

1.8 Box Plot (or Box-Whisker Diagram):

Box plot introduced by õTukeyô is a graphical representation of numerical data and based upon the following five number summary:

i) Smallest observation (Sample minimum)

ii) Lower quartile (Q_1)

iii) Median (M or Q_2)

iv) Upper quartile (Q_3)

v) Largest observation (Sample maximum)

It is an excellent tool for conveying valuable information about some descriptive measures like central tendency, dispersion, skewness etc. in data sets.

Box plot has a scale in one direction only. A rectangular box is drawn extending from the lower quartile (lies on the lower side of the rectangle) to the upper quartile (lies on the upper side of the rectangle). Thus the box plot represents the middle 50% of the data. The horizontal line within the box represents the median value. The vertical lines (whiskers) are then drawn extending from above and below the box to the largest and smallest values, respectively. Thus, the relative positions of the components illustrate how the data is distributed.

Box plots display differences between populations without making any assumption of the underlying statistical distribution and hence they are non-parametric. The spacing between the different parts of the box indicates the degree of dispersion and skewness present in data and identify outliers.

Single and Multiple Box Plots: A single box lot can be drawn for one data set. Alternately multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set. For a single box plot, the length of the box is arbitrary. For multiple box plots, the length of the box can be set proportional to the number of observations in a given group or sample. The stepwise procedure is given below:

i) Take the response variable on the vertical axis and factor of interest on the horizontal axis.

ii) Calculate M, Q_1, Q_3 and locate the largest and smallest observations.

iii) Draw the rectangular box with Q_3 and Q_1 lies on the upper and lower side of the box respectively.

iv) Draw the vertical lines (whiskers) extending from the upper and lower sides of the box to the largest and smallest values.

Detection of Outliers and Skewness: In refined box plots, the whiskers have a length not exceeding 1.5 x inter-quartile length i.e. 1.5 x (Q_3 ï Q_1). Any values beyond the ends of the whiskers are detected as outlier.

Skewed to the Left: If the box plot shows the outliers at the lower range of the data (below the box), the mean (+) value is below the median, the median line does not evenly divide the box and the lower tail of box plot is longer than the upper tail, than the distribution of data may be skewed to the left or negatively skewed.

Skewed to the Right: If the box plot shows the outliers at the upper range of the data (above the box), the mean (+) value is above the median, the median line does not evenly divide the box, and the upper tail of the box plot is longer than the lower tail, then the distribution of data may be skewed to the right or positively skewed.

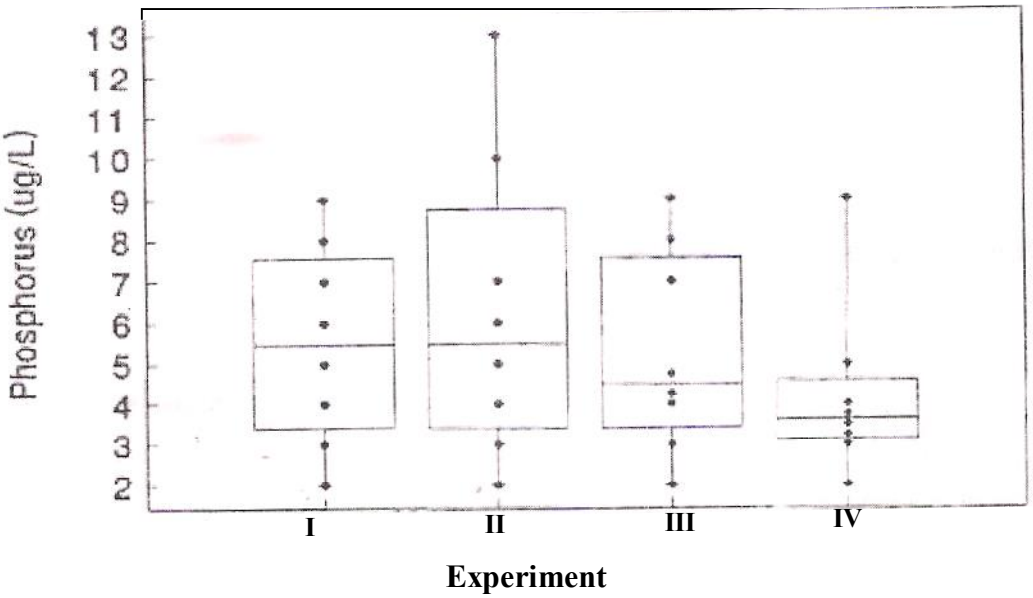
Example-13: Consider the following data on phosphorus (μg/l) recorded in four experiments:

Experiment	**Observations**
I	2, 3, 4, 5, 6, 7, 8 and 9
II	2, 3, 4, 5, 6, 7, 10 and 13
III	2, 3, 4, 4.25, 4.75, 7, 8 and 9
IV	2, 3, 3.25, 3.5, 3.75, 4, 5 and 9

Draw the box plots of the above data sets describing the relative positions of median, quartiles, maximum and minimum observations and draw your conclusions about the distribution of data.

Solution:

Box plots of four data sets (I-IV)



Experiment-1: Here the median = 5.5, Maximum and minimum values are 2 and 9 giving a range of 7. It is a symmetrical distribution since

i) The maximum and minimum observations are at equal distances from the median

ii) The maximum and minimum observations are at equal distances from the box

iii) Median lies exactly in the centre of the box

Experiment-II: Here the median is = 5.5, Maximum and minimum values are 2 and 13 giving a range of 11. It is a skewed distribution since

i) The maximum and minimum observations are not at equal distances from the median

ii) The maximum and minimum observations are not at equal distances from the box

iii) Median does not lie exactly in the centre of the box

Experiment-III: Here the median is = 4.5, Maximum and minimum values are 2 and 9 giving a range of 9. It is a skewed distribution since

i) The maximum and minimum observations are not at equal distances from the median

ii) The maximum and minimum observations are not at equal distances from the box

iii) Median does not lie exactly in the centre of the box

iv) The values 4, 4.25 and 4.75 are being clumped together.

Experiment-IV: Here the median is = 3.375, Maximum and minimum values are 2 and 9 giving a range of 7. It is a skewed distribution since

i) The maximum and minimum observations are not at equal distances from the median

ii) The maximum and minimum observations are not at equal distances from the box

iii) Median does not lie exactly in the centre of the box

iv) Clumping of low values (as in III case)

EXERCISES

1. The mean marks in statistics of 100 students of a class was 72. The mean marks of boys was 75, while their number was 70. Find the mean marks of girls in the class.

2. The following data give the electricity consumed (K.watt) by 100 families of Hisar

Electricity consumption	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**
No. of users	6	25	36	20	13

Calculate AM, median, mode, S.D., C.V. and find the range with in which middle 50% consumers fall [Hint: Range of middle 50% = Q_3 ï Q_1]

3. The mean and standard deviation of a set of 100 observations were worked out as 40 and 5 respectively by a computer which by mistake took the value 50 in place of 40 for one observation. Find the correct mean and variance.

4. Calculate S.D. and C.V. from the following data

Profits (10^7Rs.)	**Less than 10**	**Less than 20**	**Less than 30**	**Less than 40**	**Less than 50**	**Less than 60**
No. of companies	8	20	40	70	90	100

[Hint: First convert the cumulative frequencies in to simple frequencies]

5. For the following distribution of heights of 80 students in a class, find AM and S.D. by shortcut method and hence find C.V.

Heights (cm)	**150-155**	**155-160**	**160-165**	**165-170**	**170-175**	**175-180**	**180-185**
No. of students	6	9	20	23	15	5	2

6. Draw histogram, frequency polygon, frequency curve, less than and more than ogives and box plot of the following frequency distribution. Locate the mode, median, Q_1 and Q_3 from the appropriate graph.

Marks	40-50	50-60	60-70	70-80	80-90
No. of students	10	20	40	15	5

CHAPTER-II

PROBABILITY THEORY AND PROBABILITY DISTRIBUTIONS

As a subjective approach, the probability of an event is defined as the degree of oneôs belief, conviction and experience concerning the likelihood of occurrence of the event in the situation of uncertainty like:

- What is the chance that there will be rain tomorrow?
- What is the likelihood that a given project will be completed in time?
- Probably she will get above 80% marks in the final examination.

Galileo (1564-1642), an Italian mathematician was the first man to attempt at a quantitative measure of probability while dealing with some problems related to the theory of gambling. However, systematic and scientific foundation of mathematical theory of probability was laid in mid seventeenth century by two French mathematicians B. Pascal (1623-62) and Pierry de Fermat (1601-65). To begin the discussion on probability, we need to define the following terms

Experiment or Trial: An experiment is any activity that generates data. It is identified by the fact that it has several possible outcomes and all these outcomes are known in advance and can not be predicted with certainty. Such an experiment is called a trial or random experiment. For example, in tossing a fair coin, we cannot be certain whether the outcome will be head or tail.

Sample Space: The set of all possible outcomes of an experiment is known as sample space. The possible outcomes are called as sample points or elements of the sample space. It is denoted by S or Ω.

Event: A subset of the sample space is called an event. For example, getting an odd number when die is rolled is an event.

Simple and Compound Event: A simple or elementary event is a single possible outcome of an experiment. For example, in tossing a coin, the event of coming up head is an elementary event. Compound events have two or more elementary events in it. For example, drawing a black ace from a pack of cards is a compound event since it contains two elementary events of black colour and ace.

Exhaustive Events: The total number of possible outcomes of any trial is known as exhaustive events. For example, in throwing a die there are six exhaustive cases since any one of the six faces 1, 2, 3, 4, 5 or 6 may come uppermost while in throwing two dice, the exhaustive cases are $6^2 = 36$ since any of the 6 numbers 1 to 6 on the first die can associate with any of six numbers on the other die.

Mutually Exclusive Events: Events are said to be mutually exclusive if happening of any event precludes the happening of all other, i.e. no two or more events can happen simultaneously in the same trial. For example, in throwing a die all the six faces numbered 1 to 6 are mutually exclusive.

Equally Likely Events: Outcomes of a trial are said to be equally likely if there is no reason to expect one in preference to other. For example, in tossing an unbiased/uniform coin head and tail are equally likely events.

Favourable Events: The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event. For example, in drawing a card from pack of cards the numbers of cases favourable to drawing an ace are 4 while for drawing a spade are 13 and for drawing a red card are 26.

Independent Events: Events are said to be independent if the occurrence or non-occurrence of an event is not affected by the occurrence or non-occurrence of remaining events. For example, if we draw a card from a pack of well shuffled cards and replace it before drawing second, the result of the second draw is independent that of the first draw. But, however, if the first card drawn is not replaced then the result of second draw is dependent on the result of the first draw and the events are called dependent events.

Complementary Events: Two events are said to be complementary if they are mutually exclusive and exhaustive for a given sample space. For example, when a die is thrown then events representing an even number and an odd number are complementary events.

2.1 Definition of Probability: Here we shall discuss different approaches in calculating the probability of an event.

Classical Approach (Mathematical Definition): If a trial results in ón' exhaustive, mutually exclusive and equally likely cases and ómô of them are favourable to the happening of an event óEô then the probability ópô of happening of event óEô is given by

$$p = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of case}} = \frac{m}{n}$$

Since $0 \le m \le n$, so p is non-negative and cannot exceed unity, i.e. $0 \le p \le 1$

If p = 1, then event E is called a certain event and if p = 0, E is called an impossible event. However the classical definition of probability is not applicable if i) the various outcomes of the trial are not equally likely, ii) exhaustive number of cases in a trial is infinite.

Relative Frequency Approach (Statistical/Empirical Definition): If a trial is repeated a number of times essentially under homogenous and identical conditions, then the limiting value of the ratio of number of times the event happens to the total number of trials, as the trials become indefinitely large, is called the probability of happening of that event. So if in ónô trials, an event óEô happens ómô times, then the probability p of the happening of event óEô is given by:

$$p = \lim_{n \to \alpha} \frac{m}{n}$$

Axiomatic or Modern Approach: It was introduced by the Russian mathematician A.N. Kolmogorov (1933) on the basis of set theory. According to this approach, the probability is defined as a set function and is based on certain axioms or postulates given below:

i) To every outcome of the sample space, a real number p (called the probability) can be attached and is such that $o \le p \le 1$.

ii) The probability of the entire sample space (sure event) is 1, i.e. P (S) = 1.

iii) If A and B are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$

2.2 Laws of Probability:

Addition Law of Probability: The probability that atleast one of the two events A and B occurs is equal to the probability that event A occurs, plus the probability that B occurs minus the probability that both events occur. Symbolically, it can be written as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

However, if the events are mutually exclusive, then the probability of occurrence of event A or B is sum of the individual probabilities of A and B.

Symbolically, $P(A \cup B) = P(A) + P(B)$

If A, B, C are three events then addition law may be stated as:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Similarly if A_1, A_2, ·····, A_k are mutually exclusive events then probability of occurrence of either of ðkô events is given by

$$P(A_1 \cup A_2 \cup, \cdots\cdots, \cup A_k) = P(A_1) + P(A_2) + é \ .. + P(A_k)$$

Conditional Probability: It is the probability associated with the events defined on the subset of the sample space. The conditional probability of A given B, is equal to the probability of event A when it is known that another event B has already occurred. Symbolically, we may write it as:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \ ; \ P(B) \neq 0$$

Thus events A and B are independent if and only if P (A/B) = P (A) and P (B/A) = P (B) which implies that P (A ∩ B) = P (A) P(B).

Multiplication Law of Probability: If the two events A and B are independent, then the probability that both will occur together is equal to the product of their individual probabilities. Symbolically,

$$P(A \text{ and } B) = (A \cap B) = P(A)\ P(B)$$

Similarly if A_1, A_2,·····,A_k are independent event then the probability of simultaneous occurrence of ðkô events is equal to the product of probabilities of their individual occurrence.

$$P(A_1 \cap A_2 \cap, \cdots\cdots, A_k) = P(A_1)\ P(A_2)\ é \ .. \ P(A_k)$$

However, if the events are not independent then the probability of simultaneous occurrence of events A and B is given by:

$$P(A \cap B) = P(A)\ P(B/A) = P(B)\ P(A/B)$$

Where P (B/A) represents the conditional probability of occurrence of B when the event A has already happened and P (A/B) is similarly defined.

Joint and Marginal Probability:

Joint probability is the probability of simultaneous occurrence of two or more events. For example with two events A and B, there are four joint probabilities possible as explained below:

Joint event	Sample Points belonging to the Joint event	Probability
$A \cap B$	$n(A \cap B)$	$n(A \cap B)/n(S)$
$A \cap \overline{B}$	$n(A \cap \overline{B})$	$n(A \cap \overline{B})/n(S)$
$\overline{A} \cap B$	$n(\overline{A} \cap B)$	$n(\overline{A} \cap B)/n(S)$
$\overline{A} \cap \overline{B}$	$n(\overline{A} \cap \overline{B})$	$n(\overline{A} \cap \overline{B})/n(S)$

For example, consider the employment status of skilled and unskilled workers in a village, where the event A implies that the worker is employed and B that he is skilled.

Employment Status	Skilled B	Unskilled ($\overline{B}$)	Total
Employed (A)	$n(A \cap B) = 160$	$n(A \cap \overline{B}) = 240$	$n(A) = 400$
Unemployed ($\overline{A}$)	$n(\overline{A} \cap B) = 40$	$n(\overline{A} \cap \overline{B}) = 60$	$n(\overline{A}) = 100$
Total	$n(B) = 200$	$n(\overline{B}) = 300$	$n(S) = 500$

The joint probability table is as follows:

Status	B	$\overline{B}$	Marginal Probability
A	$P(A \cap B) = \frac{160}{500} = 0.32$	$P(A \cap \overline{B}) = \frac{240}{500} = 0.48$	$P(A) = 0.80$
$\overline{A}$	$P(\overline{A} \cap B) = \frac{40}{500} = 0.08$	$P(\overline{A} \cap \overline{B}) = \frac{60}{500} = 0.12$	$P(\overline{A}) = 0.20$
Marginal Probability	$P(B) = 0.40$	$P(\overline{B}) = 0.60$	1.00

The row and column probabilities are called marginal probabilities, simply because these are found in the margins of the table. It may also be noted that a marginal probability is the sum of a set of joint probabilities. For example,

$P(A)$ = Marginal Probability of employed worker

= $P(A \cap B) + P(A \cap \overline{B}) = 0.32 + 0.48 = 0.80$

$P(B)$ = Marginal Probability of skilled worker

= $P(A \cap B) + P(\overline{A} \cap B) = 0.32 + 0.08 = 0.40$

Also in the above table, we find that

$$P(A \cap B) = P(A) \times P(B); \; P(\overline{A} \cap B) = P(\overline{A}) \times P(B)$$

$$P(A \cap \overline{B}) = P(A) \times P(\overline{B}); \; P(\overline{A} \cap \overline{B}) = P(\overline{A}) \times P(\overline{B})$$

Thus the joint probabilities are the product of the relevant marginal probabilities which shows that events A and B are independent. In other words, the employment prospects are equal for skilled and unskilled workers.

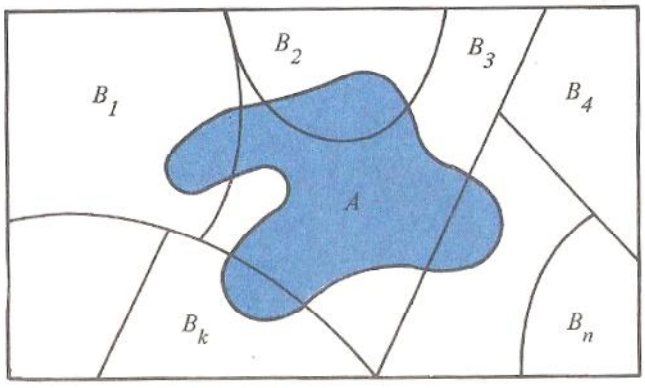
2.3 Bayes' Theorem:

It was introduced by a British Mathematician Thomas Bayes (1702-61). It is useful in the revision of the old (given) probabilities in the light of additional information supplied by the experiment or past records. It may be of extreme help to policy makers in arriving at better decisions in the face of uncertainty. It is given below:

If $B_1, B_2, \cdots, B_n$ be n mutually exclusive events whose union is the sample space 'S' and if $P(B_i) \neq 0$, $(i = 1, 2, \cdots, n)$ then for any arbitrary event A of the sample space 'S'

$$P(B_i/A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i)\,P(A/B_i)}{\sum_{i=1}^{n} P(B_i)P(A/B_i)}$$

The probabilities $P(B_1)$, $P(B_2)$, …, $P(B_n)$ are termed as 'prior probabilities because they exist before we gain any information from the experiment. $P(A/B_i)$ $(i=1, 2, \cdots, n)$ are called 'likelihoods' because they indicate how likely the event 'A' occur given each and every prior probabilities. $P(B_i/A)$ are called posterior probabilities or revised probabilities.



Graphical Representation of Bayes' Theorem

Example-1: A box contains a few red and a few blue balls. One ball is drawn randomly, find the probability of getting a red ball if:

i) We know that there are 30 red and 40 blue balls in the box.

ii) We do not know the composition of the box.

Solution: Let A be the event of getting a red ball.

When composition of the box is known, we use classical approach to find the required probability and when true composition of the box is unknown, we use the Empirical or relative frequency approach.

i) Composition of the box is known (a priori information)

$$P(A) = \frac{\text{Number of favourable cases to A}}{\text{Exhaustive No. of cases}}$$

$$= \frac{\text{Number of red balls}}{\text{Total nunber of balls}} = \frac{30}{70} = \frac{3}{7}$$

ii) The experiment of drawing one ball is replicated a large number of times. The relative frequency of the red ball is found, then relative frequency of red ball will be approximately equal to P(A).

If we make n draws with replacement then,

$$P_n(A) = \frac{\text{Number of red balls appeared in n draws}}{n} \approx P(A)$$

Example-2: What is the chance that a leap year selected at random will contain 53 Sundays?

Solution: A leap year consists of 52 complete weeks with a balance of two days. These two days can be any one of the seven possible days:

(i) M & T (ii) T & W (iii) W & Th (iv) Th & F

(v) F & Sat (vi) Sat & S (vii) S & M

Out of the seven combinations, only last two are favourable.

$\therefore$ Required chance = 2/7

Example-3: A basket of fruit contains 3 mangoes, 2 bananas and 7 pineapples. If a fruit is chosen at random, what is the probability that it is either a mango or a banana.

Solution: Let A_1, A_2 and A_3 respectively denote the events that a chosen fruit is a mango, a banana or a pineapple.

Given

Number of mangoes (m_1) = 3

Number of bananas (m_2) = 2

Number of pineapples (m_3) = 7

$\therefore$ Total number (N) = $m_1 + m_2 + m_3 = 3 + 2 + 7 = 12$

$\therefore$ $P(A_1) = \frac{m_1}{N} = \frac{3}{12}$, $P(A_2) = \frac{m_2}{N} = \frac{2}{12}$ and $P(A_3) = \frac{m_3}{N} = \frac{7}{12}$

Since the events A_1, A_2 and A_3 are mutually exclusive

P(the chosen fruit is either a mango or a banana) = P(A_1 or A_2)

$= P(A_1) + P(A_2) = \frac{3}{12} + \frac{2}{12} = \frac{5}{12}$

Example-4: In a single throw of two dice, determine the prob. of getting

(a) a total of 2, (b) a total of 12, (c) a total of 7 or 9

Solution: Two dice can be thrown in 6x 6 = 36 ways

a) A total of 2 can be obtained as (1, 1) = 1 way

$\therefore$ P (a total of 2) = 1/36

b) A total of 12 can be obtained as (6, 6) = 1 way

$\therefore$ P (a total of 12) = 1/36

c) A total of 7 can be obtained as (6, 1), (1, 6), (5, 2), (2, 5), (4, 3) (3, 4).

A total of 9 can be obtained as (6, 3), (3, 6), (5, 4), (4, 5)

$\therefore$ a total of 7 or 9 can be obtained in 10 ways.

$\therefore$ required Probability = 10/36 = 5/18.

Example-5: A card is drawn from a well-shuffled pack of 52 cards. Find the probability of drawing

(a) a black king

(b) a jack, queen, king or ace

(c) a card which is neither a heart nor a king

(d) a spade or a club.

Solution: A card can be drawn from a pack of 52 cards in ${}^{52}C_1$ ways.

(a) P (a black king) = $\frac{{}^{2}C_1}{{}^{52}C_1} = \frac{2}{52} = \frac{1}{26}$

(b) P (a jack, queen, king or ace) $\frac{^{16}C_1}{^{52}C_1} = \frac{16}{52} = \frac{4}{13}$

(c) P (a card which is neither a heart nor a king) = $\frac{^{36}C_1}{^{52}C_1} = \frac{36}{52} = \frac{9}{13}$

(d) P (a spade or a club) = $\frac{^{26}C_1}{^{52}C_1} = \frac{26}{52} = \frac{1}{2}$

Example-6: The probability that a boy will get admission in B.Sc. (Ag.) is 2/3 and the probability that he will not get admission in M.V.Sc. is 5/9. If the probability of getting admission in at least one course is 4/5, what is the probability that he will get admission in both courses?

Let event A: Boy will get admission in B.Sc. (Ag.)

and B : Boy will get admission in M.V.Sc.

The P(A)= 2/3, P(B)= 1-(5/9) = 4/9, P(A∪B) = 4/5

Thus by application of additive law, we get:

P(A∩B) = P(A) + P(B) ï P(A∪B) = 2/3 + 4/9-4/5 = 14/45

Example-7: A study showed that 65% of managers had a MBA degree and 50% had a B Tech degree. Furthermore, 20% of the managers had a MBA degree but no B Tech degree. What is the probability that a manager selected at random has a MBA degree, given that he has a B Tech degree?

Solution:

Let event A : Manager has a MBA degree

B : Manager has a B Tech degree

Then P(A) = 0.65, P(B) = 0.5 and $P(A \cap \overline{B}) = 0.20$.

We know that $P(A) = P(A \cap B) + P(A \cap \overline{B}) \Rightarrow P(A \cap B) = P(A) - P(A \cap \overline{B})$

hence P(A∩B) = 0.65 ï 0.20 = 0.45

Thus P(A/B) = $\frac{P(A \cap B)}{P(B)} = \frac{0.45}{0.5} = 0.9$

Example-8: In a locality, out of 5,000 people residing, 1200 are above 30 years of age and 3000 are females. Out of 1200 (who are above 30), 200 are females. Suppose, after a

person is chosen you are told that the person is female. What is the probability that the she is above 30 years of age?

Solution: Let event A denote the person of age above 30 years and event B that the person is a female. Then,

$$\text{Thus} \quad P(B) = \frac{3000}{5000} = 0.6, \quad P(A \cap B) = \frac{200}{5000} = 0.04$$

$$\text{Hence } P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.04}{0.60} = 0.067$$

Example-9: Given two independent events, A, B such that P(A) = 0.3 and P(B) = 0.6. Determine

(a) P (A and B), (b) (A and not B), (c) P (not A and B),

(d) (neither A nor B) (e) P (A or B)

Solution: (a) $P(A \text{ and } B) = P(A \cap B) = P(A).P(B) = 0.3 \times 0.6 = 0.18$

(b) $P(A \text{ and not } B) = P(A \cap \overline{B}) = P(A) . P(\overline{B}) = 0.3 \times (1 - 0.6)$

$= 0.3 \times 0.4 = 0.12$

(c) $P(\text{not } A \text{ and } B) = P(\overline{A} \cap B) = P(\overline{A}).P(B) = (1-0.3) \times 0.6$

$= 0.7 \times 0.6 = 0.42$

(d) $P(\text{neither A nor B}) = P(\overline{A} \cap \overline{B}) = P(\overline{A}). P(\overline{B}) = (1 - 0.3)(1 - 0.6)$

$= 0.7 \times 0.4 = 0.28$

(e) $P(A \text{ or } B) = 1 - P(\text{neither A nor B})$

$= 1 - 0.28 = 0.72$

Example-10: Find the chance of drawing 2 white balls in succession from a bag containing 5 red and 7 white balls, the balls drawn not being replaced.

Solution: Let óAô be the event that ball drawn is white in first draw.

and óBô be the event that ball drawn is white in 2nd draw.

$$P(A \cap B) = P(A). P(B/A)$$

Here $P(A) = 7/12$ and $P(B/A) = 6/11$

$$\therefore \quad P(A \cap B) = \frac{7}{12} \times \frac{6}{11} = \frac{7}{22}$$

Example-11: The probability of n independent events are p_1, p_2,·····,p_n. Find an expression for the probability that at least one of the events will happen.

Solution: Prob. of non-happening of 1st event = 1 ï p_1

Prob. of non-happening of 2nd event = 1 ï p_2

é é é é é é é

Prob. of non-happening of nth events = 1 - p_n.

Since the events are independent

∴ Prob. of non-happening of all the events

$$= (1-p_1)\,(1-p_2)\,é \;\ldots\; (1-p_n)$$

∴ Prob. of happening of at least one of the events.

$$= 1 ï (1- p_1)\,(1-p_2)\,é \;é\; (1-p_n)$$

Example-12: A problem in statistics is given independently to three students A, B and C whose chances of solving it are $\frac{1}{2}, \frac{1}{3}$ and $\frac{1}{4}$ respectively. What is the probability that problem will be solved?

Solution: Probability that A solves the problem i.e. $P(A) = \frac{1}{2}$

Probability that B solves the problem i.e. $P(B) = \frac{1}{3}$

Probability that C solves the problem i.e. $P(C) = \frac{1}{4}$

By addition law of probability:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Since the events are independent

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A) \text{ x } P(B) - P(A) \text{ x } P(C) - P(B) \text{ x } P(C) + P(A) \text{ x } P(B) \text{ x } P(C)$$

$$= \frac{1}{2} + \frac{1}{3} + \frac{1}{4} - \frac{1}{2} \text{x} \frac{1}{3} - \frac{1}{2} \text{x} \frac{1}{4} - \frac{1}{3} \text{x} \frac{1}{4} + \frac{1}{2} \text{x} \frac{1}{3} \text{x} \frac{1}{4}$$

$$= \frac{1}{2} + \frac{1}{3} + \frac{1}{4} - \frac{1}{6} - \frac{1}{8} - \frac{1}{12} + \frac{1}{24} = \frac{3}{4}$$

Alternative Method:

Sometimes, it is more convenient to find the probability of the complement of an event rather than of event itself and this approach will be used here in this example.

Consequently $P(\vec{A}) = \frac{1}{2}, P(\vec{B}) = \frac{2}{3}$ and $P(\vec{C}) = \frac{3}{4}$

Now probability that problem is solved = 1 ï Probability that problem is not solved by any one.

$= 1 - P(\overline{A} \cap \overline{B} \cap \overline{C})$ {Since events are independent}

$= 1 - P(\overline{A}) . P(\overline{B}) . P(\overline{C})$

$= 1 - \left[\frac{1}{2} \text{ x } \frac{2}{3} \text{ x } \frac{3}{4}\right] = \frac{3}{4}$ Ans.

Example-13: óAô speaks truth in 60% cases and óBô in 70% cases. In what percentage of cases are they likely to contradict each other in stating the same fact?

Solution: It is clear that they will contradict each other only if one of them speaks the truth and other tells a lie.

The probability That A speaks the truth and B tells a lie $= \frac{60}{100} \text{ x } \frac{30}{100} = \frac{9}{50}$

The probability that B speaks the truth and A tells a lie

$$= \frac{70}{100} \times \frac{40}{100} = \frac{14}{50}.$$

$\therefore$ The probability that they contradict each other = $\frac{9}{50} + \frac{14}{50} = \frac{23}{50}$

$\therefore$ The % of cases in which they contradict each other

$$= \frac{23}{50} \times 100 = 46\%$$

Example-14: The probability that a man will be alive in 25 years is $\frac{3}{5}$, and the probability that his wife will be alive in 25 years is $\frac{2}{3}$. Find the probability that

a) both will be alive
b) only the man will be alive
c) only the wife will be alive
d) at least one will be alive
e) exactly one will be alive

Solution: Define the events H and W as Husband will be alive in 25 years and Wife will be alive in 25 years respectively.

Given P (H) = $\frac{3}{5}$ and P (W) = $\frac{2}{3}$

a) P(both will be alive) = P (H$\cap$W) = P(H) P(W) (since H and W are independent)

$$= \frac{3}{5} \text{ x } \frac{2}{3} = \frac{2}{5}$$

b) P(only the man will be alive) = P(H) P($\overline{W}$) = P(H) [1- P(W)] = $\frac{3}{5}$ [1- $\frac{2}{3}$] = $\frac{1}{5}$

c) P(only the wife will be alive) = P($\overline{H}$) P(W) = [1- P(H)]P(W)

$$= [1-\frac{3}{5}] \text{ x } \frac{2}{3} = \frac{4}{15}$$

d) P(at least one will be alive) = P(H$\cup$W) = P(H) + P(W) - P(H$\cap$W)

$$= \frac{3}{5}+\frac{2}{3}-\frac{2}{5}=\frac{9+10-6}{15}=\frac{13}{15}$$

e) Probability that exactly one will be alive i.e. either of the two (not both) =

$$P(H \cap \overline{W}) + P(\overline{H} \cap W) = P(H).P(\overline{W}) + P(\overline{H}).P(W)$$

$$= \frac{3}{5}\text{x}\left(1-\frac{2}{3}\right)+\left(1-\frac{3}{5}\right)\text{x}\frac{2}{3}=\frac{3}{5}\text{x}\frac{1}{3}+\frac{2}{5}\text{x}\frac{2}{3}=\frac{7}{15}$$

Example-15: In a bolt factory machines A, B, and C manufacture respectively 25 per cent, 35 per cent and 40 per cent of the total output. Of the total of their output 5, 4, and 2 per cent are defective bolts, A bolt is drawn at random and is found to be defective. What is the probability that it was manufactured by machines A, B, or C?

Solution: Let, A_i (i = 1, 2, 3) be the event of drawing a bolt produced by machine A, B and C, respectively. It is given that

$$P(A_1) = 0.25;\ P(A_2) = 0.35, \text{ and } P(A_3) = 0.40$$

Let B = the event of drawing a defective bolt

Thus $P(B/A_1) = 0.05$, $P(B/A_2) = 0.04$ and $P(B/A_3) = 0.02$

By Bayesô Theorem, we get:

$$P(A_1/B) = \frac{P(A_1)P(B/A_1)}{\sum_{i=1}^{3} P(A_i)P(B/A_i)}$$

$$= \frac{(0.25)(0.05)}{(0.25)(0.05)+(0.35)(0.04)+(0.40)(0.02)} = 0.362$$

Similarly we obtain P (A_2/B) = 0.406 and P (A_3/B) = 0.232

Second Method:

Table of posterior probabilities:

Event	**Prior Probability $P(A_i)$**	**Conditional Probability $P(B/A_i)$**	**Joint Probability (2)x(3)**	**Posterior Probability $P(A_i/B)$**
(1)	**(2)**	**(3)**	**(4)**	**(5)**
A_1	0.25	0.05	0.0125	0.0125÷0.0345=0.362
A_2	0.35	0.04	0.0140	0.014÷0.0345=0.406
A_3	0.40	0.02	0.0080	0.008÷0.0345=0.232
Total	**1.00**		**0.0345**	**1.000**

The above table shows the probability that the given item was defective and produced by machine A is 0.362, by machine B is 0.406, and machine C is 0.232.

Example-16: Three urns are given containing white (W), black (B) and red (R) balls as Urn-I: 2W, 3B and 4R; Urn-II: 3W, 1B and 2R balls; Urn-III: 4W, 2B and 5R balls. An urn is selected at random and two balls are drawn from the selected urn and they happen to be one white and one red ball. What is the probability that the balls are drawn from the Urn-II?

Solution: Let A_1, A_2 and A_3 be the events that Urn-I, II, III are selected, respectively. Let B be the event that out of two selected balls, one is white and other is red. Thus probabilities of selecting a urn are:

$$P(A_1) = 1/3;\ P(A_2) = 1/3, \text{ and } P(A_3) = 1/3$$

Probability of selecting one white and one red ball from urn=I is:

$$P(B/A_1)\frac{{}^{2}C_1 \text{ x } {}^{4}C_1}{{}^{9}C_2} = \frac{2 \text{ x } 4 \text{ x } 2}{9 \text{ x } 8} = \frac{2}{9}$$

Similarly, $P(B/A_2)\frac{^3C_1 \text{ x } ^2C_1}{^6C_2} = \frac{3 \text{ x } 2 \text{ x } 2}{6 \text{ x } 5} = \frac{2}{5}$

And from Urn-III: $P(B/A_3)\frac{^4C_1 \text{ x } ^5C_1}{^{11}C_2} = \frac{4 \text{ x } 5 \text{ x } 2}{11 \text{ x } 10} = \frac{4}{11}$

Now the Bayesô probability that balls belong to Urn-II is:

$$P(A_2/B) = \frac{P(A_2)P(B/A_2)}{\sum P(A_i)P(B/A_i)} = \frac{1/3 \text{ x } 2/5}{\frac{1}{3}\text{x}\frac{2}{9} + \frac{1}{3}\text{x}\frac{2}{5} + \frac{1}{3}\text{x}\frac{4}{11}} = \frac{99}{244}$$

2.4 Random Variable and Probability Distribution:

Random Variable: Random variable is a rule which associates uniquely a real number to every elementary event $e_i \in S$, i = 1, 2,·····,n. In other words a R.V. is a real valued function defined over the sample space of the experiment i.e. a variable whose values are determined by the outcomes of the experiment.

If the R.V. takes only integer values such as 0, 1, 2, 3é .. then it is called discrete R.V. For example, let X be the number of heads obtained in two independent tosses of a coin. Here S = [HH, HT, TH, TT] and X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0. Thus X can take values 0, 1, 2, and is a discrete random variable.

Other examples of discrete random variables are: The number of printing mistakes in each page of a book, the number of telephone calls received by the telephone operator.

If the random variable takes on all values with in a certain interval, then it is called a continuous random variable. The amount of rainfall on a rainy day, height and weight of individuals are examples of continuous random variables.

Probability Distribution: The rule which assigns probabilities to each of the possible values of a random variable is called a probability distribution.

Definition: The distribution of a discrete random variable is called discrete probability distribution.

Let X be a discrete random variable taking values $p_1, p_2,\cdots\cdots,p_n$ respectively such that $\sum p_i = 1$. Then such a listing of outcomes of X together with their associated probabilities is called discrete probability distribution and can be represented in tabular form as

X	x_1	x_2	x_3	é ..	x_n	p_i Ó0 and $\sum p_i = 1$
p(x)	p_1	p_2	p_3	é ..	p_n	

Graphically it can be represented by a set of vertical lines whose heights are proportional to the probabilities $p_i = P(X = x_i)$. The function p(x) is called the probability function or probability mass function (p.m.f.) of the discrete R.V.

Definition: The distribution of a continuous R.V. is called continuous probability distribution.

It is represented by the curve with equation y = f(x). The probability, that X lies between infinitesimal interval (x, x + dx) is given by f(x)dx where f(x) is called the probability density function (p.d.f.) of the continuous R.V. X with the following properties:

i) f(x) is non-negative

ii) Area under the curve y = f(x) and x-axis is unity i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$

iii) $P(a < x < b) = \int_{a}^{b} f(x)dx$ = F(b) ï F(a)

Thus the probability that X lies between any two points a and b is given by the area under the curve y = f(x) from x = a to x = b. The function F(x) = P(X Òx) is called the cumulative distribution function (c.d.f.) of X.

Note: The area under the curve y = f(x) when a = b is zero and hence P (x = a) = 0. Thus for a continuous R.V.X

P (a Òx Òb) = P (a < x Òb) = P (a Òx < b) = P (a < x < b)

2.5 Expectation, Variance and Standard Deviation (SD):

If the random variable X assumes the discrete values x_1, x_2, x_3,·····,x_n with corresponding probabilities p_1, p_2, p_3,·····,p_n then we define:

$$\mu = E(X) = p_1x_1 + p_2x_2 + p_3x_3 + \text{é .} + p_nx_n = \sum_{i=1}^{n} p_i x_i$$

$$\text{Variance} = \mu^2 = E[x - \varepsilon]^2 = \sum_{i=1}^{n} p_i (x_i - \varepsilon)^2 = \sum_{i=1}^{n} p_i x_i^2 - \varepsilon^2$$

$$SD = \mu = \sqrt{Var(X)}$$

In the case of continuous R.V. X with probability density function f(x), we have

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Variance

$$\mu^2 = E[x - \varepsilon]^2 = \int_{-\infty}^{\infty} (x - \varepsilon)^2 f(x)dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - \varepsilon^2$$

$$\mu = \sqrt{Var(X)}$$

Example-17: During the course of a day, a machine turns out either 0, 1 or 2 defective pens with probabilities 1/6, 2/3 and 1/6, respectively. Calculate the mean value and the variance of the defective pens produced by the machine in a day.

Solution: The probability distribution of number of defective pens (X) is given by:

X	0	1	2
p(x):	1/6	2/3	1/6

$\therefore$ $E(X) = \sum_{i=0}^{2} p_i x_i = (1/6) \times 0 + (2/3) \times 1 + (1/6) \times 2 = 1$

$Var(X) = \sum_{i=0}^{2} x_i^2 p_i - \varepsilon^2 = (1/6 .0^2 + 2/3 . 1^2 + 1/6 . 2^2) - 1^2 = 1/3$

Example-18: Given the following probability distribution

x:	1	2	3	4	5	6	7
p(x):	2λ	2λ	λ	3λ	λ^2	$2\lambda^2$	$7\lambda^2 + \lambda$

(i) Find λ (ii) Evaluate $P(X \geq 5)$ and $P(X < 4)$

Solution:

i) Since sum of all probabilities is unity, therefore,

$2\lambda + 2\lambda + \lambda + 3\lambda + \lambda^2 + 2\lambda^2 + 7\lambda^2 + \lambda = 1$

or $10\lambda^2 + 9\lambda - 1 = 0$

or $(\lambda + 1)(10\lambda - 1) = 0$

$\therefore$ $\lambda = 1/10$ ($\lambda \neq -1$, since $0 \leq p(x) \leq 1$

ii) $P(X \geq 5) = p(5) + p(6) + p(7)$

$= \lambda^2 + 2\lambda^2 + (7\lambda^2 + \lambda) = 10\lambda^2 + \lambda$

$= 10\,(1/10)^2 + 1/10 = 1/5$

and $P(X < 4) = p(1) + p(2) + p(3)$

$= 2\lambda + 2\lambda + \lambda = 5\lambda$

$= 5 \times (1/10) = 1/2$

Example-19: Find the mean number of heads in three tosses of a coin.

Solution: Let 'X' denote a random variable which is the no. of heads obtained in three tosses of a coin.

Clearly X takes the values 0, 1, 2, 3. Let p and q be the probabilities of turning up head and tail respectively:

$p = ½,\ q = 1 - ½ = ½$

$\therefore$ $P(X=0) = ½ \times ½ \times ½ = 1/8$

$P(X=1) = 3 \times (½ \times ½ \times ½) = 3/8$

$P(X=2) = 3 \times (½ \times ½ \times ½) = 3/8$

$P(X=3) = ½ \times ½ \times ½ = 1/8$

x_i	p_i	$p_i x_i$
0	1/8	0
1	3/8	3/8
2	3/8	6/8
3	1/8	3/8

Mean $= \Sigma p_i x_i = 12/8 = 3/2 \quad \therefore \mu = 3/2$

Properties of Expectation:

i) For a constant 'a' $E(a) = a$ and $E(aX) = aE(X)$

ii) **Addition Theorem of Expectation:** If $X_1, X_2, \cdots, X_n$ are n RV's then

$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$

i.e. the expectation of sum of several RV's is equal to the sum of their individual expectations.

iii) **Multiplication Theorem of Expectation:** If $X_1, X_2, \cdots, X_n$ are n independent RVôs then

$E(X_1X_2 \text{é é .. } X_n) = E(X_1).\ E(X_2)\ .\text{é é .}\ E(X_n)$

i.e. the expectation of product of several independent R.Vôs is equal to the product of their individual expectations.

(iv) If $X_1, X_2, \cdots, X_n$ are n variables and $a_1, a_2, \cdots, a_n$ are n constants. Then the expectation of linear combination of $X_i$ôs given by

$E(a_1X_1 + a_2X_2 + \text{é ..} + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \text{é ..} + a_nE(X_n)$

Properties of Variance:

i) If óaô is a constant then $V(a) = 0$ and $V(aX) = a^2V(X)$

ii) If X_1, X_2 are two RVôs and a_1, a_2 are constants then

$V(a_1X_1 + a_2X_2) = a_1^2\ V(X_1) + a_2^2\ V(X_2) + 2a_1a_2Cov(X_1, X_2)$

iii) If X_1, X_2 are independent then since $Cov(X_1, X_2) = 0$, so that

$V(a_1X_1 + a_2X_2) = a_1^2\ V(X_1) + a_2^2\ V(X_2)$

When $a_1 = a_2 = 1$, we get $V(X_1 + X_2) = V(X_1) + V(X_2)$ which can further be generalized i.e. if $X_1, X_2, \cdots, X_n$ are n independent random variables, then

$V(X_1 + X_2, \text{..é } + X_n) = V(X_1) + V(X_2) + \text{.é .} + V(X_n)$

Hence variance of the sum of several independent RVôs is equal to the sum of their individual variances.

iv) If $X_1, X_2, \cdots, X_n$ are n independent RVôs then variance of linear combination of several independent variable is

$V(a_1X_1 + a_2X_2 + \text{é ..}\ a_kX_n) = a_1^2\ V(X_1) + \text{é ..} + a_n^2\ V(X_n)$

Co-variance in Terms of Expectation: Simultaneous variation in the two variables is called co-variance. It tells us how the two variables move together. If X_1 and X_2 are two RVôs with $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$ then covariance of X_1 and X_2 is given by

$Cov\ (X_1, X_2) = E[X_1 \text{ï } E(X_1)][X_2 \text{ï } E(X_2)]$

$= E(X_1 \text{ï } \mu_1)\ (X_2 \text{ï } \mu_2)$

If $p_i = 1/n$; i = 1, 2..é n then

$Cov\ (X_1, X_2) = \frac{1}{n}\sum X_1X_2 - \mu_1\ \mu_2$

Covariance of X_1 and X_2 denotes how the two variables vary together. Further if X_1 and X_2 are independent then

$$\text{Cov}(X_1, X_2) = E[X_1 - E(X_1)].[X_2 - E(X_2)]$$

$$= E[X_1 - E(X_1)].\, E[X_2 - E(X_2)]$$

$$= [E(X_1) - E(X_1)].\, [E(X_2) - E(X_2)] = 0$$

Hence if the two RV's are independent then their covariance is zero but the converse is not true.

Example-20: If X, Y and Z are three independent variables with EX = 3, EY = -4 and EZ = 6 $\sigma_x^2 = 4$, $\sigma_y^2 = 9$ and $\sigma_z^2 = 1$. Find the mean and variance of L = 2X + 3Y - Z + 8

Solution: L = 2X + 3Y - Z +8

Mean of L = E(L) = 2E(X) + 3E(Y) - E(Z) + E(8)

$= 2 \times 3 + 3 \times (-4) - 6 + 8$

$= 6 - 12 - 6 + 8 = -4$

$V(L) = 2^2 V(X) + 3^2 V(Y) + (-1)^2 V(Z) + V(8)$

$= 4 \times 4 + 9 \times 9 + 1 \times 1 + 0$

$= 16 + 81 + 1 + 0 = 98$

Example-21: If X_1, X_2 and X_3 are dependent random variables and having means 4, 8, 6 and variances 5, 8, 6 and Cov $(X_1, X_2) = 2$, Cov $(X_1, X_3) = 3$ and COV $(X_2, X_3) = 6$. Find the mean and variance of $Y = 2X_1 - 3X_2 + 3X_3$

Solution: Mean of Y = E(Y) = $E(2X_1 - 3X_2 + 3X_3)$

$E(Y) = 2E(X_1) + (-3) E(X_2) + 3E(X_3)$

$= 2 \times 4 - 3 \times 8 + 3 \times 6$

$= 8 - 24 + 18 = 2$

If $L = a_1X_1 + a_2X_2 + a_3X_3$, then

$V(L) = a_1^2 V(X_1) + a_2^2 V(X_2) + a_3^2 V(X_3) + 2\, a_1a_2 \text{Cov}(X_1, X_2) +$

$2\, a_2a_3 \text{Cov}(X_2, X_3) + 2\, a_1a_3 \text{Cov}(X_1, X_3)$

$V(Y) = 2^2 \times 5 + (-3)^2 \times 8 + (3)^2 \times 6 + 2 \times 2 \times (-3) \times 2 + 2 (-3) \times 3 \times 6 + 2 \times 2 \times 3 \times 3$

$= 20 + 72 + 54 - 24 - 108 + 36 = 50$

EXERCISES

1. An Urn contains 8 red, 7 white and 3 black balls. A ball is drawn at random. What is the probability that the ball drawn will be (i) red (ii) white or black (iii) red or black?

2. One candidate applies for a job in two firms A and B. the probability of his being selected in firm A is 0.7 and being rejected in B is 0.5. The probability that at least one of his applications being rejected is 0.6, what is the probability that he will be selected in one of the firms?

3. A market survey conducted in four cities pertained to preference for brand A soap. The responses are shown below:

	Delhi	**Kolkata**	**Chennai**	**Mumbai**
Yes	45	55	60	50
No	35	45	35	45
No option	5	5	5	5

   What is the probability that a consumer selected at random (i) preferred brand A (ii) preferred brand A and was from Chennai (iii) preferred brand A given that he was from Chennai (iv) given that a consumer preferred brand A, what is the probability that that he was from Mumbai.

4. A problem in business statistics is given to five students A, B, C, D and E whose chances of solving these are 1/2, 1/3, 1/4, 1/5 and 1/6 respectively. What is the probability that the problem will be solved.

   [Hint P (problem will be solved) = 1-P($\overline{A}$). P($\overline{B}$) P($\overline{C}$) P($\overline{D}$). P($\overline{E}$)] [Ans. 5/6]

5. A husband and wife appear in an interview for two vacancies on the same post. The probability of husbandôs selection is 1/7 and that of wifeôs selection is 1/5. What is the probability that (a) both will be selected (b) only one will be selected (c) none will be selected [Ans. 1/35, 2/7, 24/35]

6. An insurance company insured 2000 scooter drivers, 4000 car drivers and 6000 truck drivers. The probability of an accident involving a scooter, car and a truck is 1/100, 3/100 and 3/20 respectively. One of the insured persons meets with an accident. What is the probability that he is a (1) scooter driver (2) car driver (3) truck driver? [Ans. 1/52, 3/26, 45/52]

7. Three urns are given, each containing red and black balls as urn 1; 6 red and 4 black balls urn 2; 2 red and 6 black balls urn 3; 1 red and 8 black balls. An urn is chosen at random and a ball is drawn from the urn and found to be red. Find the probability that the ball is drawn from urn 2 or urn 3 [Ans. 65/173]

CHAPTER-III

IMPORTANT THEORETICAL DISTRIBUTIONS

We now discuss some generalized distributions which are very useful and relevant to the policy makers. These distributions are based on certain assumptions. Using these distributions, predictions can be made on theoretical grounds.

3.1 Discrete Probability Distributions:

Discrete Uniform Distribution: The probability distribution in which the random variable assumes all its values with equal probabilities, is known as discrete uniform distribution. If the random variable X assumes the value x_1, $x_{2,}\cdots\cdots,x_k$ with equal probabilities, then its probability function is given by:

$$P(X = x_i) = \frac{1}{k} \text{ for } i = 1,2,.....,k$$

Bernoulli Distribution: A random variable X which takes only two values 1 and 0 (called as success and failure) with respective probabilities p and q such that P(X=1) = p, and P(X = 0) = q and p + q = 1, is called a Bernoulli Variate and its distribution is called Bernoulli Distribution.

Binomial Distribution: This distribution was given by James Bernoulli (1713) and is based on the assumptions of Bernoulli trials.

Bernoulli Trials: A series of independent trials which can result in one of the two mutually exclusive outcomes called success or failure such that the probability of success (or failure) in each trial is constant, then such repeated independent trials are called Bernoulli trials.

If we perform a series of n Bernoulli trials such that for each trial, p is the probability of success and q is the probability of failure (p + q = 1), then probability of x successes in a series of n independent trials is given by

$p(x) = {}^nC_x\, p^x\, q^{n-x}$ where $x = 0, 1, 2,\cdots\cdots,n$

This is known as binomial distribution and the probabilities $p(0) = q^n$; $p(1) = {}^nC_1 pq^{n-1}$; $p(2) = {}^nC_2 p^2 q^{n-2},\cdots\cdots$, $p(n) = q^n$ of 0 success, 1 success, 2 successes $\cdots\cdots$n successes are nothing but the first, the second, the thirdé é the $(n+1)^{th}$ terms in the

binomial expansion $(q + p)^n$. For this reason, the distribution is called binomial distribution.

Definition: A random variable óXô is said to follow Binomial Distribution if it assumes only non-negative values and its probability mass function is given by:

$$P(X = x) = p(x) = {}^nC_x\, p^x\, q^{n-x}; \quad x = 0, 1, 2\cdots\cdots n \text{ and } p + q = 1$$

Assumptions:

i) There are n repeated independent trials.

ii) Each trial results in two mutually exclusive outcomes called success and failure.

iii) The probability of success ópô is constant for each trial.

Properties of Binomial Distribution:

i) $$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} {}^nC_x p^x q^{n-x} = (q+p)^n = 1$$

ii) The constants n and p are parameters of the Binomial distribution and we denote the distribution as X ~ B(n, p)

iii) Mean = np and variance = npq. Since q < 1 being probability, hence for a Binomial distribution mean is always greater than variance.

Coefficient of skewness $(\gamma_1) = \dfrac{q-p}{\sqrt{npq}}$ and Coefficient of Kurtosis $(\gamma_2) = \dfrac{1-6pq}{npq}$

iv) If $X \sim B(n_1, p)$ and $X_2 \sim B(n_2, p)$ are two independent binomial variates with same probability p, then $Y = X_1 + X_2 \sim B(n_1 + n_2, p)$ i.e. sum of two independent Binomial variates with same probability is again a binomial variate. This property is called **Additive Property of Binomial Distribution**.

v) For Binomial Distribution, the recurrent relationship is

$$p(x+1) = \left(\frac{n-x}{x+1} \times \frac{p}{q}\right) p(x); \quad x = 0, 1, 2é \ .. \ n-1$$

Frequency Function of Binomial Distribution: Suppose an experiment with n outcomes is repeated N times, then expected frequency function of binomial distribution is given by

$$f(x) = Np(x) = N.{}^nC_x p^x q^{n-x}$$

and is used for fitting the binomial distribution to the observed data.

Example-1: Find the Binomial distribution whose mean is 9 and standard deviation is 3/2.

Solution: Mean = np = 9, variance = npq = $(SD)^2$ = 9/4

Thus $\frac{npq}{np} = \frac{9/4}{9}$ or q = 1/4

p = 1 ï q = 1 ï 1/4 = 3/4

Now np = 9

Thus n(3/4) = 9 $\Rightarrow$ n = 12

Thus the parameters of binomial distribution are n = 12 and p = 3/4.

Example-2: If 20% animals in a region are sick, then find the probability that out of 6 animals chosen at random the number of sick animals is i) zero ii) One iii) at the most 1 iv) at least 2.

Solution: Given n = 6 and p = 20/100 = 1/5

Let X be the number of sick animals out of 6, then X ~ B (6, 1/5)

i) $P(X = 0) = {}^6C_0\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^6 = \left(\frac{4}{5}\right)^6 = 0.262$

ii) $P(X = 1) = {}^6C_1\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^5 = 6 \times \left(\frac{4}{5}\right)^1\left(\frac{4}{5}\right)^5 = 0.393$

iii) P (at the most 1) = $P(X \le 1)$ = P (X = 0 or 1)

= P (X = 0) + P (X = 1)

= 0.26 + 0.39 = 0.65

iv) $P(X \ge 2)$ = P (at least 2) = 1 ï P (X = 0 or 1) = 1-0.65 = 0.35

Example-3: The probability that a bulb will fail before 100 hours is 0.2. If 15 bulbs are tested for life length, what is the probability that the number of failures before 100 hours (i) does not exceed 2; (ii) are atleast 1.

Solution: Using binomial distribution we have:

n = 15, p = 0.2 and q = 1 ï 0.2 = 0.8

Let X = No. of bulbs failing before 100 hours

Therefore, X ~ B (15, 0.2)

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$= {}^{15}C_0(0.2)^0 (0.8)^{15} + {}^{15}C_1(0.2)^1 (0.8)^{14} + {}^{15}C_2(0.2)^2 (0.8)^{13}$

$= (0.8)^{15} + 15 (0.2) (0.8)^{14} + 105 (0.2)^2 (0.8)^{13} = 0.648$

$P(X \geq 1) = 1 - P(X = 0)$

$= 1 - q^n = 1 - (0.8)^{15} = 1 - 0.035 = 0.965$

Example-4: In a farm, 200 rows each of 6 papaya seedlings are sown. Assuming equal chance of a seedling to be male or female, how many rows do you expect to have 4 or 5 female plants?

Solution: Let X be the number of female plants from 6 seedlings in a row.

Here N = 200, n = 6 and p = 1/2

$\therefore$ $X \sim B(6, 1/2)$

P (4 or 5 female plants in a row) = P (X = 4 or 5)

$= P(X = 4) + P(X = 5) = {}^6C_4 (½)^4 (½)^2 + {}^6C_5 (½)^5 (½)^1$

$= \frac{15}{64} + \frac{6}{64} = \frac{21}{64}$

Thus, the number of rows with 4 or 5 female plants = N x P(X = 4 or 5)

$= 200x\frac{21}{64} \cong 66$

Example-5: Assuming equal probabilities of girls and boys in a family of 4 children find the probability distribution of number of boys and also find the expected number of families out of 2000 having (i) at most two boys (ii) at least one boy.

Solution: Let X be the number of boys in a family of 4 children. Possible values of X can be 0, 1, 2, 3 and 4. The required probabilities can be obtained using binomial distribution.

Given n = 4 and p = 1/2 = q

Thus, $X \sim B(4, ½)$

Probability of having x boys

$P(X = x) = p(x) = {}^nC_X p^X q^{n-X}$ x = 0, 1, 2 é .. n

$P(X=0) = {}^4C_0 (½)^0 (½)^{4-0} = 1/16$

$P(X=1) = {}^4C_1 (½)^1 (½)^{4-1} = 1/4$

$P(X=2) = {}^4C_2 (½)^2 (½)^{4-2} = 3/8$

$P(X=3) = {}^4C_3 (½)^3 (½)^{4-3} = 1/4$

$P(X=4) = {}^4C_4 (½)^4 (½)^{4-4} = 1/16$

The probability distribution of random variable X (number of boys) is given by

x:	0	1	2	3	4
p(x):	1/16	1/4	3/8	1/4	1/16

i) Probability of getting at most 2 boys = $P(X \leq 2) = P(X = 0, 1 \text{ or } 2)$

$= P(X = 0) + P(X = 1) + P(X = 2)$

$= 1/16 + 1/4 + 3/8 = 11/16$

Expected number of families having at most two boys = $N. P(X \leq 2)$

$= 2000 \times 11/16 = 1375$

ii) Probability of getting at least one boy = $P(X \geq 1)$

$= P(X = 1, 2, 3 \text{ or } 4) = 1 - P(X = 0) = 1 - 1/16 = 15/16$

Expected number of families having at least one boy = $N. P(X \geq 1)$

$= 2000 \times 15/16 = 1875$

Note: Since p = q = 1/2, here the binomial distribution is symmetrical and thus the probability distribution of number of boys and number of girls will be same.

Example-6: The screws produced by a certain machine were checked by examining samples of sizes 7. The following table shows the distribution of 128 samples according to the number of defective screws.

No. of Defectives(X):	0	1	2	3	4	5	6	7	**Total**
No. of Samples(f) :	4	6	19	35	30	26	7	1	**128**

Fit a binomial distribution and find the expected frequencies. Also find the mean and variance of the distribution.

Solution: Here $N = \sum f_i = 128, n = 7, \sum_{i=1}^{n} f_i x_i = 448$

$$\overline{X} = \frac{1}{N}\sum f_i x_i = \frac{448}{128} = \frac{7}{2}$$

If the data follows Binomial distribution, then the mean $\overline{X}$ should be equal to the mean of the Binomial distribution np.

$$\overline{X} = np = \frac{7}{2}; \qquad p = \frac{7}{2n} = \frac{7}{14} = \frac{1}{2} \quad \text{and } q = 1 - p = \frac{1}{2}$$

Thus the parameters of the Binomial distributions are n = 7 and p = $\frac{1}{2}$

The probability that 0, 1, 2, 3, 4, 5, 6, 7 screws will be defective is given by various terms in the expansion of $(p + q)^n = (½ + ½)^7$

$$(½ + ½)^7 = {}^7C_0(½)^0(½)^7 + {}^7C_1(½)^1(½)^6 + {}^7C_2(½)^2(½)^5 + {}^7C_3(½)^3(½)^4 + {}^7C_4(½)^4(½)^3 + {}^7C_5(½)^5(½)^2 + {}^7C_6(½)^6(½)^1 + {}^7C_7(½)^7(½)^0$$

$$= (½)^7 [1 + 7 + 21 + 35 + 35 + 21 + 7 + 1]$$

The expected frequencies of defective screws are obtained by multiplying with N

$128 (½ + ½)^7 = 128 (½)^7 [1 + 7 + 21 + 35 + 35 + 21 + 7 + 1]$

Thus expected frequencies are

x :	0	1	2	3	4	5	6	7
f :	1	7	21	35	35	21	7	1

Mean of Binomial distribution = np = 7(½) = 3.5

Variance of Binomial distribution = npq = 7(½) (½) = 1.75

Negative Binomial Distribution:

In the last section, binomial distribution is discussed where mean>variance and may be used to find the probability of x success in n trials (fixed). Here the Negative Binomial Distribution is discussed in which mean<variance and may be used to find that how many trials are required to obtain x successes (fixed).

Suppose n trials are required to obtain exactly x successes ($n \geq x$, i.e. n = x, x + 1, x + 2, é .). It will happen only if the n^{th} trial (last trial) results in x^{th} success and the previous (n-1) trials result in (x-1) successes and (n-1) ï (x-1) = (n-x) failures. It can happen in ${}^{n-1}C_{x-1}$ ways.

Thus the probability that exactly n trials will be needed to obtain x successes is given by ${}^{n-1}C_{x-1}\, p^{x-1}\, q^{n-x} \cdot p = {}^{n-1}C_{x-1}\, p^x q^{n-x}$.

Definition: A discrete random variable N (number of trials needed to obtain x successes) is said to follow negative binomial distribution if its probability function is given by:

$$P(n; x, p) = {}^{n-1}C_{x-1}\, p^x q^{n-x};\ n = x, x+1, \text{é}\ ..$$

where $0 < p < 1$ and $x \geq 1$ is fixed.

Another Form:

Negative binomial distribution may also be defined as the distribution of the number of failures preceding the x^{th} success. If r is the number of failures preceding the x^{th} success, then x+r is the total number of trials required to produce x successes. It will happen only if the last trial i.e. $(x+r)^{th}$ trial results in success and the first (x+r-1) trials result in (x-1) successes for which the probability is:

$P(r, x, p) = {}^{x+r-1}C_{x-1}\, p^x q^r$ n = x, x+1, .é . r = 0, 1, 2, é .. and may be obtained from p (n, x, p) by putting n = x + r.

Properties:

i) It is a discrete distribution and has two parameters x (number of successes) and p (probability of success)

ii) Mean = x q/p and Variance = xq/p^2, thus mean < variance

iii) It is positively skewed and leptokurtic distribution

Remarks:

i) The salient feature of negative binomial distribution is that number of successes is fixed while the number of trials (N) and number of failures is a random variable.

ii) The expected number of trials = x/p whereas expected number of failures = xq/p.

iii) This distribution is also called as waiting time distribution as it arises in problems where the success takes place as a waiting time phenomenon. For example, it arises in problems concerning inverse sampling where the sampling of items is continued till required numbers (fixed number) of items of a particular type are obtained in the sample. For x = 1, this distribution is called **Geometric Distribution**.

iv) This distribution is also called as Pascalôs distribution after the name of French mathematician B. Pascal (1623-62).

Example-7: Find the probability that a man tossing a coin gets the 4^{th} head on the 9^{th} toss.

Solution: No. of trials n = 9

Number of successes x = 4

p = 0.5

Therefore P (4^{th} head on 9^{th} trial) $^{9-1}C_{4-1}\ (0.5)^4\ (0.5)^5$

$= {}^{8}C_3\ (.5)^9$

= 0.1094

Example-8: A recruitment agency conducts interviews by telephone found from the past experience that only 40% calls are being answered. Assuming it a Bernoulli process, find:

i) The probability that the agencyôs 5^{th} answer comes on tenth call

ii) Expected number of calls necessary to obtain eight answers

iii) The probability that the agency receives the first answer on third call

Solution:

i) Consider the answer to a call as a success. Here probability of success (p) = 40/100 = 0.40

x = 5, n = 10

P (5^{th} answer coming on 10^{th} call)

$= {}^{10-1}C_{5-1}\ (0.4)^5\ (1-0.4)^{10-5}$

$= {}^{9}C_4\ (0.4)^5\ (0.6)^5 = (126)\ (0.24)^5 = 0.1003$

ii) Expected number of calls necessary to obtain 8 answers $= \frac{x}{p} = \frac{8}{0.4} = 20$

iii) P (1^{st} answer on the 3^{rd} call) $= {}^{3-1}C_{1-1}\ (0.4)^1\ (0.6)^{3-1}$

$= {}^{2}C_0\ (0.4)\ (0.6)^2$

$= 0.144$

Example-9: Suppose in a region, 20% animals are suffering from tuberculosis. If the animals are medically examined until 4 are found to have TB, then (i) what is the probability that 10 animals are examined; (ii) What are the expected number of animals required to be examined to have 5 positive cases of TB.

Solution:

i) Considering an animal with TB as success

Here p = 20/100 = 0.2

x = 4 n = 10

P (10 animals are examined in order to have 4 positive cases of TB)

$= {}^{10-1}C_{4-1}\,(0.2)^4\,(0.8)^{10-4}$

$= {}^{9}C_{3}\,(0.2)^4\,(0.8)^{6} = 0.0352$

ii) Expected number of animals required to be examined to have 5 positive cases = x/p = 5/0.2 = 25

Poisson Distribution:

It is discrete probability distribution and was developed by a French mathematician S.D. Poisson (1837). It is also a limiting case of binomial distribution. If number of trials (n) is very large and probability of success (p) is very small so that np = λ (say) is finite then the Binomial distribution can be approximated by Poisson distribution i.e. B (n, p) ~ P (ॐ) where ॐ = np

Note: P.D. is good approximation to B.D. when $n \geq 20$ and $p \leq 0.05$.

Definition: A discrete R.V. X is said to follow Poisson distribution if its probability function is given by

$$P(X = x) = p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0, 1, 2, \cdots\cdots, \infty$$

Assumptions:

i) The occurrence or non-occurrence of an event does not affect the occurrence and non-occurrence of any other event.

ii) Probability of occurrence of more than one event in a quite small interval/region is very small.

iii) The probability of success for a short interval or a small region is proportional to the length of the interval or space.

Properties:

i) The Poisson distribution is identified by only one parameter λ and distribution of Poisson variate is denoted as $X \sim P(\lambda)$.

ii) Mean = variance = λ

iii) If $X_1 \sim P(\lambda_1)$ and $X_2 \sim P(\lambda_2)$ are two independent Poisson random variables, then $Y = X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$ and it is often known as called the additive property of Poisson Distribution.

iv) For different values of λ, the values of $e^{-\lambda}$ may be obtained from Fisher and Yates tables, or may be computed by using a calculator. Then various probabilities can be computed using the recurrence relation

$$p(x+1) = \frac{\lambda}{x+1} p(x) \text{ and } p(0) = e^{-\lambda}$$

Importance of Poisson Distribution:

Poisson distribution is used in wide variety of problems concerning rare events with respect to time, area, volume or similar unit. Some of the practical situations where Poisson distribution can be used are given below

i) Number of road accidents per month

ii) Number of printing mistakes per page

iii) Number of defective items produced by a standard process per batch etc.

Fitting of Poisson Distribution: Construct the frequency distribution of data (if not given) and find its mean $\overline{X} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$. If the data follows Poisson distribution then $\lambda = \overline{X}$ and the expected frequencies are given by

$$f(x) = Np(x) = N\frac{e^{-\lambda}\lambda^x}{x!}; \qquad x = 0, 1, 2\text{é } ..$$

and is called the fitting of Poisson distribution to the observed data.

Example-10: Find the distribution of a Poisson random variable X for which 2P (X=1) = P (X=2). Also compute the probability that (i) X = 2 (ii) X <2, given $e^{-4} = 0.0183$

Solution: Probability mass function of a Poisson distribution is

$$P(X = x) = p(x) = \frac{e^{-\lambda}.\lambda^x}{x!} \qquad x = 0, 1, 2\text{ é } ..$$

Also $P(X=1) = \frac{e^{-\lambda}.\lambda^1}{1!} = \lambda.e^{-\lambda}$ and

$$P(X=2) = \frac{e^{-\lambda}.\lambda^2}{2!} = \frac{\lambda^2 e^{-\lambda}}{2}$$

Given $2P(X=1) = P(X=2)$

$$\Rightarrow \quad 2(\lambda e^{-\lambda}) = \frac{\lambda^2 e^{-\lambda}}{2} \text{ or } \lambda = 4$$

Now $P(X = 2) = \frac{e^{-4}4^2}{2!} = \frac{16 \times e^{-4}}{2}$

$= 8 \times 0.0183 = 0.1465$

$P(X < 2) = P[X = 0 \text{ or } 1] = P(X=0) + P(X=1)$

$= \left[\frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!}\right]$

$= [e^{-4} + 4e^{-4}]$

$= 5 \times 0.0183$

$= 0.0915$

Example-11: If 1.5% animals are suffering from a rare disease, find the probability that out of 200 animals in a region (i) at least two animals are suffering (ii) No animal is suffering from rare disease use $[e^{-3} = 0.0498]$.

Solution: Total number of animals (n) = 200

Let X be the number of suffering animals

Since 1.5% animals are suffering from the disease

$\therefore \quad p = \frac{1.5}{100} = 0.015$

Then $X \sim B(200, 0.015)$

Since n is large and p is small, using Poisson approximation to binomial distribution. The mean of the resulting Poisson distribution is

$\lambda = np = 200 \times 0.015 = 3$

$\therefore \quad X \sim P(3)$

Probability function for Poisson distribution is

$f(x) = \frac{e^{-\partial}\partial^x}{x!} \qquad x = 0, 1, 2, \text{é} \ ..$

i) P (at least two animals are suffering) = $P(X \geq 2)$

$= 1 - P(X < 2) = 1 - [P(X=0) + P(X=1)]$

$= 1 - \left[\frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!}\right] = 1 \text{ ï } [e^{-3} + 3e^{-3}]$

$= 1 - 4e^{-3} = 1 - 4 \times 0.0498 = 0.8008$

ii) P (no animal is suffering) = P (X = 0)

$$= \frac{e^{-\lambda}\lambda^0}{0!} = \frac{e^{-3}3^0}{1} = e^{-3} = 0.0498$$

Example-12: On an average, one bulb in 400 bulbs is defective. If the bulbs are packed in boxes of 100, what is the probability that any given box of bulbs will contain (i) no defective, (ii) less than two defective bulbs (iii) one or more defective bulbs.

Solution: Probability of a defective bulb p = 1/400 is small and number of bulbs in a given packet n = 100 are large so using Poisson approximation to binomial distribution, $\lambda = np = 100/400 = 0.25$ is finite.

Thus the required probabilities can be obtained through Poisson distribution having mean 0.25. If X denotes the number of defective bulbs in a batch of 100 bulbs, then X ~ P (0.25).

i) $P(X= 0) = e^{-\lambda} = e^{-0.25} = 0.779$

ii) $P(X < 2) = e^{-\lambda} + \lambda e^{-\lambda} = 0.779\ (1 + 0.25) = 0.973$

iii) $P(X \geq 1) = 1 - P(X=0) = 1 - 0.779 = 0.221$

Example-13: A manufacturer who produces bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug company buys 100 boxes from the producer. Using Poisson distribution, find how many boxes will contain (i) exactly one defective (ii) at least two defectives.

Solution: Let X denotes the number of defective bottles in a pack of 500 bottles.

Here p = 0.001 and n = 500 so that $\lambda = np = 500 \times 0.001 = 0.5$

Therefore, X ~ P (0.5)

i) $P(X = 1) = \lambda e^{-\lambda} = 0.5e^{-0.5} = 0.5(0.6065) = 0.3032$

Thus the number of boxes containing exactly one defective

= 100 x 0.3032 = 30.32 or 30

ii) $P (X \geq 2) = 1 - P (X=0) - P (X=1)$

$= 1 - (e^{-\lambda} + \lambda e^{-\lambda}) = 1 - e^{-\lambda}(1 + \lambda)$

$= 1 - 0.6065\ (1.5) = 1 - 0.9097 = 0.0903$

$\therefore$ Required number of boxes = 100 x 0.0903 = 9.03 i.e. approximately equal to 9

Example-14: The distribution of typing mistakes committed by a typist is given below

No. of mistakes/page (x)	:	0	1	2	3	4	5
No. of pages (f)	:	142	156	69	27	5	1 = Total 400

Fit a Poisson distribution and find the expected frequencies.

Solution: fx	:	0	156	138	81	20	5 Σ fx = 400

$\overline{X} = \lambda = \Sigma fx/N = 400/400 = 1$

using the recurrence relation

$p(x+1) = \frac{\lambda}{x+1} p(x)$ and $p(0) = e^{-\lambda}$

$N\ P(X=0) = 400\ e^{-\lambda} = 400 \times 0.3679 = 147.16$

$N\ P(X=1) = 400\ P(X=0)\ \lambda = 147.16$

$N\ P(X=2) = 400\ P(X=1)\ \lambda/2 = 73.58$

$N\ P(X=3) = 400\ P(X=2)\ \lambda/3 = 24.53$

$N\ P(X=4) = 400\ P(X=3)\ \lambda/4 = 6.13$

$N\ P(X=5) = 400\ P(X=4)\ \lambda/5 = 1.23$

Thus the expected frequencies are:

No. of Mistakes	:	0	1	2	3	4	5
Expected No. of Pages	:	147	147	74	25	6	1

3.2 Continuous Probability Distributions:

Normal Distribution:

While dealing with quantities whose magnitude is of continuous nature such as weight, height etc. a continuous probability distribution is needed. Also when an experiment involves discrete phenomena, the appropriate discrete model may be difficult to use if the number of observations are large. In such cases, it is often convenient to use a continuous model to approximate the discrete model. Normal distribution is one of the most useful theoretical distributions for continuous variables. Most of statistical data concerning biological and agricultural research can be assumed to have a normal distribution.

The normal distribution was first described by Abraham De Moivre (1733) as the limiting form of the Binomial model. Normal distribution was rediscovered by Gauss in

1809 and by Laplace in 1812. The normal model has become the most important probability model in statistical analysis.

Definition: A continuous random variable X is said to have a normal distribution with parameters μ (mean) and σ^2 (variance) if its probability density function is given by:

$$f(x;\varepsilon,\mu) = \frac{1}{\mu\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\varepsilon}{\sigma}\right)^2\right] \qquad -\infty < x < \infty$$

Here $\pi = 3.141$ and $e = 2.718$

If variable X is normally distributed and it has mean 'μ' and variance σ^2 then the variable

$$Z = \frac{X-\varepsilon}{\mu}$$

is called standard normal variate with mean 0 and variance 1, and probability density function of Z is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \qquad -\infty < z < \infty$$

Normal Probability Curve: The probability curve of normally distributed variable 'X' is called normal probability curve.

Properties of Normal Distribution:

i) The curve of normal distribution is bell shaped and is symmetrical about $x = \mu$. It can have different shapes depending upon different values of μ and σ but there is one and only one normal distribution for any given pair of μ and σ.

ii) The maximum height of the normal curve is $\frac{1}{\mu\sqrt{2\pi}}$ and is attained at $x = \mu$.

iii) Mean, median and mode of the normal distribution coincide.

iv) Probability that X lies between two values a and b ($a < b$) is given by the area under normal probability curve between $x = a$ and $x = b$.

v) The two tails of the normal probability curve extend indefinitely and never touches the horizontal axis, which implies a positive probability for all values of random variable ranging from $-\infty$ to $+\infty$.

vi) Areas covered by the various ranges of variable are given by:

$P[\mu - \sigma < X < \mu + \sigma] = 0.6826$

$P[\mu - 2\sigma < X < \mu + 2\sigma] = 0.9544$

$P[\mu - 3\sigma < X < \mu + 3\sigma] = 0.9973$

vii) Linear combination of normal variates is also distributed normally.

Relationship between Poisson, Binomial and Normal distributions

If in a binomial distribution the number of trials 'n' tends to infinity and neither p nor q is small then the binomial distribution can be approximated by the normal distribution i.e. X is approximately normal with mean np and variance npq and probability density function of $Z = \frac{X - np}{\sqrt{npq}}$ tends to $\frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$ which is probability density function of standard normal variate.

Since binomial distribution is discrete while normal distribution is continuous so continuity correction is applied while working probabilities.

i.e. $P\{X = x) = P\left\{x - \frac{1}{2} \le X \le x + \frac{1}{2}\right\}$

and $P\{x_1 \le X \le x_2) = P\left\{x_1 - \frac{1}{2} \le X \le x_2 + \frac{1}{2}\right\}$

Similarly if variable X has Poisson distribution with parameter λ and if λ tends to infinity then the p.d.f. of variable

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \text{ tends to } \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$$

Importance of Normal Distribution:

i) Most of the distributions occurring in practice like binomial, Poisson, hyper-geometric distributions etc. can be approximated by normal distribution. Moreover, many of the sampling distributions like student 't' Snedcor's F, chi-square distribution etc. tend to normality for large samples.

ii) Even if variable is not normally distributed it can be sometimes brought to normal form by simple transformation of variables.

iii) If a variable X is normally distributed then the property that $P[\mu - 3\sigma < X < \mu + 3\sigma] = 0.9973$, i.e. $P[|Z| < 3] = 0.9973$ form the basis of entire large sample theory.

iv) The sampling distributions of important statistics have been derived on the fundamental assumption that the population from which sample is drawn is normal.

v) Normal distribution finds large applications in statistical quality control.

Computation of probabilities

Let $A(z_1)$ be the area between $z = 0$ and $z = z_1$ under the standard normal curve. For different values of z_1, the areas from 0 to z_1 are given in the tables of standard normal distribution. Further since the curve is symmetrical i.e. $A(-z_1) = A(z_1)$ and 50% (=0.5) area lies on either side of $z = 0$, then with the help of these properties, we can find various probabilities. If X is normally distributed with mean μ and variance σ^2, then after converting X into standard normal variate $Z = (X-\mu)/\sigma$, we can find various probabilities as shown in the following figure:

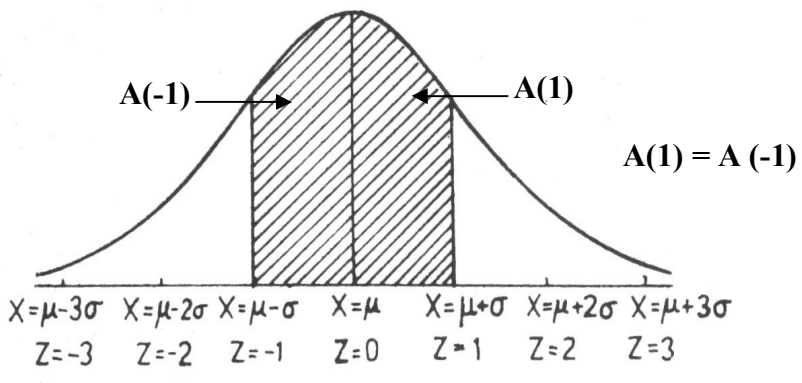
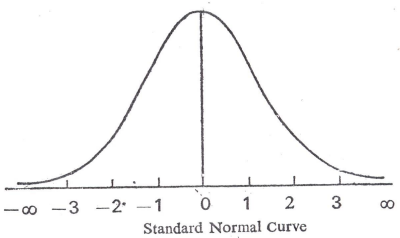


Figure showing area under Normal Curve and Standard Normal Curve

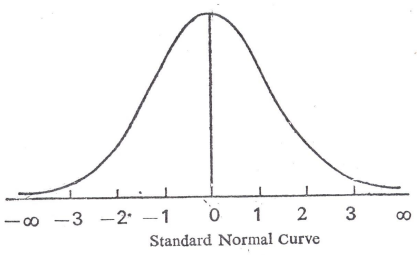
Example-15: Find the probability that standard normal value of Z lies (i) between 0 and 1.8 (ii) from -1.8 to 1.8 (iii) from 1.2 to 2.3 (iv) -2.3 to -1.2 (v) to the right of -1.50 (vi) to the left of -1.82.

Solution: From the shape and tables of area under standard normal curve we find the probabilities

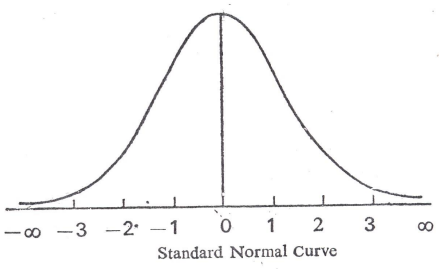
i) $P(0 \le Z \le 1.8) = A(1.8) = 0.4656$



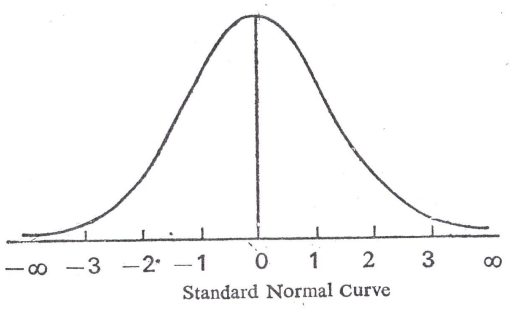
ii) $P(-1.8 \leq Z \leq 1.8) = A(1.8) + A(1.8) = 0.4641 + 0.4641 = 0.9282$



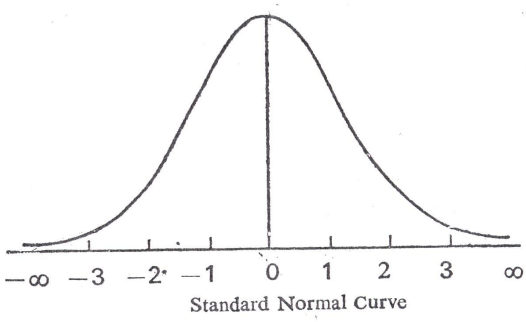
iii) $P(1.2 \leq Z \leq 2.3) = A(2.3) - A(1.2) = 0.4893 - 0.3849 = 0.1044$



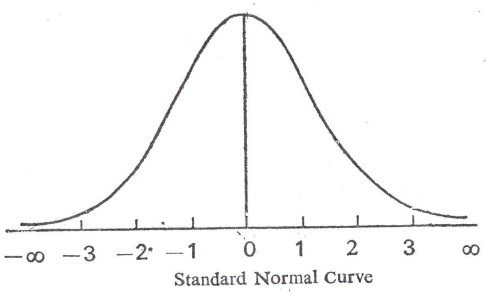
iv) $P(-2.3 \leq Z \leq -1.2) = A(2.3) - A(1.2) = 0.4893 - 0.3849 = 0.1044$



v) $P(Z > -1.50) = 0.5 + A(1.5) = 0.5 + 0.4332 = 0.9332$

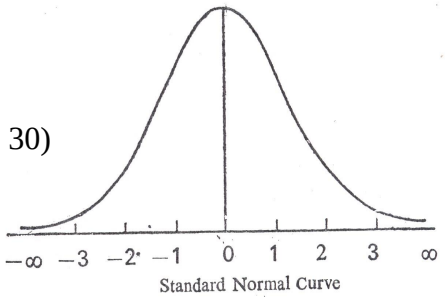


vi) P(Z < -1.82) = 0.5 ï A(1.82) = 0.5 ï 0.4656 = 0.0344



Standard Normal Curve

Example-16: Weight of onion bulb is assumed to be normally distributed with mean 200 gm and standard deviation (σ) = 30 gm. What is the probability that a bulb selected at random have weight between 215 gm and 260 gm?



Standard Normal Curve

Solution: Given μ = 200 gm; σ = 30 gm

Let X be the weight of onion bulbs, then X ~ N (200, 30)

Also $Z = \frac{X-\varepsilon}{\mathring{u}} \sim N(0, 1)$

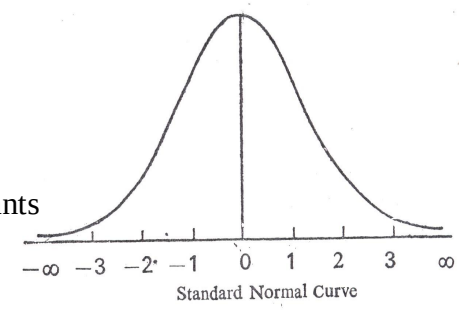
$$z_1 = \frac{x_1-\varepsilon}{\mathring{u}} = \frac{215-200}{30} = \frac{15}{30} = \frac{1}{2} = 0.5$$

and $$z_2 = \frac{x_2-\varepsilon}{\mathring{u}} = \frac{260-200}{30} = \frac{60}{30} = 2$$

P (215 ≤ X ≤ 260) = P[0.5 ≤ Z ≤ 2] = P (0 ≤ Z ≤ 2) ï P (0 ≤ Z ≤ 0.5)

= 0.4773 ï 0.1915 = 0.2858

Example-17: Plants height is assumed to be normally distributed with mean 60 cm and s.d.= 5 cm. If 20% shortest plants are to be selected for crossing, find the maximum height of the selected plants.



Standard Normal Curve

Solution: Given μ = 60 cm and σ = 5 cm

20% shortest plants are selected for crossing

Let k be the maximum height of the selected plants

Then P (X ≤ k) = 0.2

or $$P\left(\frac{X-\varepsilon}{\mathring{u}} \le \frac{k-60}{5}\right) = 0.2$$

We know from table that P (Z ≤ 0.84) = 0.2

$$\Rightarrow \qquad \frac{k-60}{5} = -0.84$$

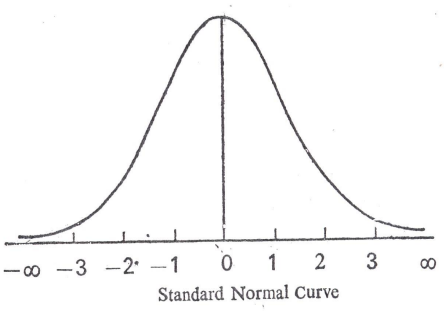
k = 60 ï 5 x 0.84 = 60 - 4.20 = 55.80

Thus maximum height of the selected plant is 55.80 cm.

Example-18: The mean marks for an examination is 72 and the standard deviation is 9. The top 10% of the students are to be awarded Aôs. Find the minimum marks a student must get in order to receive an A.

Solution: Given $\mu = 72$ and $\sigma = 9$

If X denote the marks obtained by a student, then X ~ N (72, 81). Let x_1 denote the minimum marks a student must get for an A grade.



Standard Normal Curve

Then, $P(X \geq x_1) = \frac{10}{100} = 0.1$

or $$P\left(\frac{x-\varepsilon}{\mathring{u}} \geq \frac{x_1 - 72}{9}\right) = 0.1$$

From standard normal table we have $\frac{x_i - 72}{9} = 1.28$

or $x_1 = 9 \text{ x } 1.28 + 72 = 83.52$

Thus the minimum marks for an A grade are $83.52 \cong 84$

Example-19: The lifetimes of certain kinds of electronic devices have a mean of 300 hours and standard deviation of 25 hours. Assuming normal distribution

a) Find the probability that any one of these electronic devices have a lifetime of more than 325 hours.

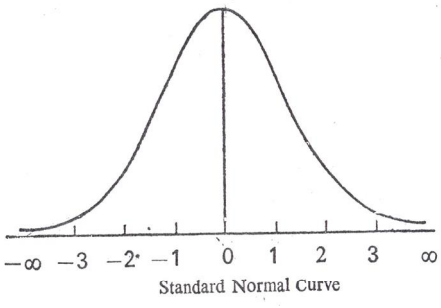
b) What percentage will have lifetimes of 300 hours or less?

c) What percentage will have lifetime between 220 and 260 hours?

Solution: Let the random variable X denotes the life time of an electronic device.

a) Given $\mu = 300$, $\sigma = 25$ and $x = 325$

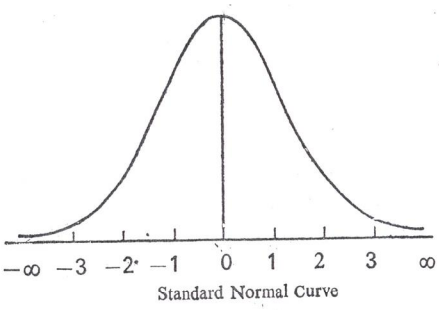
Converting X into standard normal score



Standard Normal Curve

$$z = \frac{x - \varepsilon}{\mathring{u}} = \frac{325 - 300}{25} = 1$$

Thus P(X > 325) = P (Z > 1) = 0.5 ï A(1) = 0.5 ï 0.3413 = 0.1587



Standard Normal Curve

b) $$z = \frac{x - \varepsilon}{\mathring{u}} = \frac{300 - 300}{25} = 0$$

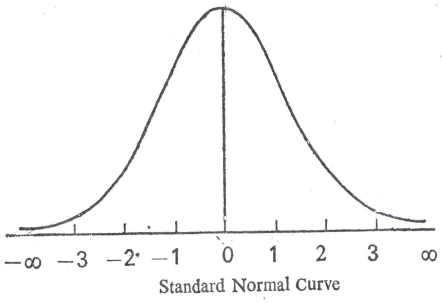
Thus P[X ≤ 300] = P[Z ≤ 0] = 0.5

Therefore, the required percentage is

0.5000 x 100 = 50%

c) $$z_1 = \frac{x_1 - \varepsilon}{\mathring{u}} = \frac{220 - 300}{25} = -3.2$$

and $$z_2 = \frac{x_2 - \varepsilon}{\mathring{u}} = \frac{260 - 300}{25} = -1.6$$



Standard Normal Curve

P(220 ≤ X ≤ 260) = P(-3.2 ≤ Z ≤ -1.6)

= A(3.2) ï A (1.6)

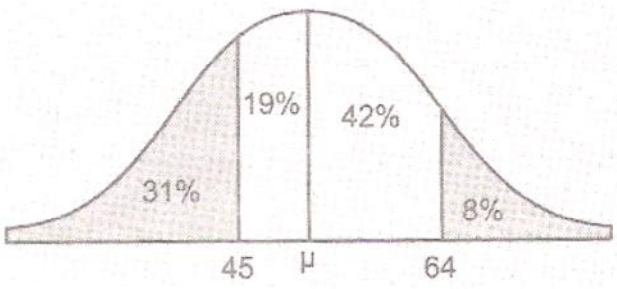
= 0.4993 ï 0.4552 = 0.0541

Therefore, the required percentage is

0.0541 x 100 = 5.41%

Example-20: In a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and SD of the distribution.

Solution: Let μ and σ be the mean and SD. Since 31% of the items are under 45, therefore, the area to the left of ordinate at x = 45 is 0.31. Thus the area to the right of ordinate upto mean is (0.5 ï 0.31) = 0.19. The value of Z corresponding to this area is -0.5.



Hence z = (45 ï μ)/σ = -0.5 or μ - 0.5σ = 45

As 8% of items are above 64, therefore, the area from x = 64 upto mean is (0.5 ï 0.08) = 0.42 and the value of Z corresponding to this area is 1.4.

Hence z = (64 ï μ)/σ = 1.4 or μ + 1.4σ = 64

From the above two equations we get 1.9 σ = 19 i.e. σ = 10. Putting σ = 10 in first equation, we get μ = 50

The mean of distribution = 50 and SD = 10

Example-21: A machine produces bolts which are 10% defective. Find the probability that a random sample of 400 bolts produced by the machine the number of defective bolts found (i) will be at most 30 (ii) will be between 30 and 50 (iii) will exceed 55.

Solution: Since the probability of a defective bolt is 0.1 which is considered as large and number of bolts 400 are large, therefore, it is more appropriate to use normal distribution as a limiting case of binomial distribution.

Thus μ (mean) = np = 400 x 0.1 = 40

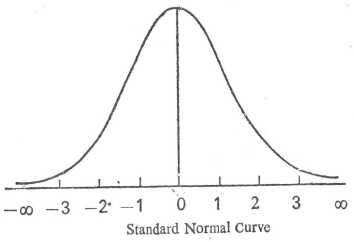
$\mu = \sqrt{npq} = \sqrt{400 \times 0.1 \times 0.9} = 6$

Let X = No. of defective bolts

i) $P(X \le 30) = P(X \le 30.5)$ (Continuity correction)

$= P(Z \le (30.5 - 40)/6) = P(Z \le -1.58)$

$= 0.5 - A(1.58) = 0.5 - 0.4429 = 0.0571$



Standard Normal Curve

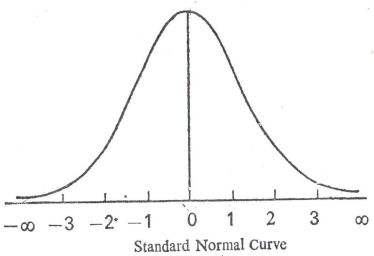
ii) $P(30 < X < 50) = P(29.5 < X < 50.5)$

$$= P\left[\frac{29.5 - 4.0}{6} \le Z \le \frac{50.5 - 40}{6}\right]$$

$= P(-1.75 < Z < 1.75)$

$= A(1.75) + P(1.75)$

$= 0.4599 + 0.4599 = 0.9198$



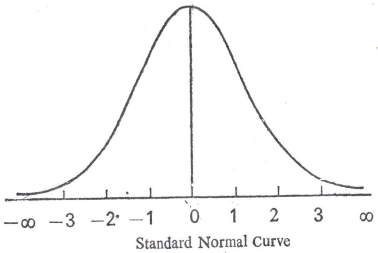
Standard Normal Curve

iii) $P(X > 55) = P(X > 54.5) = P\left[Z > \frac{54.5 - 40}{6}\right]$

$= P(Z > 2.42)$

$= 0.5 - A(2.42)$

$= 0.5 - 0.4922 = 0.0078$



Standard Normal Curve

Gamma Distribution:

Definition: A continuous random variable X is said to follow gamma distribution with parameters a and ə > 0, if the p.d.f. is given by:

$$f(x) = \begin{cases} \frac{a^{ə}}{\tilde{u}(ə)} e^{-ax} x^{ə-1}; Ŭ > 0, ə > 0, 0 < x < \infty \\ 0, \quad \text{otherwise} \end{cases}$$

It is generally denoted by ɔ (a, ə)

If Ŭ = 0, gamma distribution simply denoted by ɔ (ə).

Properties:

i) It is a continuous distribution with mean $\frac{\lambda}{a}$ and variance $\frac{\lambda}{a^2}$

ii) If Ŭ < 1, variance > mean; if Ŭ > 1, variance < mean and if Ŭ = 1, variance = mean

iii) Gamma distribution is positively skewed and leptokurtic in shape

Beta Distribution:

Definition-1: A continuous random variable X is said to have a beta distribution of first kind with parameters μ and ɜ (μ > 0, ɜ > 0) if its p.d.f. is given by:

$$f(x) = \begin{cases} \frac{1}{ɓ(ε, ɜ)} x^{ε-1} (1-x)^{ɜ-1}; & (ε, ɜ) > 0, 0 < x < 1 \\ 0, \quad \text{otherwise} \end{cases}$$

where, ɓ (ε, ɜ) is the beta function.

Properties:

It is a continuous distribution with mean $\frac{ε}{(ε+ɜ)}$ and variance $\frac{εɜ}{(ε+ɜ)^2(ε+ɜ+1)}$

Definition-2: A random variable X is said to have a beta distribution of second kind with parameters μ and ɜ (μ > 0, ɜ > 0) if its p.d.f. is given by:

$$f(x) = \begin{cases} \frac{1}{ɓ(ε, ɜ)} \frac{x^{ε-1}}{(1+x)^{ε+ɜ}} & (ε, ɜ) > 0, 0 < x < \infty \\ 0, \quad \text{otherwise} \end{cases}$$

Properties:

It is a continuous distribution with mean $\frac{ε}{ɜ-1}$ and variance $\frac{ε(ε+ɜ-1)}{(ν-1)^2(ɜ-2)}$

Note: Both beta type-I and type-II distributions are same except in respect of range. Beta type-I has the range (0, 1) and beta type-II has range (0, Ð).

EXERCISES

1. The probability that an evening college student will become postgraduate is 0.4. Determine the probability that out of 5 students a) none, b) one, c) atleast one will become postgraduate.

2. The incidence of a certain disease is such that on an average 20% of workers suffer from it. If 10 workers are selected at random, find the probability that (i) exactly 2 workers (ii) Not more than 2 workers suffer from the disease.

3. The mean of binomial distribution is 40 and standard deviation 6. Calculate n, p and q.

4. The appearance of 2 or 3 on a die is considered as success. Five dice are thrown 729 times and the following results are obtained

Number of Successes	**:**	0	1	2	3	4	5
Frequency	**:**	45	195	237	132	81	39

   Assuming dice to be unbiased, fit the binomial distribution.

5. The average number of telephone calls received in an exchange between 2 PM and 3 PM is 2.0 per minute. Find the probability that number of calls received during a particular minute are

   (i) 0; (ii) 1; (iii) atleast 2 (Hint: Apply Poisson distribution with $\lambda = 2.0$)

6. If the probability that an individual suffers a bad reaction of a given medicine is 0.002, determine the probability that out of 2000 individuals (i) Exactly 3 individuals (ii) more than 2 individuals will suffer from reaction.

7. The following table gives the number of days in a 50 day period during which automobile accidents occurred in a city.

No. of Accidents	:	0	1	2	3	4
No. of Days	:	21	18	7	3	1

   Fit a Poisson distribution to the data.

8. Assuming the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches^2. How many soldiers in a regiment of 1000 would you expect to be over 6 feet tall?

9. In a certain examination 10% of the students got less than 30 marks and 97% of the students got less than 62 marks. Assuming distribution to be normal, find the mean and standard deviation of marks.

10. How many workers have a salary above Rs 6750 in the distribution whose mean is Rs 5000 and standard deviation is Rs 1000 and the number of workers in the factory is 15000. Assume that salary of workers follows the normal law.

CHAPTER-IV

STATISTICAL INFERENCE

Statistical Inference is a branch of statistics which deals with drawing conclusions about the population on the basis of few observations (sample) and sampling is a process of selecting a small fraction (sample) from the population so that it possesses the characteristics of the population.

In applied investigations especially in agricultural and allied sciences, it may not be possible to study the whole population and thus the investigator is forced to draw inferences about the population on the basis of the information obtained from the sample data (due to high operational cost and time consideration), which is called as statistical inference. For example, it is out of question to harvest and record the produce from all the fields growing the wheat crop which constitute the population under study.

4.1 Some Basic Concepts:

Population: The word population in Statistics is used to refer to any aggregate collection of individuals possessing a specified characteristic e.g. population of teachers in HAU, plants in a field etc. A population containing a limited number of individuals or members is called a finite population, whereas a population with unlimited number of individuals or members is known as infinite population. The population of books in a library, population of Indian students in UK are examples of finite population, whereas fish population in Pacific Ocean and the number of stars in the sky are infinite **populations.**

Sample: A portion or a small section selected from the population by some sampling procedure is called a sample.

Census: The recording of all the units of a population for a certain characteristic is known as census or complete enumeration.

Parameter: Parameter are the numerical constants of the population, e.g. population mean (μ), population variance (σ^2) etc.

Statistic: Any function of sample observations is called a statistic. Its value may vary from sample to sample, e.g. sample mean (x) and, sample variance (s^2) are the statistics.

Some commonly used Sampling Techniques: For drawing a sample from the population, several sampling designs are used, some of which are discussed as under:

Probability and Non-Probability Sampling

Probability Sampling: This is a method of selecting a sample according to certain laws of probability in which each unit of population is assigned some definite probability of being selected in the sample. The followings are some such sampling methods:

Simple Random Sampling (SRS): It is the simplest and most commonly used method in which the sample is drawn unit by unit with equal probability of selection at each draw for each unit. Hence simple random sampling is a method of selecting ón' units out of a population of size óN' by assigning equal probability to all units. It is a sampling procedure in which all possible combinations of ón' units that may be formed from the population of óN' units have the same probability of selection. The procedure, where a selected unit is replaced in the population and ón' units are drawn successively is called simple random sampling with replacement (SRSWR). If sample is drawn without replacing the units selected at each draw, it is called simple random sampling without replacement (SRSWOR). In SRSWR, there are N^n possible samples of size n each with probability of selection $\frac{1}{N^n}$ that can be drawn from a population of size N, while in SRSWOR, there are ${}^{N}C_{n}$ possible samples each with probability of selection as $1 / {}^{N}C_{n}$.

Methods of Drawing a Random Sample:

i) Lottery Method: Suppose we wish to draw a random sample of size n from a finite population of N units. We take N pieces of paper of the same size and shape and number them from 1 to N such that each unit in the population corresponds to one piece of paper. These pieces are then put in a container and mixed up thoroughly and a sample of n pieces is drawn either one by one or in a single stroke. The sampling units bearing the numbers on the selected pieces will constitute the desired random sample. Lottery method cannot be applied if the study population is infinite.

ii) Random Number Table Method: Suppose a random sample of n units is to be taken from a population of N units. We mark all the units serially from 1 to N, and take any page of random number table. Starting from anywhere either row-wise or column-wise, random numbers are selected in the sample ignoring the values greater than N. In this method all the digits greater than N are rejected.

Sampling with and without Replacement: In lottery method, while drawing a piece of paper from the container, we may have the choice of replacing or not replacing the selected piece into the container before the next draw is made. In the first case the

numbers can come up again and again, while in the second case they can come up only once. The sampling in which each member of a population may be chosen more than once is called sampling with replacement, whereas the sampling in which no member can be chosen more than once is called sampling without replacement.

Stratified Sampling: Stratified sampling technique is generally followed when the population is heterogeneous and where it is possible to divide it into certain homogeneous sub-populations, say k, called strata. The strata differ from one another but each is homogeneous within itself. The units are selected at random from each of these strata. The number of units selected from different strata may vary according to their relative importance in the population. The sample, which is the aggregate of the sampled units from various strata, is called a stratified random sample and the technique of drawing such a sample is known as stratified sampling or stratified random sampling.

Advantages of Stratified Sampling over Random Sampling:

i) The cost per observation in the survey may be reduced

ii) Estimates of the population parameters may be obtained for each sub-population

iii) Accuracy at given cost is increased

iv) Administrative control is much better as compared to simple random sampling

Systematic Sampling: If the sampling units are arranged in a systematic manner and then a sample is drawn not at random but by taking sampling units systematically at equally spaced intervals along some order. The sample obtained in this manner is called a systematic sample and the technique is called the systematic sampling.

Cluster Sampling: In some situations the elementary units are in the form of groups, composed of smaller units. A group of elementary units is called a cluster. The procedure in which sampling is done by selecting a sample of clusters and then carrying out the complete enumeration of clusters is called cluster sampling. For example in taking a sample of households we select a few villages and then enumerate them completely. Cluster sampling is typically used when the researchers cannot get a complete list of the members of a population they wish to study but can get a complete list of groups or 'clusters' of the population. It is also used when a random sample would produce a list of individuals so widely scattered that surveying them would prove to be much expensive. This sampling technique may be more practical and/or economical than simple random

sampling or stratified sampling. The systematic sampling may also be taken as the cluster sampling in which a sample of one cluster is taken and then it is completely investigated.

Multistage Sampling:

The cluster sampling is more economical but the method restricts the spread of the sample over the population which increases the variance of the estimator. The method of sampling which consists in first selecting the clusters and then selecting specified elements from each cluster is known as two stage sampling. Here clusters which form the units of sampling at the 1^{st} stage are called first stage units (fsu) or primary stage units (psu) and the elements within clusters are called second stage unit (ssu). This procedure can be generalized for more than two stages which is termed as multistage sampling. For example, in crop survey, district may be fsu, blocks as ssu and village may be considered as third stage units or ultimate stage units (usu).

Non-Probability Sampling: It is a sampling procedure in which the sample units are selected not according to law of chance but according to some prior judgement. This procedure is adopted when we wish to collect some confidential information or quick information at low cost. Popular non-probability sampling schemes are purposive, judgement sampling and quota sampling

4.2 Sampling Distributions:

If all possible samples of size n are drawn from a given population and for each sample the value of a statistic, such as the mean, variance etc. is calculated. The value of the statistic will vary from sample to sample and resulting distribution of the statistic is called its sampling distribution. If the particular statistic is the sample mean, the distribution is called the sampling distribution of the mean and so on. The standard deviation of a sampling distribution of a statistic is often called its standard error.

Sampling Distribution of Sample Mean ($\overline{X}$):

Suppose X_1, X_2,é .., X_n is a random sample from normal population N (μ, σ^2). Then sample mean is $\overline{X} = \frac{1}{n}$Ł$X_i$

We assume that the variance σ^2 is known. Since $\overline{X}$ is a linear combination of normal variates, therefore, distribution of $\overline{X}$ is also normal having mean $\varepsilon_{\bar{x}}$ and variance $\mathring{u}^2_{\bar{x}}$ where.

$$\varepsilon_{\bar{x}} = E[\bar{X}] = E\left[\frac{X_1 + X_2, ..., X_n}{n}\right] = \frac{1}{n}\left[E(X_1) + E(X_2) + ... + E(X_n)\right]$$

$$= \frac{1}{n}\left[\varepsilon + \varepsilon + ... + \varepsilon\right] = \frac{n\varepsilon}{n} = \varepsilon$$

and $\sigma^2 = V(\bar{X}) = V\left[\frac{X_1 + X_2 + ..., + X_n}{n}\right] = \frac{1}{n^2}\left[V(X_1) + V(X_2) + ..., + V(X_n)\right]$

$$= \frac{1}{n^2}\left[\mathring{u}^2 + \mathring{u}^2 + ... + \mathring{u}^2\right] = \frac{n\mathring{u}^2}{n^2} = \frac{\mathring{u}^2}{n}$$

Thus, sampling distribution of $\bar{X}$ is also normal with mean μ and variance σ^2/n.

or $\bar{X} \sim N\left(\varepsilon, \frac{\mathring{u}^2}{n}\right)$ and $Z = \frac{\bar{X} - \varepsilon_{\bar{X}}}{\mathring{u}/\sqrt{n}} \sim N(0, 1)$

When parent population is not normal, the sampling distribution will depend on the form of the parent population. However, as n gets large, the form of the sampling distribution will become more and more like a normal distribution, no matter what the parent distribution is. This is stated in the following theorem, which is popularly known as central limit theorem.

Central Limit Theorem: If random samples of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ will have a distribution approximately normal with mean μ and variance σ^2/n. The approximation becomes better as n increases.

If σ^2 is unknown, then an estimate of σ^2 is obtained from the sample, which is given by $s^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$. Further if n < 30, then $t = \frac{\bar{X} - \varepsilon_{\bar{X}}}{s/\sqrt{n}}$ follows a student's t-distribution with (n-1) degrees of freedom and when n > 30, $Z = \frac{\bar{X} - \varepsilon_{\bar{X}}}{s/\sqrt{n}}$ follows a standard normal distribution i.e. N (0, 1). Also, the standard error of the statistic sample mean ($\bar{X}$) is given by SE ($\bar{X}$) $= \frac{s}{\sqrt{n}}$.

Sampling Distribution of Difference of Means: Let $x_1, x_2, \cdots\cdots, x_{n1}$ and $y_1, y_2, \cdots\cdots, y_{n2}$ be two independent random samples of sizes n_1 and n_2 from N (μ_1, σ_1^2) and N (μ_2, σ_2^2), respectively.

Then the statistics $\overline{X} = \frac{1}{n_1}\sum X_i$ and $\overline{Y} = \frac{1}{n_2}\sum Y_i$ have sampling distributions $N\left(\varepsilon_1, \frac{\sigma_1^2}{n_1}\right)$ and $N\left(\varepsilon_2, \frac{\sigma_2^2}{n_2}\right)$ respectively provided the population variances σ_1^2 and σ_2^2 are known. The sampling distribution of the difference of means $\overline{X} - \overline{Y}$ is normal with mean $\varepsilon_{\bar{x}-\bar{y}}$ and variance $\sigma^2_{\bar{x}-\bar{y}}$ where $\varepsilon_{\bar{x}-\bar{y}} = \mu_1 - \mu_2$ and $\sigma^2_{\bar{x}-\bar{y}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Also standard error of the difference of means $\overline{X} - \overline{Y}$

$$SE_d(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ and thus the statistic } Z = \frac{(\overline{X} - \overline{Y}) - (\varepsilon_1 - \varepsilon_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

If σ_1^2 and σ_2^2 are unknown but equal and samples are small, then the pooled estimate s_p^2 of common variance σ^2 is given by:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \text{where } s_1^2 = \frac{1}{n_1 - 1}\sum_i (x_i - \bar{x})^2; s_2^2 = \frac{1}{n_2 - 1}\sum_i (y_i - \bar{y})^2$$

Then $SE(\overline{X} - \overline{Y}) = s_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ and the statistic $t = \frac{(\bar{x} - \bar{y}) - (\varepsilon_1 - \varepsilon_2)}{s_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ follows

student's t-distribution with $(n_1 + n_2 - 2)$ d.f. In case of large samples the statistic

$$Z = \frac{(\overline{X} - \overline{Y}) - (\varepsilon_1 - \varepsilon_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

Remarks:

1) For Simple Random Sampling without Replacement (SRSWOR) from a finite population, we have

   $E(\bar{x}) = \mu$ and $S.E_m = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}}$ where N is the population size.

2) If the original population is not normal then also for large samples ($n \geq 30$), the sample mean ($\overline{X}$) is approximately normally distributed

   Thus if $X \sim (\mu, \sigma^2)$, then for large samples $\overline{X} \sim N(\mu, \sigma^2/n)$ approximately.

Sampling Distribution of Sample Proportion: Suppose a population is divided into two non-overlapping classes C and C' i.e. individuals possessing a characteristic are put in class C and those not possessing the characteristic in class C' e.g. Smokers and Non-Smokers, Defectives and Non-Defectives, Males and Females etc.

Let A be the number of individuals possessing a particular characteristic in a population of size N and let a be the number of individuals in a sample of size n possessing the characteristic C. Then

$P = \frac{A}{N}$ denotes proportion of units in the population possessing characteristic C

and $p = \frac{a}{n}$ denotes proportion of units in the sample possessing characteristic C

Then the sampling distribution of sample proportion is as follows:

If all possible samples of size n are drawn from a population of size N, then sample proportion (p) is distributed with mean P and variance $\frac{PQ}{n}$ i.e. $p \sim \left(P, \frac{PQ}{n}\right)$ and for large samples (n>30), the distribution of p is approximately normal i.e. $p \sim N\left(P, \frac{PQ}{n}\right)$; where Q = 1 ï P or $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$

The Z statistic obtained above is used for testing the null hypothesis that the population proportion (P) is equal to some specified value P_0 i.e. H_0: $P = P_0$ for large samples.

Sampling Distribution of difference between two Sample Proportions:

Let p_1 and p_2 be two sample proportions obtained from samples of sizes n_1 and n_2 from two populations with population proportions P_1 and P_2, respectively. If all possible samples of sizes n_1 and n_2 are drawn from two populations with population proportions P_1 and P_2, respectively, then the difference between sample proportions p_1 and p_2 i.e. p_1 ï p_2 is distribution with mean P_1 ï P_2 and variance $\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$

For large samples i.e. n_1>30 and n_2>30, the distribution is approximately normal i.e. $p_1 - p_2 \sim N\left(P_1 - P_2, \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)$ which implies

$$Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

The Z statistic obtained above is used for testing the significance of difference between two population proportions for large samples.

The mean and standard error of some of the important statistics are given below:

Statistic	Parameter	S.E.	Estimated S.E.	Remarks
Sample mean ($\overline{X}$)	μ	$\mathring{u}/\sqrt{n}$	$s/\sqrt{n}$	
Sample proportion (p)	P	$\sqrt{PQ/n}$	$\sqrt{pq/n}$	P is the population proportion and Q = 1-P
Difference of two sample means ($\overline{X} - \overline{Y}$)	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	σ_1^2 and σ_2^2 are variances of two populations
Difference of two sample proportions (p_1-p_2)	P_1-P_2	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	P_1 and P_2 are population proportions and $Q_1 = 1-P_1$; $Q_2 = 1-P_2$

Uses of Standard Error:

i) The reciprocal of the S.E. gives an estimate of the reliability or precision of the statistic.

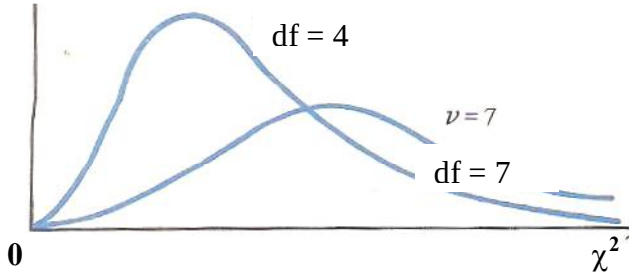
ii) S.E. enables us to determine the confidence limits, which are expected to contain the population parameter.

iii) If t is any statistic, then for large samples $Z = \frac{t - E(t)}{S.E(t)}$ is a standard normal variate having mean zero and variance unity (in case the population is normal, then there is no restriction on the sample size). This Z value forms the basis for testing of hypothesis.

Chi-square (χ^2) Distribution: A continuous random variable X is said to follow χ^2 distribution if its probability density function is:

$$f(x) = ke^{-x/2}(x)^{n/2-1}$$

where k is a constant such that the area under probability density curve is unity. The only parameter (positive integer) of the χ^2 (pronounced as Ky) distribution is ónô which is called the ónumber of degrees of freedomô. Range of χ^2 is from 0 to ∞ i.e. $0 \le \chi^2 \le \infty$.



Shape of the Chi-square distribution curve

Theorem: If $Z_1, Z_2, \ldots, Z_n$ are ónô independent standard normal variates i.e. $Z_i = \frac{x_i - \varepsilon}{\sigma}$ then $Z_1^2 + Z_2^2 + \ldots\ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$ follows a χ^2 distribution with n degrees of freedom. However, if Z_i is taken as $\frac{X_i - \bar{x}}{\sigma}$ where $\bar{x}$ is the sample mean instead of μ then $\sum Z_i^2$ follows χ^2 distribution with (n-1) degrees of freedom.

Properties of χ^2 Distribution:

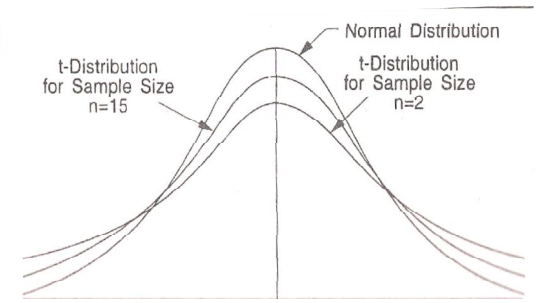
i) The mean and S.D. of χ^2 distribution with n d.f. are n and $\sqrt{2n}$ respectively

ii) χ^2 distribution is positively skewed but with an increased degree of freedom, the χ^2 curve approaches more and more close to the normal curve and for $n \geq 30$ i.e. in the limiting case, χ^2 curve/distribution can be approximated by the normal curve/distribution

iii) Sum of independent χ^2- variates is also a χ^2- variate

Student's 't' distribution: The form of the probability density function for ótô distribution is given by

$$f(t) = K\left(1 + \frac{t^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}; (-\infty < t < \infty)$$



Normal Distribution and t-Distribution curves for sample sizes n = 2 and n = 15

where k is a constant such that total area under the curve is unity. The only parameter n (a positive integer) is called the ñnumber of degrees of freedomò. This distribution was found out by W.S. Gosset (1908) who used the pen-name ñStudentò and hence it is known as studentôs átô distribution. A variable which follows Studentôs distribution is denoted by átô.

If a random sample of size n is drawn from a normal population with mean μ and S.D. σ (unknown), then $\frac{\bar{x} - \varepsilon}{s/\sqrt{n}}$ follows átô distribution with (n-1) d.f. where $\bar{x}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_1 - \bar{x})^2$ denote sample mean and sample variance respectively.

Properties:

i) The curve is leptokurtic with $\beta_2 > 3$ but curve approaches more and more close to the normal curve and for n > 30. t-curve approaches to the normal curve.

ii) t-curve is symmetrical about t = 0 hence mean = mode = median = 0. Its standard deviation is $\sqrt{\frac{n}{n-2}}, (n > 2)$. Also the t-curve extends from -∞ to +∞ like the normal curve.

iv) The degrees of freedom n is the only parameter of the t-distribution.

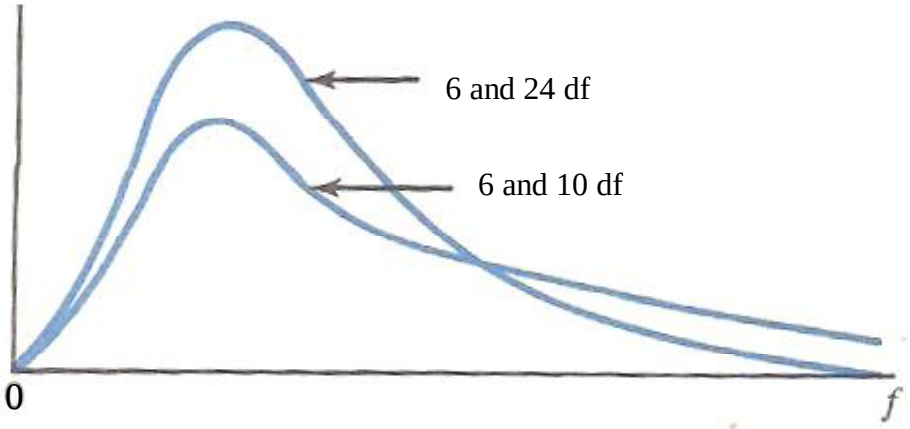
Snedcor's F-Distribution:

F-distribution is defined as the ratio of two independent chi-square variates, each divided by their respective degrees of freedom. Let χ_1^2 and χ_2^2 are independent random variables having Chi-square distributions with degrees of freedom v_1 and v_2 respectively, then distribution of the ratio $F = \frac{\chi_1^2/v_1}{\chi_2^2/v_2}$ follows F-distribution with v_1 and v_2 degrees of freedom.

If we have independent random samples of sizes n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2 respectively, then $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ has F distribution with degrees of freedom (n_1-1) and $(n_2 - 1)$. F-distribution is also known as the variance ratio distribution. This distribution is primarily applied in the analysis of variance, where

we wish to test the equality of several means simultaneously. F-distribution is also used to make inferences concerning the variance of two normal populations.

Definition: A random variable follows F distribution if the probability density function is defined by $f(f) = Kf^{\left(\frac{v_1}{2}\right)-1}(3_2 + 3_1 f)^{-\left(\frac{3_1+3_2}{2}\right)}$ where ófô ranges between 0 to ∞ and K is constant. The parameters 3_1 and 3_2 gives the degrees of freedom (3_1, 3_2) of the distribution. The distribution was proposed by Snedecor and named óFô in honour of the distinguished statistician Sir R.A. Fisher.



F-distribution curve

Properties:

i) The F-distribution is positively skewed with ϭ$_1$ but for g_1 and g_2 both greater than 30, F-curve/distribution can be approximated by the normal curve

ii) Mean of F-distribution is $\frac{\upsilon_2}{\upsilon_2 - 1}$ and variance is $\frac{2\upsilon_2^{\,2}(\upsilon_1 + \upsilon_2 - 2)}{\upsilon_1(\upsilon_2 - 2)^2(\upsilon_2 - 4)}$

iii) If $X \sim Fg_1, g_2$ then $\frac{1}{X} \sim Fg_2, v_1$

iv) υ_1 and υ_2 are the only two parameters of the F-distribution

4.3 Point and Interval Estimation:

The use of sample statistic such as sample mean ($\overline{X}$) sample variance (s^2), sample proportion (p) etc. to draw conclusions about the population parameters such as mean (μ), variance (σ^2), proportion (P) is of fundamental importance in statistical inference.

The following two concepts are used for drawing valid inferences about the unknown population parameter based upon random samples

i) **Estimation:** A procedure of using a sample statistic to approximate a population parameter is called estimation. A statistic used to estimate a population parameter value is called an **estimator** and the value taken by the estimator for a particular sample is called an **estimate.**

A few examples are given below for illustration:

a) A breeder needs to know the average yield of a newly released variety V_1

b) A manufacturer needs to know the proportion of second quality items of his product.

c) A social worker needs to know the proportion of families having three or more children in a particular region.

ii) **Testing of Hypothesis:** It is a procedure to test the claim or belief about an unknown parameter value and will be discussed in the next chapter.

There are two types of estimators: (i) Point estimator and (ii) confidence interval estimator.

Point Estimation: When the estimator of unknown population parameter is given by a single value or point value then it is called **Point Estimator**. For example, sample mean $(\overline{X})$ and sample variance (s^2) are point estimators of the population mean (μ) and population variance (σ^2) respectively.

There are several alternative estimators which might be used for estimating the same parameter. For example, the population mean is a measure of central tendency of the population values and sample mean, sample median and sample mode may be considered as the possible estimators of the population mean. Now the question arises which sample statistic should be used as the estimator of the population parameter. The best estimator is one that is more suitable to a given problem, and has same desirable properties like unbiasedness, consistency, efficiency and sufficiency.

Properties of a Good Estimator: A good estimator is one which is close to the parameter being estimated. Some of the desirable properties of a estimator are discussed below:

i) **Unbiasedness:** An estimator is a random variable and it is always a function of sample observations. If the expected value of an estimator is equal to the population parameter, then it is called an **unbiased estimator**. Thus an estimator

(of a parameter) is said to be unbiased for parameter θ, if E(t) = θ; if E(t) ≠ θ, then it is biased estimator of θ and Bias of the estimator is given by B(t) = E(t) ï θ. For example, sample mean is an unbiased for population mean and sample variance computed with the divisor (n-1) is also unbiased for σ^2.

ii) Consistency: It is a limiting property of an estimator i.e. it concerns with the behavior of the estimator for large sample sizes. If the difference between estimator and the corresponding population parameter continues to become smaller and smaller as the sample size increases, i.e. the estimator coverages in probability to the population parameter then it is called consistent estimator of that parameter. Symbolically, if t_n is an estimator computed from a sample of size n and θ is the parameter being estimated and Pr $\left[\left|t_n - \theta\right| < \in\right] \to 1$, as $n \to \infty$ for any positive ε, however small, then t_n is said to be consistent estimator of θ. It is true for $\overline{X}$ and s^2 which are consistent estimators of μ and σ^2 respectively.

iii) Efficiency: An estimator t_1 is said to be more efficient than t_2 for parameter θ if MSE (t_1) < MSE (t_2) where MSE (t) = E $(t-\theta)^2$ is the mean square error of the estimator t. It can be proved that sample mean is more efficient estimator of population mean μ then the sample median.

iv) Sufficiency: A sufficient estimator is one that uses all the information about the population parameter contained in the sample. It ensures that all information that a sample can furnish with respect to the estimation of a parameter is utilized. It may be noted that sample mean $\overline{X}$ and sample proportion (p) are sufficient estimators of corresponding parameters since all the information in the sample is used in their computation. On the other hand, sample mid-range is not a sufficient estimator since it is computed by averaging only the highest and lowest values in the sample.

Interval Estimation:

Point Estimates provides no information regarding the reliability of estimates i.e. how close an estimate is to the true population parameter. Thus, point estimators are rarely used alone to estimate the population parameters. It is always better to construct an

interval estimator which contains the population parameter so that the reliability of the estimator can be measured. This is the purpose of interval estimation.

When the estimator of a parameter is given in the form an interval (A, B) with a specified level of confidence instead of a single value, it is called **Confidence Interval Estimator**.

Definition: The interval (A, B) is said to be (1-α)x100% Confidence Interval for the population parameter θ if P (A ≤ θ ≤ B) = 1-α.

The quantities A and B are called **Confidence Limits** or **Fiducial Limits** while (1-Ŭ) 100% is called **level of confidence**.

Remarks:

i) By putting α = 0.05 and 0.01 we have 95% and 99% confidence intervals.

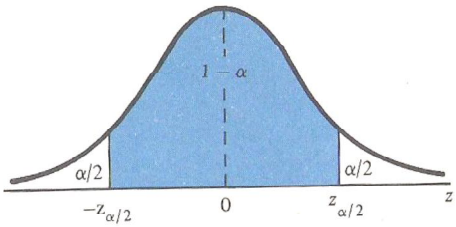
ii) By 95% confidence interval, we mean that the interval will contain the parameter in atleast 95% cases if the experiment is repeated with different samples.

Construction of C.I. for Population Mean μ:

Case-1: When population variance σ^2 is known or sample size n ≥ 30

Let $x_1, x_2, \cdots, x_n$ be a random sample of size n from N (μ, σ^2), then $Z = \frac{\overline{X} - \varepsilon}{\mathring{u}/\sqrt{n}} \sim N(0,1)$ is a standard normal variate. The (1-α) 100% C.I. for μ is given by

$$P\left[-z_{\check{U}/2} \le \frac{\overline{X} - \varepsilon}{\mathring{u}/\sqrt{n}} \le z_{\check{U}/2} \right] = 1 - \check{U}$$



where $z_{\alpha/2}$ is the Z-value for which the area on the right tail of standard normal curve is α/2

or $$P\left[\overline{X} - z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}} \le \varepsilon \le \overline{X} + z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}} \right] = 1 - \check{U}$$

Thus (1-α) 100% upper and lower confidence limits for μ are $\overline{X} \pm z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}}$ and the quantity $z_{\check{U}/2} \frac{\mathring{u}}{\sqrt{n}}$ is called margin of error whereas $\frac{\mathring{u}}{\sqrt{n}}$ is called standard error of sample mean.

Values of the Standard Normal Variate

Confidence level (1-α) 100%	**90%**	**95%**	**98%**	**99%**
Level of significance α	0.10	0.05	0.02	0.01
z_α (for one tailed test)	1.28	1.64	2.05	2.33
$z_{\alpha/2}$ (for two tailed test)	1.64	1.96	2.33	2.58

Thus the 95% and 99% confidence limits for μ are $\overline{X} \pm 1.96$ ů/$\sqrt{n}$ and $\overline{X} \pm 2.58$ ů/$\sqrt{n}$ respectively.

Remarks:

i) Even if σ^2 is unknown but for large sample size ($n \geq 30$), we can substitute sample standard deviation s in place of σ and the interval estimator for μ is given by $\overline{X} \pm z_{\alpha/2} s/\sqrt{n}$.

ii) The confidence limits as given above are applicable for sampling from infinite population or sampling with replacement from finite population. For sampling without replacement from finite population, standard error will be multiplied by the factor $\sqrt{(N-n)/(N-1)}$ known as **finite population correction factor**.

Case-2: When σ^2 is unknown and n<30: For this situation, the statistic $\frac{\overline{x} - \varepsilon}{s/\sqrt{n}}$ follows student's t-distribution with (n-1) d.f. In this case Z-values are replaced by t-distribution values with (n-1) d.f. Thus the (1-α) 100% confidence limits for μ are $\overline{X} \pm t_{\alpha/2,(n-1)}\ s/\sqrt{n}$ where $t_{\alpha/2,(n-1)}$ is the tabulated value of t leaving an area α/2 on the right tail of t- distribution curve with (n-1) d.f. and $s^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2$ is the sample variance.

	Confidence interval for μ (summary)	
Sample size	**σ known**	**σ unknown**
Large	$\overline{X} \pm z_{\alpha/2}\ \sigma/\sqrt{n}$	$\overline{X} \pm z_{\alpha/2} s/\sqrt{n}$
Small	-do-	$\overline{X} \pm t_{\alpha/2,(n-1)}\ s/\sqrt{n}$

Confidence interval for population proportion (P):

For large sample, the confidence interval for P with (1-α)100% level of confidence is given by $p \pm z_{\alpha/2}\sqrt{PQ/n}$ where Q = 1-P. Since P and Q are unknown and

they are estimated by sample statistics p and q respectively and the confidence limits for P can be taken as $p \pm z_{\alpha/2}\sqrt{pq/n}$.

Example-1: 400 labourers were selected from a certain district and their mean income was found to be Rs. 700 per week with a standard deviation of Rs. 140. Set up 95% and 98% confidence limits for the mean income of the labour community of the district.

Solution: Given $\bar{x} = 700$; $s = 140$ and $n = 400$

$SE_m = s/\sqrt{n} = 140/20 = 7$

The confidence limits for population mean (average weekly income)

$= \bar{x} \pm z_{\alpha/2}\ s/\sqrt{n}$

Thus 95% confidence limits are: $700 \pm 1.96(7) = 700 \pm 13.72 = (686.28, 713.72)$

and 98% confidence limits are: $700 \pm 2.33(7) = 700 \pm 16.31 = (683.69, 716.31)$

Example-2: A random sample of 144 families shows that 48 families have two or more children. Construct a 90% confidence interval for the proportion of families having two or more children.

Solution: Sample proportion $p = 48/144 = 1/3$

The confidence limits for population proportion P are given by $p \pm z_{Ŭ2}\sqrt{\frac{pq}{n}}$

where $p = 1/3$, $q = 2/3$, $z_{0.05} = 1.645$, $n = 144$

$SE_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(1/3)(2/3)}{144}} = 0.0393$

Thus the 90% confidence interval is

$\frac{1}{3} \pm 1.645(0.0393) = 0.333 \pm 0.065 = (0.268, 0.398)$

4.4 Testing of Hypothesis:

One of the prime objectives of experimentation whether it is in field or laboratory, is the comparison of treatments means and variances under study. A researcher is usually interested in the following comparisons for drawing logical conclusions of the study undertaken by him. The statistical tests which are applicable in different situations are also given:

i) Comparison of a treatment mean with its hypothetical value (one sample Z-test and one sample t-test)

ii) Comparison of two treatment means (two sample Z-test, two sample t-test and paired t-test)

iii) Comparison of two population variances (F-test)

Some Basic Concepts to Hypotheses Testing:

Any statement or assumption about the population or the parameters of the population is called as **statistical hypothesis**.

The truth or falsity of a statistical hypothesis is never known with certainty unless we examine the entire population. This, of course, would be impractical in most situations. Instead, we take a random sample from the population of interest and use the information contained in this sample to decide whether the hypothesis is likely to be true or false. Evidence from the sample if is inconsistent with the stated hypothesis leads to the rejection of the hypothesis, whereas evidence supporting the hypothesis leads to its acceptance. The investigator should always state his hypothesis in a manner so as to test it for possible rejection. If the investigator is interested in a new vaccine, he should assume that the new vaccine is not better than the vaccine now in the market and then set out to reject this contention. Similarly, to test if the new ploughing technique is superior to old one, we test the hypothesis that there is no difference between these two techniques.

The hypothesis which is being tested for possible rejection is referred to as **Null hypothesis** and is represented by H_0 where as the hypothesis complementary to the null hypothesis is referred to as **Alternative hypothesis** and is represented by H_1.

These hypotheses are constructed such that the acceptance (or rejection) of one leads to the rejection (or acceptance) of the other. Thus if we state the null hypothesis as H_0 : μ (yield of c.v. WH-542) = 65 q/ha, then the alternative hypothesis might be H_1 : $\mu \neq 65$ q/ha or $\mu > 65$ q/ha or $\mu < 65$ q/ha.

Two Types of Errors in Hypothesis Testing: In order to make any decision to accept or to reject the null hypothesis we have to draw a random sample $x_1, x_2, ..é , x_n$ of size n from the population under study and on the basis of the information contained in the

sample we have to decide whether to accept or reject the null hypothesis. Because of the random nature of the sample the above decision could lead to two types of errors.

	Decision	
	Accept	**Reject**
H_0 is true	Correct decision (no error)	Type ï I error
H_0 is false	Type ï II error	Correct decision (no error)

Type I error occurs when we reject a true null hypothesis, i.e. we reject the null hypothesis when it should be accepted. **Type II error** is committed when we accept a false null hypothesis, i.e. we accept the null hypothesis when it should be rejected.

The relative importance of these two types of errors depends upon the individual problem under study. For instance, in the above example, it is expensive to replace the existing variety V_1 and so one should be very careful about the type I error. Whatever may be the relative importance of these errors, it is preferable to choose a test for which the probability of both types of error is as small as possible. Unfortunately, when the sample size n is fixed in advance, it is not possible to control simultaneously both types of errors. What is possible is to choose a test that keeps the probability of one type of error a minimum when the probability of other type is fixed. It is customary to fix type I error and to choose a test that minimize the probability of type II error.

Level of significance is the probability of committing a type I error i.e. it is the risk of rejecting a true null hypothesis. It is denoted by the symbol Ŭ

On the other hand, the probability of committing a type II error is denoted by the symbol ɓ and consequently (1- ɓ) is called the **power of the test**. There is no hard and fast rule for the choice of Ŭ, it is customary to choose Ŭ equal to 0.05 or 0.01. A test is said to be significant if H_0 is rejected at Ŭ= 0.05 and is considered as highly significant if H_0 is rejected at Ŭ= 0.01.

P-value: It indicates the strength of evidence for rejecting the null hypothesis H_0, rather than simply concluding 'reject H_0' or 'do not reject H_0'. Small p-values suggest that the null hypothesis is unlikely to be true. The smaller, it is the more convincing, is the rejection of the null hypothesis.

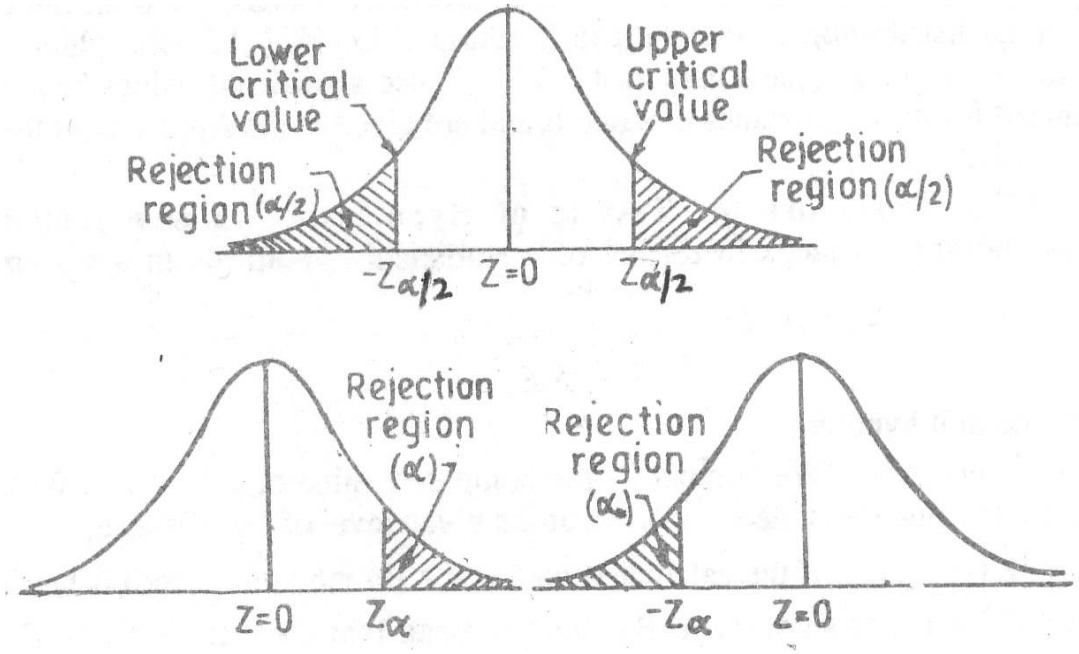
Test statistic: It is the statistic whose value is calculated from the sample data and then compared with critical or table value to decide whether to reject or accept H_0.

The procedure of testing any hypothesis consists of partitioning the total sample space in two regions. One is referred to as **region of rejection** or the **critical region** and other as region of acceptance. If the test statistic on which we base our decision falls in the critical region, then we reject H_0. If it falls in the acceptance region, we accept H_0.

One tailed and two tailed tests:

A test of any statistical hypothesis where the alternative is one sided (right sided or left sided) is called a **one tailed test**. For example, a test for testing the mean of a population H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$ (right tailed) or H_1: $\mu < \mu_0$ (left tailed) is a one tailed test. A test of statistical hypothesis where the alternative is two sided such as: H_0: $\mu = \mu_0$ against H_1 : $\mu \neq \mu_0$ ($\mu > \mu_0$ or $\mu < \mu_0$), is known as two tailed test. The critical or table value of Z for one tailed test at level of significance α is the same as the critical value of Z for a two tailed test at level of significance 2α as shown in the figure.

Two-tailed Test



Right-tailed Test **Left-tailed Test**

4.5 One Sample Tests for Mean:

Here we will discuss tests for determining whether we should reject or accept H_0 about the population mean μ.

Case (i): Population S.D. (σ) is known (One Sample Z-test)

Let a random sample $x_1, x_{2,}$.é ., x_n of size n be drawn from a normal population whose SD(σ) is known. We want to test the null hypothesis that the population mean is equal to the specified mean μ_0 against the alternative hypothesis that the population mean is not equal to μ_0.

Assumptions:

i) Population is normal

ii) The sample is drawn at random

iii) Population S.D. (σ) is known

Procedure:

1. Formulate the null hypothesis $H_0: \mu = \mu_0$
2. Formulate the alternative hypothesis $H_1: \mu \neq \mu_0$

   Situation (i) $H_1: \mu \neq \mu_0$ (Two tailed test)

   Situation (ii) $H_1: \mu > \mu_0$ (Right tailed test)

   Situation (iii) $H_1: \mu < \mu_0$ (Left tailed test)
3. Choose the level of significance $\alpha = 0.05$ or 0.01
4. Compute the test statistic value

$$Z_{cal} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

5. **Conclusion:**

   For two tailed test if $|Z_{cal}| \geq z_{\alpha/2}$, reject H_0 otherwise accept H_0.

   For right tailed test if $Z_{cal} \geq z_{\alpha}$, reject H_0 otherwise accept H_0.

   For left tailed test if $Z_{cal} \leq - z_{\alpha}$, reject H_0 otherwise accept H_0.

Example-3: The average number of mango fruit per tree in a particular region is known from a past experience as 520 with a standard deviation 4. A sample of 20 trees gives an average number of fruit 450 per tree. Test whether the average number of fruit selected in the sample is in agreement with the average production in that region.

Solution: A stepwise solution is as follows:

1. $H_0 : \mu = \mu_0 = 520$ fruit
2. $H_1 : \mu \neq$ 520 fruit (Two tailed test)

3. $\alpha = 0.05$

4. $\bar{x} = 450$, $\mathring{u} = 4$ and $n = 20$

$$|Z_{cal}| = \left|\frac{\bar{x} - \varepsilon_0}{\mathring{u}/\sqrt{n}}\right| = \left|\frac{450 - 520}{4/\sqrt{20}}\right| = 78.26$$

Conclusion: $|Z_{cal}| > Z_{tab}$ i.e. 1.96 at $\alpha = 0.05$. Therefore, we reject H_0 and conclude that average number of fruit per tree in the sample is not in agreement with the average production in the region.

Case (ii): If the population S.D. (σ) is not known but sample size is large (say > 30). Still we can use the one sample Z-test.

Assumptions:

i) Population is normal

ii) The sample is drawn at random

iii) Population S.D. (σ) is unknown

iv) Sample size is large

Test statistic, we can use sample S.D. (s) in place of (σ), then

$$Z_{cal} = \frac{\bar{x} - \varepsilon_0}{s/\sqrt{n}}; \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Conclusion: same as in case (i).

Example-4: Certain gram variety tested on 64 plots gave an average yield as 985 kg/ha, and variance 1600 kg^2/ha. Test at 5% level of significance that the experiment agreed with the breeders claim that the average yield of the variety is 1000 kg/ha. Also construct 95% confidence interval for population mean.

Solution: Here $n = 64$, $\bar{X} = 985$ kg/ha and $s^2 = 1600$ kg^2/ha or $s = 40$ kg/ha

$H_0 : \mu_0 = 1000$ kg/ha

$H_1 : \mu_0 \neq 1000$ kg/ha

Level of significance $\alpha = 0.05$

Population variance is unknown and sample is large so, Z-test is used

$$Z_{cal} = \frac{\bar{X} - \varepsilon_0}{s/\sqrt{n}} = \frac{985 - 1000}{40/\sqrt{64}} = \frac{985 - 1000}{40/8} = \frac{985 - 1000}{5}$$

$$= \frac{-15}{5} = -3 \text{ or } \quad |Z_{cal}| = 3.0$$

because $|Z_{cal}| > 1.96$ so, we reject the null hypothesis at 5% level of significance. Hence it can be concluded that experiment does not confirm breederôs claim that average yield of variety is 1000 kg/ha.

95% confidence interval for mean (μ) = $\bar{x} \pm z_{\ddot{y}/2} \frac{s}{\sqrt{n}} = 985 \pm \frac{40}{\sqrt{64}} \times 1.96$

$= 985 \pm 5 \text{ x } 1.96 = (975.2, 994.8)$

Example-5: A sample of 121 tyres is taken from a lot. The mean life of tyres is found to be 39350 kms with a standard deviation of 3267 kms. Could the sample come from a population with mean life of 40000 kms considering $\alpha = 0.02$?

Solution:

H_0 : Mean life of tyres in the population (μ) = 40,000 kms

H_1 : $\mu \neq 40000$ kms (Two tailed test)

Since the sample size (n = 121) is large, therefore, we apply Z-test

$Z = \frac{\overline{X} - \varepsilon_0}{s/\sqrt{n}}$, since σ is not known and therefore, it is replaced by sample standard deviation (s) = 3267 kms

Thus $Z_{cal} = \frac{39350 - 40000}{3267/\sqrt{121}} = \frac{650}{297} = 2.19$

$Z_{\alpha/2} = Z_{005} = 2.33$

Since $|Z_{cal}| < 2.33$, therefore, H_0 cannot be rejected. Hence we conclude that mean life time of tyres in the population is equal to 40000 kms or the sample has been drawn a population with life time equal to 40000 kms.

Case (iii): Population SD (σ) is unknown and sample size is small i.e. < 30 (one sample t-test) this case is important in the sense that it is always feasible and less expensive to have a small sample size.

Let $X_1, X_2, .é ., X_n$ be a random sample of size n from a normal population with SD σ (unknown) and we want to test the null hypothesis that the population mean μ is

equal to a specified value μ_0 against the alternative hypothesis. The stepwise testing procedure is as follows:

Assumptions:

i) Population is normal

ii) The sample is drawn at random

iii) Population SD (σ) is unknown and sample size is small.

Procedure:

1. $H_o : \mu = \mu_0$

2. Situation (i) $H_1 : \mu \neq \mu_0$ (Two tailed test)

   Situation (ii) $H_1 : \mu > \mu_0$ (Right tailed test)

   Situation (iii) $H_1 : \mu < \mu_0$ (Left tailed test)

3. Choose the level of significance $\alpha = 0.05$ or 0.01

4. Test Statistic

   Obtain sample mean $\bar{x}$ and sample SD's

   $\bar{x} = \frac{1}{n}\sum x$ and $s^2 = \frac{1}{n-1}\sum (x_i - x)^2$ and finally compute $t = \frac{\bar{x} - \varepsilon_0}{s/\sqrt{n}}$

5. Obtain tabulated value of 't' distribution with (n-1) df at the level of significance α

6. **Conclusion:**

   For two tailed test if $|t_{cal}| \geq t_{\alpha/2, n-1}$ reject H_0 otherwise accept H_0.

   For right tailed test if $t_{cal} \geq t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

   For left tailed test if $t_{cal} \leq - t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

Example-6: The life time (000ô hours) of a random sample of 11 electric bulbs from a large consignment are 4.2, 3.6, 3.9, 3.1, 4.2, 3.8, 3.9, 4.3, 4.4, 4.6 and 4.0. Can we accept the hypothesis at 5% level of significance that the average life time is more than 3.75.

Solution: H_0 : Average life time (μ) = 3.75

i) $H_1 : \mu > 3.75$ (Right tailed test)

Test Statistic

$$t = \frac{\bar{x} - \varepsilon_0}{s/\sqrt{n}}$$

Calculation of $\bar{x}$ and s

Here n = 11; $\Sigma x = 44$ and $\bar{x} = 4$

x_i	4.2	3.6	3.9	3.1	4.2	3.8	3.9	4.3	4.4	4.6	4.0	**Total**
$x-\bar{x}$	0.2	-0.4	-0.1	-0.9	0.2	-0.2	-0.1	0.3	0.4	0.6	0	
$(x-\bar{x})^2$	0.04	0.16	.01	.81	.04	.04	.01	.09	.16	.36	0	**1.72**

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} = \sqrt{0.172} = 0.415$$

$$t_{cal} = \frac{4 - 3.75}{0.415}\sqrt{11} = 2.00$$

ii) Since $t_{cal} > t_{0.05,\ 10} = 1.812$, therefore H_0 is rejected and it can be concluded that the average life time of bulbs is more than 3750 hours.

Example-7: Ten individuals are chosen at random from a normal population and their heights are found as follows: 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71 inches, respectively. Test whether the mean height is 69.6 inches in the population (use α = 0.05). Also construct the 95% confidence interval for population mean (μ).

Solution: From the given data, we obtain n = 10, $\Sigma x = 678$, $\Sigma x^2 = 46050$, $\bar{x} = 67.8$

$H_0 : \mu_0 = 69.6$

$H_1 : \mu_0 \neq 69.6$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

$$= \frac{1}{9}\left[46050 - \frac{(678)^2}{10}\right] = \frac{1}{9}[46050 - 45968.4] = \frac{81.6}{9} = 9.07$$

Test statistic

$$t_{cal} = \frac{\bar{x} - \varepsilon}{\sqrt{s^2/n}} = \frac{67.8 - 69.6}{\sqrt{\frac{9.07}{10}}} = \frac{-1.8}{\sqrt{0.907}} = -1.89$$

$|t_{cal}| = 1.89$ and $t_{0.05}(9)$ df = 2.26

As $|t_{cal}|$ is less than table t-value at Ŭ= 0.05 at 9 df, the test provides no evidence against null hypothesis and we conclude that the mean of the population is 69.6ò.

95% confidence interval for population mean

$$= \bar{x} \pm t_{\ddot{U}/2} \frac{s}{\sqrt{n}} = 67.8 \pm \frac{3.01}{\sqrt{10}} \times 2.26 = 67.8 \pm \frac{3.01}{3.16} \times 2.26$$

$$= 67.8 \pm 0.95 \text{ x } 2.26 = 67.8 \pm 2.15 = (65.65, 69.95)$$

Example-8: A new feed was given to 25 animals and it was found that the average gain in weight was 7.18 kg with a standard deviation 0.45 kg in a month. Can the new feed be regarded having similar performance as that of the standard feed, which has the average gain weight 7.0 kg?

Solution:

$H_0 : \mu = 7.0$ kg

$H_1 : \mu \neq 7.0$ kg $\alpha = 0.05$ (Two tailed test)

Here $\bar{x} = 7.18$ kg and $s = 0.45$ kg and $n = 25$

$$t_{cal} = \frac{7.18 - 7.0}{0.45/\sqrt{25}} = 2.0$$ t_{tab} at $\alpha = 0.05$ with 24 d.f. = 2.06

Conclusion:

As $|t_{cal}| < t_{tab}$, we accept H_0 at 5% level of significance therefore, we conclude that new feed do not differ in performance than the existing feed.

4.6 Two Sample Tests for Means:

The Comparison of a sample mean with its hypothetical value is not a problem of frequent occurrence. A more common problem is the comparison of two population means. For example, we may wish to compare two training methods or two diets to see their effect on the increase in weight. Here we will like to test the null hypothesis whether the two population means are same (H_0: $\mu_1 = \mu_2$) against the alternative hypothesis that the two population means are different.

Case (i): Population SD's are known (Two sample Z test)

Let X_1, X_2,é .., X_{n1} and Y_1, Y_2,é ..,Y_{n2} be the two independent random samples of sizes n_1 and n_2 from the two normal populations with known standard deviations σ_1 and σ_2 and we want to test the null hypothesis $H_0 : \mu_1 = \mu_2$.

Assumptions:

i) Populations are normal.

ii) Samples are drawn independently and at random.

iii) Population SD's are known.

Stepwise procedure is as follows:

1. Formulate the null hypothesis H_0: $\mu_1 = \mu_2$

2. Formulate the alternative hypothesis that the two means are not equal

   Situation (i) $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

   Situation (ii) $H_1 : \mu_1 > \mu_2$ (Right tailed test)

   Situation (iii) $H_1 : \mu_1 < \mu_2$ (Left tailed test)

3. Choose the level of significance $\alpha = 0.05$ or 0.01

4. Test statistic

Obtain $\overline{X}$ and $\overline{Y}$ from the two independent random samples of size n_1 and n_2 respectively and compute.

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\mathring{u}_1^2}{n_1} + \frac{\mathring{u}_2^2}{n_2}}}$$

Conclusion:

For two tailed test if $|Z_{cal}| \geq z_{\alpha/2}$, reject H_0 otherwise accept H_0.

For right tailed test if $Z_{cal} \geq z_{\alpha}$, reject H_0 otherwise accept H_0.

For left tailed test if $Z_{cal} \leq - z_{\alpha}$, reject H_0 otherwise accept H_0.

Case (ii): Population SD's are unknown but the samples sizes are large (Two sample Z-test).

If the sample sizes are large, then we can replace the population SD's with corresponding sample values s_1 and s_2.

Assumptions:

i) Populations are normal.

ii) Samples are drawn independently and at random.

iii) Population SD's are unknown.

iv) Sizes of the samples are large.

Procedure is same as in case (i) except the test statistic

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ where } s_1^2 = \frac{1}{n_1}\sum (x_i - \bar{x})^2; \qquad s_1^2 = \frac{1}{n_2}\sum (y_i - \bar{y})^2$$

Conclusion: same as in case (i)

Example-9: A random sample of the heights in inches of adult males living in two different countries gave the following results

$n_1 = 640$ $\bar{X} = 67.35$ò and $s_1 = 2.56$ò

$n_2 = 160$ $\bar{Y} = 68.56$ and $s_2 = 2.52$

Test at 0.01 level of significance whether the average height of males in the two countries differ significantly?

Solution: Given

	Country – I	**Country – II**
No. of observation	640	160
Mean	67.35	68.56
s.d	2.56	2.52

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$$Z_{cal} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{67.85 - 68.56}{\sqrt{\frac{(2.56)^2}{640} + \frac{(2.52)^2}{160}}} = \frac{-0.7}{\sqrt{0.01024 + 0.03969}}$$

$$= \frac{-0.7}{\sqrt{0.04993}} = \frac{-0.7}{0.22} = -3.18$$

$|Z_{cal}| = 3.18$

$|Z_{cal}| > Z_{tab} = 2.58$, H_0 is rejected. Hence we conclude that the average heights of males in two countries are different.

Example-10: I.Q. Test of two groups of boys and girls gave the following results

Boys: $\bar{x} = 80$, SD = 10, $n_1 = 30$

Girls: $\bar{y} = 75$, SD = 13, $n_2 = 70$

Is there a significant difference in the mean scores of boys and girls at 5% level of significance?

Solution: H_0 : $\mu_1 = \mu_2$ i.e. mean scores of boys and girls is same.

H_1 : μ_1 Í μ_2 (Two tailed test)

Since both the sample sizes are large, therefore, two sample Z-test is applicable.

$$Z_{cal} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{80 - 75}{\sqrt{\frac{100}{30} + \frac{169}{70}}} = \frac{5}{\sqrt{5.75}} = 2.08$$

Since $|Z_{cal}| > Z_{\alpha/2}$ ($\alpha = 0.05$) = 1.96, therefore, H_0 is rejected and it is concluded that there is significant difference in the mean scores of boys and girls.

Example-11: A random sample of 90 birds of one breed gave on average production of 240 eggs per bird per year with a SD of 18 eggs. Another random sample of 60 birds of another breed gave an average production of 195 eggs per bird/year with a SD of 15 eggs. Is there any significant difference between the two breeds with respect to their egg production?

Solution:

Stepwise solution is as follows:

1. $H_0 : \mu_1 = \mu_2$
2. $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)
3. $\alpha = 0.05$
4. $\bar{x} = 240$ $n_1 = 90$ $s_1 = 18$

   $\bar{y} = 195$ $n_2 = 60$ $s_2 = 15$

   Given test statistic

$$Z_{cal} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{240 - 195}{\sqrt{\frac{(18)^2}{90} + \frac{(15)^2}{60}}} = 16.61$$

Conclusion:

Since $|Z_{cal}| > Z_{tab} = 1.96$ at 5% level of significance, therefore, we reject H_0 and conclude that there is a significant difference between the two breeds of birds with respect to egg production.

Case (iii): Population SD's are unknown but assumed same and sample sizes are small (Two sample t-test)

Let $X_1, X_{2,}\cdots\cdots, X_{n1}$ and $Y_1, Y_{2,}\cdots\cdots, Y_{n2}$ in be two independent random samples of sizes n_1 and n_2 (small) from two normal populations with unknown but equal standard deviations σ_1 and σ_2. Here we want to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative that the population means are different.

Assumptions:

1. Populations are normal.
2. Samples are drawn independently and at random.
3. Population SD's are unknown but assumes to be the same.
4. Sample sizes are small

We proceed by the following steps:

1. $H_0 : \mu_1 = \mu_2$
2. Situation (i) $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

   Situation (ii) $H_1 : \mu_1 > \mu_2$ (Right tailed test)

   Situation (iii) $H_1 : \mu_1 < \mu_2$ (Left tailed test)
3. $\alpha = 0.05$ or 0.01
4. **Test statistic**

   Let $\overline{X}$, $\overline{Y}$ denote the sample means and s_1^2 and s_2^2 be sample variances , then the statistic $t = \dfrac{\overline{X} - \overline{Y}}{s}\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}$ follows Studentôs t-distribution with $(n_1 + n_2 ï 2)$ d.f.

   where $s^2 = \dfrac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2}{n_1 + n_2 - 2}$ is called pooled variance.
5. Obtain t_{tab} at $(n_1 + n_2 ï 2)$ at α level of significance.
6. **Conclusion:**

   For two tailed test if $|t_{cal}| \geq t_{Ï2, n_1+n_2-2}$ reject H_0 otherwise accept H_0.

   For right tailed test if $t_{cal} \geq t_{Ï, n_1+n_2-2}$ reject H_0 otherwise accept H_0.

   For left tailed test if $t_{cal} \leq -\ t_{Ï, n_1+n_2-2}$ reject H_0 otherwise accept H_0.

Example-12: Two types of drugs X and Y were tried on certain persons for increasing weight, 5 persons were given drug X and 7 persons were given drug Y. The increase in weight is given below

Drug X	**:**	7	11	12	8	2	
Drug Y	:	12	10	14	17	8	10	13

Do the two drugs differ significantly with regard to their effect in increasing weight?

Solution: H_0: $\mu_1 = \mu_2$ i.e. there is no significant difference in the efficacy of two drugs.

H_1: $\mu_1 \neq \mu_2$ (Two tailed test)

Since population variances are unknown and sample sizes are small (we assume the population variances to be equal).

Therefore, applying two sample t-test

$$t_{cal} = \frac{\bar{x} - \bar{y}}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Calculation of $\bar{x}, \bar{y}$ and s

	x	x- $\bar{x}$	$(x-\bar{x})^2$	y	y- $\bar{y}$	$(y-\bar{y})^2$
	7	-1	1	12	0	0
	11	3	9	10	-2	4
	12	4	16	14	2	4
	8	0	0	17	5	25
	2	-6	36	8	-4	16
				10	-2	4
				13	1	1
Total	**40**	**0**	**62**	**84**	**0**	**54**

$\bar{x} = 40/5 = 8$; $\bar{y} = 84/7 = 12$

$$s = \sqrt{\frac{\sum (x-\bar{x})^2 + \sum (y-\bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{62+54}{5+7-2}} = 3.406$$

$$t_{cal} = \frac{8-12}{3.406}\sqrt{\frac{5x7}{5+7}} = -2.0$$

Since $|t_{cal}| < t_{\alpha/2}$ ($\alpha = 0.05$, df = 10) = 2.23, therefore, we accept H_0 and conclude that there is no significant difference in the efficacy of two drugs in increasing the weight.

Example-13: An experiment was conducted to compare the effectiveness of two sources of nitrogen, namely ammonium chloride and urea, on grain yield of paddy. The results on the grain yield of paddy (kg/plot) under the two treatments are given below.

Ammonium chloride (x): 13.4, 10.9, 11.2, 11.8, 14.0, 15.3, 14.2, 12.6, 17.0, 16.2, 16.5, 15.7

Urea(y): 12.0, 11.7, 10.7, 11.2, 14.8, 14.4, 13.9, 13.7, 13.7, 16.9, 16.0, 15.6, 16.0

Which source of nitrogen is better for paddy?

Solution: $H_0 : \mu_1 = \mu_2$ The effect of the two source of nitrogen on paddy yield are same

$H_1 : \mu_1 \neq \mu_2$ The effects of two sources of nitrogen on paddy yield are not same.

Let $\alpha = 0.05$

Ammonimum Chloride

$n_1 = 12$

$\Sigma x = 168.8$

$\Sigma x^2 = 2423.72$

$\bar{x} = 14.07$

$$s_1^2 = \frac{1}{n_1 - 1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

$$= \frac{1}{11}\left[2423.72 - \frac{(168.8)^2}{12}\right]$$

$$= \frac{1}{11}[2423.72 - 2374.45]$$

$$= \frac{1}{11}[49.27]$$

$= 4.48$

Urea

$n_2 = 12$

$\Sigma y = 166.8$

$\Sigma y^2 = 2369.09$

$\bar{y} = 13.91$

$$s_2^2 = \frac{1}{n_2 - 1}\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]$$

$$= \frac{1}{11}\left[2369.09 - \frac{(166.9)^2}{12}\right]$$

$$= \frac{1}{11}[2369.09 - 2321.30]$$

$$= \frac{1}{11}[47.79]$$

$= 4.34$

Before applying two-sample t-test, it is required to test the equality of variability in populations, first use F-test.

$$F_{cal} = \frac{s_1^2}{s_2^2} = \frac{4.48}{4.34} = 1.03$$

Table F value at (11, 11) degrees of freedom and $\alpha = 0.05$ level of significance is 2.82.

Here $F_{cal} < F_{tab}$., F is not significant. Therefore, the variances are equal, and we can pool them. The pooled variance is

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{11 \text{ x } 4.48+11 \text{ x } 4.34}{12+12-2} = \frac{11\ (4.48+4.34)}{22} = \frac{97.02}{22} = 4.41$$

$$t_{cal} = \frac{\bar{x}-\bar{y}}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{14.07-13.91}{\sqrt{4.41\left(\frac{1}{12}+\frac{1}{12}\right)}} = \frac{0.16}{\sqrt{0.735}} = \frac{0.16}{0.857} = 0.186$$

Table t-value for 22 df at $\alpha = 0.05$ is 2.074. Since $|t_{cal}|$ is less than table t-value, we accept the null hypothesis and conclude that both the sources of nitrogen have similar effect on the grain yield of paddy.

Example-14: Descriptive summary for samples obtained from two electric bulb manufacturing companies is as under:

	Company A	**Company B**
Mean life (in hours)	1234	1136
Standard deviation (in hours)	36	40
Sample size	8	7

Which brand of bulbs are you going to purchase if you can take a risk of 5%?

Solution:

H_0 : $\mu_1 = \mu_2$ i.e. there is no significant difference in the mean life of two brands of bulbs.

H_1 : $\mu_1 \neq \mu_2$ (Two tailed test)

Applying two sample t-test

$$t = \frac{\bar{x}-\bar{y}}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

$\bar{x} = 1234$; $\bar{y} = 1136$; $n_1 = 8$, $n_2 = 7$, $s_1 = 36$, $s_2 = 40$

$$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{(8-1)(36)^2+(7-1)(40)^2}{8+7-2}} = \sqrt{\frac{9072+9600}{13}} = 37.9$$

$$t_{cal} = \frac{1234-1136}{37.9}\sqrt{\frac{8x7}{8+7}} = 4.99$$

Since $|t_{cal}| > t_{\alpha/2}$ ($\alpha = 0.05$, df = 13) = 2.16, therefore, we reject H_0. Thus bulbs of brand A should be purchased as their mean life time is significantly greater than that of B.

Example-15: For a random sample of 10 animals fed on diet A, the increase in weight in kg in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 kg and for another random sample of 12 animals of the same species fed on diet B, the increase in weights for same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 kg.

i) Test whether diet A and B differs significantly as regards the effect on increase in weight is concerned.

ii) How will we modify the testing procedure if the population variances are known to be 5 and 9 Kg^2

Solution:

1. $H_0 : \mu_1 = \mu_2$

2. $H_1 : \mu_1 \neq \mu_2$ (Two tailed test)

3. $\alpha = 0.05$

4. $$\bar{x} = \frac{\sum x}{n_1} = \frac{120}{10} = 12$$

   $$\bar{y} = \frac{\sum y}{n_2} = \frac{180}{12} = 15$$

   $(n_1-1)\, s_1^2 = \Sigma(x-x)^2 = 120$

   $(n_2-1)s_2^2 = \Sigma \quad (y-y)^2 = 314$

   $$s^2 = \frac{120+314}{10+2-2} = 21.1$$

   $$t_{cal} = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} + \frac{12-15}{\sqrt{21.1\left(\frac{1}{10}+\frac{1}{12}\right)}} = -1.6$$

5. t_{tab} at 5% level of significance with 20 d.f. is 2.086.

6. As $|t_{cal}| < 2.086$, we accept H_0 and conclude that the two diets do not differ significantly.

iii) Here the population variances are known, therefore, we can apply two sample Z-test [case (i)] where $\mathring{u}_1^2 = 5$ and $\mathring{u}_2^2 = 9$

1. $H_0 : \mu_1 = \mu_2$

2. $H_1 : \mu_1 \neq \mu_2$

3. $\alpha = 0.05$

4. Test Statistic

$$Z_{cal} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\mathring{u}_1^2}{n_1} + \frac{\mathring{u}_2^2}{n_2}}} = \frac{12 - 15}{\sqrt{\frac{5}{10} + \frac{9}{12}}} = -2.68$$

Conclusion: As $|Z_{cal}| > Z_{tab} = 1.96$ at 5% level of significance, therefore, we reject H_0 and conclude that diet A differs significantly from diet B as far as increase in weight is concerned.

Case (iv): Population variances are unknown but different

For testing the significance of the differences between two means, we have made the assumption that the variances of two populations are same. Before applying the t-test, it is desirable to test this assumption by F-test by comparing variance ratio s_1^2/s_2^2 ($s_1^2 > s_2^2$) against F distribution with (n_1-1, n_2-1) degrees of freedom. If the two variances are different then in this case we find out t_{cal} as:

$$t_{cal} = \frac{\bar{x} - \bar{y} - (\varepsilon_1 - \varepsilon_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

and compare it with t_{tab} with v d.f. where

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

Case (v) Test for the paired samples (paired t-test): In above tests, we have assumed that the two random samples are independent, but some times in practice we find that two random samples may be correlated. For instance due to the shortage of material, the experiment have to be carried out on same set of units on two different occasions or two varieties/fertilizers are tested on adjacent plots. Other examples where paired t-test may be used are to see (i) effect of coaching in securing good marks (ii) the effect of medical practice in changing the blood pressure etc.

Let there be n pairs and the observations be denoted by (x_1, y_1), (x_2, y_2), é .., (x_n, y_n) and let d_1, d_2,é .., d_n represents the differences of n related pairs of measurements, where $d_i = x_i$ ï y_i, i = 1, 2,·····,n.

Assumptions:

1. Populations are normal.
2. Samples are dependent and taken at random.
3. Population SD's are unknown but equal
4. Sizes of the samples are small.

Procedure:

1. H_0: $\mu_d = 0$ i.e. mean differences in the population are zero.
2. Situation (i) H_1: $\mu_d \neq 0$ (Two tailed test)

   Situation (ii) H_1: $\mu_d > 0$ (Right tailed test)

   Situation (iii) H_1: $\mu_d < 0$ (Left tailed test)
3. Choose the level of significance α
4. Test Statistic Computation

   Compute $\bar{d} = \frac{\sum d_i}{n}$, $s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1}$ and find out $t_{cal} = \frac{\sqrt{n}\,\bar{d}}{s_d}$
5. Obtain t_{tab} at α level of significance with (n-1) d.f.
6. **Conclusion:**

   For two tailed test if $|t_{cal}| \geq t_{\alpha/2, n-1}$ reject H_0 otherwise accept H_0.

   For right tailed test if $t_{cal} \geq t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

   For left tailed test if $t_{cal} \leq - t_{\alpha, n-1}$ reject H_0 otherwise accept H_0.

Example-16: Eleven employees of a company were given training for accounts test. The marks obtained by them before and after training are given below:

Before Training	32	29	28	30	27	29	27	26	32	25	28
After Training	34	29	31	28	31	33	30	30	33	30	37

Test at 5% level of significance that the training has contributed towards increase in marks.

Solution: H_0: $\mu_d = 0$ i.e. the training has no effect on marks of employees

H_1: $\mu_d > 0$ (Right tailed test)

Test statistic:

$t = \frac{\bar{d}\sqrt{n}}{s_d}$, where $\bar{d}$ = mean increase in marks after training

Calculation of $\bar{d}$ and s

												Total
d	2	0	3	-2	4	4	3	4	1	5	9	$\Sigma d=33$
d^2	4	0	9	4	16	16	9	16	1	25	81	$\Sigma d^2=181$

$\bar{d} = 33/11 = 3$ and $s_d = \sqrt{\frac{1}{n-1}\left[\sum d^2 - n(\bar{d})^2\right]} = \sqrt{\frac{1}{10}(181 - 11(3)^2} = 2.863$

$t_{cal} = \frac{3\sqrt{11}}{2.863} = 3.475$

Since $t_{cal} > t_{0.05,10} = 1.812$, therefore, we reject H_0 and conclude that training has contributed significantly towards increase in marks.

Example-17: A drug is given to 10 patients, and the increments in their blood pressure were recorded to be 3, 6, -2, 4, -3, 4, 6, 0, 0, 2. Is it reasonable to believe that the drug has no effect on blood pressure? Use $\alpha = 0.05$.

Solution: H_0: $\mu_d = 0$ i.e. drug has no effect on the blood pressure of patients

H_1: μ_d Í 0 (Two tailed test)

Applying paired t-test

$t = \frac{\bar{d}\sqrt{n}}{s_d}$

Calculation of $\bar{d}$ and s_d

											Total
d	3	6	-2	4	-3	4	6	0	0	2	$\Sigma d=20$
d^2	9	36	4	16	9	16	36	0	0	4	$\Sigma d^2=130$

$\bar{d} = 20/10 = 2$

$s_d = \sqrt{\frac{1}{9}(130 - 10(2)^2} = 3.162$

$$t_{cal} = \frac{2\sqrt{10}}{3.162} = 2$$

Since $|t_{cal}| < t_{0.025,\ 9\ df} = 2.262$, therefore, we do not reject H_0 and conclude that drug has no significant effect on the blood pressure of patients.

Testing of Hypothesis about a Population Proportion:

If all possible samples of size n are drawn from a population of size N, then sample proportion (p) is distributed with mean P and variance PQ/n and for large samples (n Ó30), p is approximately normally distributed, i.e. p ~ N (P, PQ/n) where Q = 1 ï P

or $$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$$

The Z-statistic obtained above is used for testing H_0: $P = P_0$ for large samples.

Testing Hypothesis about Difference of Proportions:

Let p_1, p_2 be two sample proportions obtained from independent samples of sizes n_1 and n_2 from two populations with population proportions P_1 and P_2 respectively.

Here H_0: $P_1 = P_2$ and H_1: P_1 Í P_2

For large samples (i.e. n_1 Ó 30 and n_2 Ó 30), the distribution of p_1 ï p_2 is approximately normal

i.e. p_1 ï $p_2 \sim N(P_1$ ï $P_2, P_1Q_1/n_1 + P_2Q_2/n_2)$

or $$Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0,1)$$

Under H_0, p_1 and p_2 are independent unbiased estimators of the same parameter $P_1 = P_2 = P$. Thus we use the weighted mean of p_1 and p_2 as estimator of P. i.e.

$$p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \text{ and } q = 1 - p$$

thus $$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and is used for testing H_0: $P_1 = P_2$ for large samples.

Example-18: A coin is tossed 900 times and heads appears 480 times. Does this result support the hypothesis that the coin is unbiased at (i) $\alpha = 0.05$ (ii) $\alpha = 0.01$.

Solution:

H_0: Coin is unbiased i.e. the proportion of heads (P) = 0.5

H_1: P Í 0.5 (Two tailed test)

Sample proportion (p) = 480/900 = 0.533

Standard error of (p) = $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.5x0.5}{900}} = 0.0167$

Test statistic

$$Z_{cal} = \frac{p - P}{SE(p)} = \frac{0.533 - 0.500}{0.0167} = 1.98$$

i) Since $|Z_{cal}| > z_{\alpha/2}$ (α=0.05) = 1.96, therefore, H_0 is rejected and it is concluded that the coin is biased.

ii) Since $|Z_{cal}| < z_{\alpha/2}$ (α=0.01) = 2.58, therefore H_0 is accepted and it is concluded that the coin is unbiased.

Example-19: A sales clerk in the departmental store claims that 60% of the customers entering the store leave without buying anything. A random sample of 50 customers showed that 35 of them left without making any purchase. Test the claim of the sales clerk at 5% level of significance.

Solution:

H_0: P = 0.60

H_1: P Í 0.60 (Two tailed test)

Sample proportion (p) = 35/50 = 0.70

Standard error of (p) = $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.60 \text{ x } 0.4}{50}} = 0.0693$

Test statistic

$$Z_{cal} = \frac{p - P}{SE(p)} = \frac{0.70 - 0.60}{0.0693} = 1.44$$

Since $|Z_{cal}| < z_{0.025} = 1.96$, therefore, H_0 is not rejected and it supports the claim of sales clerk.

Example-20: In a random sample of 100 persons taken from a village A, 60 are found consuming tea. In another sample of 200 persons taken from village B, 100 persons are found consuming tea. Do the data reveal significant difference between the two villages so far as the habit of taking tea is concerned?

Solution: H_0: $P_1 = P_2$ Tea habit in the two villages is same

H_1: $P_1 \neq P_2$ (Two tailed test)

$n_1 = 100$ $p_1 = 60/100 = 0.6$

$n_2 = 200$ $p_2 = 100/200 = 0.5$

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = (60 + 100)/(100 + 200) = 0.53$$

$$Z_{cal} = \frac{0.6 - 0.5}{\sqrt{(0.53)(0.47)\left(\frac{1}{100} + \frac{1}{200}\right)}} = \frac{0.1}{\sqrt{(0.53)(0.47)(0.015)}} = 1.64$$

Since $|Z_{cal}| < z_{\alpha/2}$ (α=0.05) = 1.96 therefore H_0 is accepted and it is concluded that there is no significant difference in the habit of taking tea in the two villages A and B.

4.7 Chi-square (χ^2) Test and its Applications:

The Chi-square test (written as χ^2-test) is one of the simplest and most widely used non-parametric test which was given by Karl Pearson (1900).

Assumption:

i) Totals of observed and expected frequencies are same i.e. $\Sigma O_i = \Sigma E_i = N$ and $N > 50$ and

ii) No observed frequency should be less than 5. If any frequency is less than 5, then for the application of Chi-square test it to be pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjusted for the degrees of freedom lost in pooling.

Here, we shall discuss following applications of χ^2 test

i) Test of goodness of fit

ii) Test of independence

iii) Test for the population variance

iv) Test for the homogeneity of several population variances (Bartlettôs test)

Applications of Chi-square:

(1) Test of Goodness of Fit of a Distribution:

H_0 : observed and expected frequencies are in complete agreement i.e. fit is good.

H_1 : observed and expected frequencies are not in agreement.

The goodness of fit of any set of data to a probability distribution can be tested by a chi-square test. For carrying out the goodness of fit test, we calculate expected frequencies E_i corresponding to the observed frequencies O_i on the basis of given distribution i =1, 2,·····, k and compute the value of the test statistic:

$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, which follows Chi-square distribution with (k-1) d.f. and gives the magnitude of discrepancy between expected and observed patterns.

Conclusion: Reject H_0 if χ^2_{cal} is greater than table value of χ^2 at α level of significance with (k-1) d.f. and conclude that fit is not good i.e. the observed frequencies are not in agreement with expected frequencies.

Example-21: The data relate to the distribution of printing mistakes. Fit the Poisson distribution and test for goodness of fit.

x :	0	1	2	3	4	5	6
f :	275	72	30	7	5	2	1

Solution:

H_0: Printing mistakes follow Poisson law

H_0: Printing mistakes does not follow Poisson law

The mean of the distribution $\bar{X} = \frac{\Sigma fx}{N}$ where N=Σf, determines the estimate of λ value equal to $\bar{X}$, since in Poisson distribution mean is equal to λ.

i) Compute the expected frequencies corresponding to observed frequencies by the Poisson distribution as explained in Example-14 (Chapter-3).

ii) Compute $\chi^2_{cal} = \sum_{i=1}^{7} (O_i - E_i)^2 / E_i$ and compare with tabulated value of χ^2.

X	f (Observed frequency)	fx	Expected frequency	Expected Frequency after rounding
0	275	0	242.10	242
1	72	72	116.69	117
2	30	60	28.12	28
3	7	21	4.52	5
4	5	20	0.54 =5.12	
5	2	10	0.05	
6	1	6	0.01	
Total	**392**	**189**		

$$\overline{X} = \frac{189}{392} = 0.482$$

Note that last four classes have expected frequency less than 5. Hence to maintain the continuity of χ^2 distribution, the last four classes are pooled so that the expected frequency of the last class becomes (4.52 + 0.54 + 0.05 + 0.01) = 5.12 $\simeq$ 5. In this way, the number of classes reduced to four and d.f. will be reduced to 3.

$$\chi^2_{cal} = \sum (O_i - E_i)^2 / E_i$$

= (275 - 242)^2/242 + (72 ï 117)^2 /117 + (30 - 28)^2/28 + (15 ï 5)^2 /5

= 4.5 + 17.31 + 0.14 + 20 = 41.95

Tabulated value $\chi^2_{3,\,0.05} = 7.82$

Since $\chi^2_{cal} > \chi^2_{tab}$, we reject H_0 and conclude that the Poisson distribution did not fit well to the given data.

Example-22: The following figures shows the distribution of digits in numbers chosen at random from a telephone directory

Digit	0	1	2	3	4	5	6	7	8	9	**Total**
Frequency	180	200	190	230	210	160	250	220	210	150	**2000**

Test whether digits may be taken to occur equally frequently in the directory.

Solution: H_0: Digits occur equally frequently in the directory

H_1: Digits do not occur equally frequently in the directory i.e. fit is not good

The expected frequency (E) for each digit

0, 1, 2,·····,9 is 2000/10 = 200

We will now use χ^2 ï test of goodness of fit

$\chi^2 = \Sigma(O ï E)^2/E$

O	E	$(O ï E)^2/E$
180	200	2.0
200	200	0
190	200	0.5
230	200	4.5
210	200	0.5
160	200	8.0
250	200	12.5
220	200	2.0
200	200	0
150	200	12.5

$\chi^2_{cal} = \sum(O-E)^2/E = 42.5$

χ^2_{tab} (for k-1 = 9 df at 5% level of significance) is 16.9. Thus $\chi^2_{cal} > \chi^2_{tab}$ and thus H_0 is rejected. Thus it can be concluded that digits are not uniformly distributed in the directory.

Example-23: The following table gives the number of road accidents that occurred during the various days of the week. Find whether the accidents are uniformly distributed over the week?

Days	**Sunday**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Total**
No. of accidents	16	18	10	13	13	11	17	98

Solution: H_0 : Road accidents are uniformly distributed over the week.

H_1 : Road accidents are not uniformly distributed over the week.

Under H_0, expected frequency = 98/7 = 14

Let O and E represent the observed and expected frequencies, then

Day	**O**	**E**	**O^2**	**O^2/E**
Sunday	16	14	256	18.29
Monday	18	14	324	23.14
Tuesday	10	14	100	7.14
Wednesday	13	14	169	12.07
Thursday	13	14	169	12.07
Friday	11	14	121	8.64
Saturday	17	14	289	20.64
Total	**98**	**98**		**101.99**

$N = \Sigma O_i = \Sigma E_i = 98$

$$\chi^2_{cal} = \sum \frac{O_i^{\ 2}}{E_i} - N = 101.99 \text{ ï } 98 = 3.99$$

Table value of χ^2 at $\alpha = 0.05$ and (7-1) = 6 degree of freedom is 12.6. Since $\chi^2_{cal} < \chi^2_{tab}$ so we do not reject the null hypothesis and conclude the accidents are uniformly distributed over the week.

Example-24: In experiments on pea breeding, Mendal obtained the following frequencies of seeds : 315 round and yellow, 101 wrinkled and yellow; 108 round and green; 32 wrinkled and green, total 556. Theory predicts that the frequencies should be in the ration 9 : 3 : 3 : 1. Find χ^2 and examine correspondence between theory and experiment.

Solution: H_0: The data follow the ratio 9 : 3 : 3 : 1 (or the fit is good)

H_1: The data does not follow the ratio 9 : 3 : 3 : 1 (or the fit is not good)

Expected frequencies for group I = $\frac{9}{16} \times 556 = 313$

Expected frequencies for group II = $\frac{3}{16} \times 556 = 104$

Expected frequencies for group III = $\frac{3}{16} \times 556 = 104$

Expected frequencies for group IV = $\frac{1}{16} \times 556 = 35$

Observed and expected frequencies	**I**	**II**	**III**	**IV**	**Total**
O :	315	101	108	32	**556**
E :	313	104	104	35	**556**

$$\chi^2_{cal} = \frac{(315-313)^2}{313} + \frac{(101-104)^2}{104} + \frac{(108-104)^2}{104} + \frac{(32-35)^2}{35} = 0.51$$

The table value $\chi^2_{3,\,0.05}$ = 7.82

The calculated value is less than 7.82, hence we do not reject H_0. We conclude that there is a correspondence between the theory and experiment or the data follows the ratio 9 : 3 : 3 : 1

(ii) Test of Independence of Attributes in Contingency Tables:

Another application of the Chi-square test is in testing independence of attributes A and B in a m x n contingency table, which contains mn cell frequencies in m rows and n columns, where m and n are the categories of the attributes A and B respectively. For testing independence of row and column classifications, we define the null and alternative hypothesis as follows

H_0 : Attributes A and B are independent

H_1 : Attributes A and B are not independent

Let O_{ij} denote the observed frequency in the (i, j) cell and E_{ij} be the expected frequency under the null hypothesis.

When H_0 is true $E_{ij} = \frac{R_i \text{ x } C_j}{N}$ where R_i is the i^{th} row total C_j the j^{th} column total and N is the total frequency.

Test Statistic:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{O_{ij}^2}{E_{ij}} - N, \text{ is distributed as } \chi^2 \text{ with (m ï 1) (n ï 1) d.f.}$$

Conclusion: Reject H_0 if $\chi^2_{cal} > \chi^2_{\alpha,\,(m-1)(n-1)}$ with (m-1)(n-1) d.f. at α per cent level of significance, otherwise we do not have sufficient evidence for rejection of H_0 and hence accept H_0.

Yates' Correction:

If any cell frequency is <5, then Yatesô correction of continuity is to be applied and we get modified χ^2 as:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left[\,|\, O_{ij} - E_{ij} \,|\, - 0.5\right]^2}{E_{ij}}$$

Alternately, we can merge the nearby classes for attributes A or B or both so that no cell frequency in the modified table remains less than 5. Compute the value of χ^2 for the modified table and adjust the d.f. as per new dimensions.

Example-25: Show that the conditions at home have a bearing on the condition of the child on the basis of the following observed table:

	Condition at home		
Conditions of child	**Clean**	**Not clean**	**Total**
Clean	75	40	115
Fairly clean	35	15	50
Dirty	25	45	70
Total	**135**	**100**	**235**

Solution: H_0 : Condition of child is independent of condition at home. H_1: condition of the child depends on condition at home:

The expected frequencies are computed as follows:

E_{11} = (115x135)/235 = 66.01

E_{21} = (50x135)/235 = 28.7 and so on

Expected frequency table

	Clean	**Not clean**	**Total**
Clean	66.1	48.9	115
Fairly clean	28.7	21.3	50
Dirty	40.2	29.8	70
Total	135	100	235

$$G^2_{cal} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ji})^2}{E_{ij}}$$

$$= \frac{(75 - 66.1)^2}{66.1} + \frac{(35 - 28.7)^2}{28.7} + \frac{(25 - 40.2)^2}{40.2} + \frac{(40 - 48.9)^2}{48.9}$$

$$+ \frac{(15 - 21.3)^2}{21.3} + \frac{(45 - 29.8)^2}{29.8} = 18.95$$

The table value of χ^2 at 2 d.f. and at 5% level of significance ($G_2^2, 0.05$) = 5.99. The calculated value is more than the table value, hence the null-hypothesis is rejected. Our decision is that the condition at home has a bearing on the condition of the child.

Example-26: The data relate to the sample of married women according to their level of education and marriage adjustment score.

Level of education	**Marriage adjustment score**				
	Very low	**Low**	**High**	**Very high**	**Total**
Post Graduate	24	97	62	58	**241**
Matriculate	22	28	30	41	**121**
Illiterate	32	10	11	20	**73**
Total:	78	135	103	119	**435**

Can you say that two attributes are independent?

Solution: Here null and alternative hypotheses are

H_0 : The two attributes are independent i.e. marriage adjustment is independent of education level.

H_1 : The two attributes are associated i.e. adjustment in marriage is a function of education.

We compute the expected frequency corresponding to observed ones using formula: $E_{ij} = \frac{R_i \text{ x } C_j}{N}$ which is given as follows:

$$E_{11} = \frac{78 \text{ x } 241}{435} = 43.2 \quad E_{12} = \frac{135 \text{ x } 241}{435} = 74.8 \quad E_{13} = \frac{103 \text{ x } 241}{435} = 57.1 \quad E_{14} = \frac{119 \text{ x } 241}{435} = 65.9$$

$$E_{21} = \frac{78 \text{ x } 121}{435} = 21.7 \quad E_{22} = \frac{135 \text{ x } 121}{435} = 37.6 \quad E_{23} = \frac{103 \text{ x } 121}{435} = 28.7 \quad E_{24} = \frac{119 \text{ x } 121}{435} = 33.1$$

$$E_{31} = \frac{78 \text{ x } 73}{435} = 13.1 \quad E_{32} = \frac{135 \text{ x } 73}{435} = 22.7 \quad E_{33} = \frac{103 \text{ x } 73}{435} = 17.3 \quad E_{34} = \frac{119 \text{ x } 73}{435} = 20.1$$

Compute $\chi^2_{cal} = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{(O_{ij} - E_j)^2}{E_{ij}} = 57.26$

also χ^2_{tab} (at (4-1) (3-1) = 6 d.f.) = 12.59

Since calculated value of χ^2 is greater than tabulated value, hence we reject H_0 and say that the two attributes i.e., level of education and marriage adjustment score are related to each other. That is, higher the level of education, greater is the adjustment in marriage.

Fisher Exact Test for 2 x 2 Contingency Table:

If two attributes are divided into only two classes to form a 2 x 2 contingency table

Attribute A	**Attribute B**		**Total**
	B_1	**B_2**	
A1	a	b	a + b
A2	c	d	c + d
Total	**a + c**	**b + d**	**N = a + b + c + d**

In this case, the value of χ^2 can be calculated directly from the observed frequencies by the formula:

$$G^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{N(ad-bc)^2}{R_1R_2C_1C_2}, \text{ where } N = (a+b+c+d)$$

which follows chi-square distribution with one degree of freedom

Yates Correction for 2 x 2 Contingency Table:

If some of the cell frequencies in 2 x 2 contingency table are less than 5, the continuity of χ^2 distribution is not maintained. So, Yatesô correction should be used to remove this discrepancy. For applying Chi-square test Yates suggested that add 0.5 in the frequency which are less than 5 and add or 0.5 to the remaining cell frequencies in such a way that the marginal totals remain the same. Specially for 2 x 2 contingency table the value of χ^2 under Yatesô correction can be obtained from the formula $G^2 = \frac{N\left[\,|\,ad-bc\,|\,-N/2\right]^2}{(a+b)(c+d)(a+c)(b+d)}$ which follows χ^2ï distribution with 1 d.f.

Example-27: The following data relate to the height of fathers and their first sons at the age of 35 years.

		Height of Fathers		**Total**
Height of Sons	Tall	8	2	10
	Short	7	6	13
	Total	15	8	23

Test whether the height of sons is independent of the height of the fathers.

Solution: Since one of the cell frequency is < 5, therefore, the value of χ^2 is calculated using the formula of Yatesôcorrection i.e.

$$\chi^2 = \frac{N\left[\,|\,ad - bc\,| - N/2\right]^2}{(a+b)\,(c+d)\,(a+c)\,(b+d)} = \frac{23\left[\,|\,48 - 14\,| - 11.5\right]^2}{10 \times 13 \times 15 \times 8} = 0.746$$

and $\chi^2_{0.05}$ at 1 df = 3.84

Since the calculated value of χ^2 is less than tabulated value of χ^2, we do not reject H_0 and conclude that the height of sons is independent of the height of their fathers.

Example-28: From the following data test at 5% level of significance if literacy depends upon the region.

Education	**Region**		**Total**
	Rural	**Urban**	
Literature	10	46	56
Illiterate	40	4	44
Total	**50**	**50**	**100**

Solution: H_0 : Literacy is independent of region.

H_1 : Literacy is not independent of region.

Since frequency in one cell is less than 5, so using Yatesô Corrected Formula for chi-square

$$\chi^2_{cal} = \frac{\left[\left|ad - bc)\right| - \frac{N}{2}\right]^2 N}{R_1 R_2 C_1 C_2} = \frac{\left[\left|10 \times 4 - 40 \times 46\right| - \frac{100}{2}\right]^2 100}{50 \times 50 \times 56 \times 44}$$

$$\chi^2_{cal} = \frac{\left[\left|40 - 1840)\right| - 50\right]^2}{25 \times 56 \times 44} = \frac{(1750)^2}{25 \times 56 \times 44} = \frac{3062500}{61600} = 49.72$$

Table χ^2_{tab} at 1 df at 5% level of significance is 3.84. Since $\chi^2_{cal} > \chi^2_{tab}$, therefore we reject the null hypothesis and conclude that literacy is dependent on region.

Example-29: From the following results regarding eye colour of mother and son, test at Ŭ = 0.05 if the colour of sonôs eyes is associated with that of mother?

Mother's eye colour	**Son's eye colour**		
	Light blue	**Not light blue**	**Total**
Light blue	47	16	63
Not light blue	4	33	37
Total	**51**	**49**	**100**

Solution:

H_0: Colours of motherôs and sonôs are independent i.e. not associated

H_1: Colours of motherôs and sonôs are not independent (i.e. associated)

Since one cell frequency is less than 5, therefore, applying Yatesô correction for 2 x 2 contingency table, we get:

$$\chi^2_{cal} = \frac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{R_1R_2C_1C_2} = \frac{100(|1551-64|-50)^2}{63x37x51x49} = 35.45$$

Since $\chi^2_{cal} > \chi^2_{0.05,1} = 3.84$, therefore, we reject H_0 and conclude that the attributes are not independent i.e. the colour of sonôs eye depend on the colour of motherôs eye.

Chi-square test for the Population Variance:

This test is used to test the null hypothesis that whether the sample has been drawn from population with the specified variance σ_0^2.

Let $x_1, x_2,$ é x_n be n independent observations from N (ε, σ^2)

1. Formulate H_0 and H_1 as:

$$H_0 : \sigma^2 = \sigma_0^2 \quad Vs \quad H_1 : \sigma^2 > \sigma_0^2$$

2. Choose Ǔ

3. Compute χ^2 statistic

If H_0 is true then:

$$\chi^2 = \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2} \text{ which follows } \chi^2 \text{ with n -1 d.f.}$$

where, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is the sample variance

4. Hence if $\chi^2_{cal} \geq \chi^2_{\alpha,n-1}$, reject H_0 and not otherwise; where $\chi^2_{\alpha,n-1}$ = tabulated value of χ^2 at 5% level of significance with n ï 1 d.f.

Approximation of χ^2-distribution for large sample size (n > 30):

If the sample size n is large (>30), then we can use Fisherôs approximation i.e. $\sqrt{2\chi^2}$ follows normal distribution with mean $\sqrt{2n-1}$ and variance 1.

i.e. $\sqrt{2\chi^2} \sim N(\sqrt{2n-1}, 1)$

$$\text{Thus } Z = \sqrt{2\chi^2} - (\sqrt{2n-1}) \sim N(0,1)$$

And usual Z (standard normal) test can be applied.

Example-30: The variability in the yield of a crop variety by the conventional method (measures in terms of standard deviation) for a random sample of size 30 was 2.8 qha^{-1}. Can it be concluded at $\alpha = 0.05$ that it is not more than that of standard method which is believed to be equal to 2.2 q ha^{-1}.

Suppose:

1. $H_0 : \sigma = 2.2$ q ha^{-1}
2. $H_1 : \sigma > 2.2$ q ha^{-1}
3. $\alpha = 0.05$
4. Test statistic: Here s = 2.8 $\sigma_0 = 2.2$ n = 30

$$\chi^2_{cal} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{29(2.8)^2}{(2.2)^2} = \frac{227.36}{4.84} = 46.98$$

Since $\chi^2_{cal} > \chi^2_{tab}$ (at $\alpha = 0.05$) for 29 d.f. = 42.56, therefore H_0 is rejected. Thus, it is concluded that variability in yields by conventional method is more than that of standard method.

By using Fisherôs approximation for large sample size:

Test Statistic:

$$Z_{cal} = \sqrt{2\chi^2} - (\sqrt{2n-1})$$

$$= \sqrt{2 \text{ x } 46.98} - (\sqrt{60-1})$$

= 9.693 ï 7.681 = 2.012

Since $Z_{cal} > Z_{tab} = 1.645$, therefore H_0 is rejected.

Snedcor's F-test: It is used

i) As a test for the equality of two population variances i.e. whether the two samples may be regarded as drawn from the normal populations having the same variance.

ii) As a test for the equality of several population means.

Testing the equality of two population variances:

The F-test may be used to test the equality of two population variances. Let x_1, x_2, é ..xn_1 and y_1, y_2,·····,yn_2 be two independent samples drawn randomly from two normal

populations with variances σ_1^2 and σ_2^2. Let s_1^2 and s_2^2 be the estimates of population variances. We want to test the null hypothesis that the population variances are equal.

Assumptions:

i) Populations are normal.

ii) Samples are drawn independently and at random

Stepwise testing procedure is as follows

1. $H_0: \sigma_1^2 = \sigma_2^2$

2. $H_1: \sigma_1^2 > \sigma^2_2$

3. Choose level of significant $\alpha = 0.05$ or 0.01

4. Test statistic $F_{cal} = \frac{s_x^2}{s_y^2}$ where $s_x^2 = \frac{1}{n_1 - 1}\sum\left(x_i - \bar{x}\right)^2$ $s_y^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2$ are unbiased estimators of σ_1^2 and σ_2^2 respectively. Here numerator corresponds to greater variance

If F_{cal}>F-tabulated at ($v_1 = n_1$-1, $v_2 = n_2$-1) d.f. and at α level of significance, then we reject H_0 and conclude that the population variances are significantly different, otherwise we accept H_0.

Example-31: Test the assumption for equality of two population variances in the example 15 on two sample t-test.

Solution: $n_1 = 10$ $\bar{x} = 12$ $s_1^2 = \frac{1}{n_1 - 1}\sum(x_i - \bar{x})^2 = 120/9 = 13.33$

$n_2 = 12$ $\bar{y} = 15$ $s_2^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2 = 314/11 = 28.55$

$H_0: \sigma_1^2 = \sigma_2^2$; $H_1: \sigma_2^2 > \sigma_1^2$ (since the larger variance is in numerator)

$$F_{cal} = \frac{s_2^2}{s_1^2} = \frac{28.55}{13.33} = 2.14$$

Since $F_{cal} < F_{(11, 9)} = 3.59$ at $\alpha = 0.05$, therefore, we do not reject H_0 and conclude that population variances are equal. Thus, the usual two sample t-test can be applied.

Example-32: The life times for random samples of batteries (type A and B) were recorded and the following results were obtained.

Type of Battery	No. of Batteries	Mean life (hours)	Sum of squares of deviations from mean
A	10	500	1800
B	12	555	2160

Test if there is any significant difference between the life times of two types of batteries.

Solution: Equality of means will be tested by applying two sample t-test where we assume that $\sigma_1^2 = \sigma_2^2$, therefore, we first apply F-test for the equality of two population variances.

1. $H_0 : \sigma_1^2 = \sigma_2^2$
2. $H_1 : \sigma_1^2 > \sigma_2^2$
3. $\alpha = 0.05$ or 0.01
4. Test statistic $F_{cal} = \frac{s_1^2}{s_2^2}$

$n_1 = 10,\ \bar{x} = 500 \qquad \Sigma (x_i - \bar{x})^2 = 1800 \qquad s_1^2 = \frac{1800}{9} = 200$

$n_2 = 12,\ \bar{y} = 555 \qquad \Sigma (y_i - \bar{y})^2 = 2160 \qquad s_1^2 = \frac{2160}{11} = 196.3$

$$Fcal = \frac{200}{196.3} = 1.02$$

Tabulated $F_{(9, 11)}$ at $\alpha = 0.05$ is equal to 3.92

Since $F_{cal} < F_{tab}$, therefore, we do not reject H_0 and conclude that the population variances are not significantly different.

Two sample t-test: After testing the null hypothesis of equality of population variances, we now apply the t-test.

1. $H_0 : \mu_1 = \mu_2$
2. $H_0 : \mu_1 \neq \mu_2$
3. $\alpha = 0.05$ or 0.01
4. Test Statistic

$$t_{cal} = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_i - \bar{x})^2 + (y_i - \bar{y})^2\right] = \frac{1}{20}[1800 + 2160] = 198$$

$$t_{cal} = \frac{500 - 555}{\sqrt{198}} \sqrt{\frac{10 \times 12}{10 + 12}} = -21.32 \qquad \Rightarrow |t_{cal}| = 21.32$$

Conclusion: Since $|t_{cal}| > t_{0.025,\ 20} = 2.08$ therefore we reject H_0 and conclude that there is significant difference in the life times of two types of batteries.

Example-33: In a test given to two groups of students, the marks obtained are as follows:

First Group	18	20	36	50	49	36	34	49	41
Second Group	36	31	29	35	37	30	40		

Examine the significance of difference in the mean marks secured by students of two groups.

Solution:

$H_0 : \mu_1 = \mu_2$ There is no significant difference between the mean marks

$H_1 : \mu_1 \neq \mu_2$

$\alpha = 0.05$

Calculation of $\bar{X}, \bar{Y}$ and s_1, s_2

First group X	**$(X - \bar{X})$ = X – 37**	**$(X - \bar{X})^2$**	**Second Group X**	**$(Y - \bar{Y})$ = Y – 34**	**$(Y - \bar{Y})^2$**
18	-19	361	36	2	4
20	-17	289	31	-3	9
36	-1	1	29	-5	25
50	13	169	35	1	1
49	12	144	37	3	9
36	-1	1	30	-4	16
34	-3	9	40	6	36
49	12	144			
41	4	16			
333	0	1134	238	0	100

$$\overline{X} = \frac{333}{9} = 37 \qquad \overline{Y} = \frac{238}{7} = 34$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum (X_i - \overline{X})^2 = \frac{1134}{8} = 141.75 \qquad s_2^2 = \frac{1}{n_2 - 1}\sum (Y_i - \overline{Y})^2 = \frac{100}{6} = 16.66$$

For testing the equality of population variances, use F-test:

$$F_{cal} = \frac{s_1^2}{s_2^2} = \frac{141.75}{16.66} = 8.51$$

F_{tab} for (8, 6) d.f. at Ŭ= 0.05 = 3.58

Here $F_{cal} > F_{tab}$, therefore, we conclude that population variances are significantly different.

Remark: Since population variances $\mathring{u}_1^2$ and $\mathring{u}_2^2$ are unknown and differ significantly, therefore, it falls under case (iv).

Test Statistic: $$t_{cal} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{37 - 34}{\sqrt{\frac{141.75}{9} + \frac{16.66}{7}}} = \frac{3}{\sqrt{18.13}} = 0.704$$

$$\text{d.f.} = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_1^2/n_1\right)^2}{n_2 - 1}} = \frac{(18.13)^2}{31.01 + 0.94} = \frac{328.70}{31.95} = 10 \text{ (after rounding off)}$$

t_{tab} for 10 d.f. at Ŭ= 0.05 (Two tailed test) = 2.228

Here $|t_{cal}| < t_{tab}$, thus H_0 is not rejected and it is concluded that mean marks secured by two groups do not differ significantly.

Testing the equality of several Population Means:

The F-test can also be used to test the equality of several population means in the analysis of variance technique. The F-test has wider application as it provides an overall test for the equality of several population means where as t-test may be used to test the equality of only two population means.

Here we shall discuss ANOVA for i) Completely Randomized Design ii) Randomized Block Design. The null hypotheses we want to test here

H_0 : All treatment effects (means) are equal

H_1 : Atleast two treatments means differ

Completely Randomized Design (CRD):

It is a simplest design that uses only two basic principles of experimental design. In this design, total experimental area is divided into a number of experimental units and the treatments are allotted to units entirely at random. CRD is used when there is homogenous experimental material e.g. in agriculture field experiments when all the plots have same soil fertility, soil texture and uniform agronomical practices and in animal science experiments when the animals are of same breed, grown in similar conditions etc. CRD is used in laboratory, pot and green house experiments etc.

It has easy layout, its analysis is the simplest one, it has flexibility with respect to number of treatments and number of replications, it allows maximum number of degree of freedom in error. In CRD missing data can be handled easily. It is suited for only small number of treatments because large number of treatments needs large material in which variation increases, so we opt for other designs.

Layout:

The layout for CRD is the simplest one. The whole experimental area is divided into a no. of units $N = \sum_{i=1}^{t} r_i$ and all the treatments are allotted randomly to all the units. Assign t treatments randomly to N units such that T_i is allotted to r_i units. Suppose we have 5 treatments T_1, T_2, T_3, T_4 and T_5 with replication 4, 3, 4, 4 and 5 respectively. The whole experimental area is to be divided into

$\sum r_i = 4 + 3 + 4 + 4 + 5 = 20$ plots

T_1	T_4	T_2	T_5	T_3
T_3	T_5	T_1	T_3	T_4
T_2	T_5	T_3	T_2	T_1
T_1	T_4	T_5	T_4	T_5

Model for CRD

Let Y_{ij} is jth unit in ith treatment.

$Y_{ij} = \mu + t_i + e_{ij}$

where μ = General Mean

$t_i = i^{th}$ treatment effect

e_{ij} = random error ~ $N(0, \sigma^2)$

For the analysis purpose the data is written systematically

	Treatments					
Replications	1	2	é é é é é .	i	é é é .	t
	Y_{11}	Y_{21}	é é é é ..	Y_{i1}	é é é	Y_{t1}
	Y_{12}	Y_{22}	é é é é ..	Y_{i2}	é é é	Y_{t2}
	.					
	.					
	Y_{1r_1}	Y_{1r_2}	é é é é ..	Y_{ir_i}	é é é	Y_{tr_t}
Total	$\mathbf{T_1}$	$\mathbf{T_2}$	…….	$\mathbf{T_i}$	……..	$\mathbf{T_t}$
Mean	$\bar{Y}_1$	$\bar{Y}_2$	…….	$\bar{Y}_3$	……	$\bar{Y}_t$

$N = \sum_{i=1}^{t} r_i$ Let $G = \sum_i T_i = \sum_i \sum_j y_{ij}$ be the grand total.

i) Correction factor = G^2/N

ii) Total sum of squares = $\sum\sum Y_{ij}^2$ ï CF

iii) Treatment sum of squares = $\sum \frac{T_i^2}{r_i} - CF$

iv) Error sum of squares = Total SS ï treatments SS

ANOVA

Source	d.f.	SS	M.S.S.	F_{cal}
Treatments	t-1	$\sum \frac{T_i^2}{r_i} - CF = SS_T$	$SS_T/(t-1) = T$	T/E
Error	N-t	Total SS- Treat. SS = SS_E	$SS_E/(N-t) = E$	
Total	N-1	$\sum\sum Y_{ij}^2$ ï CF = Total SS		

It the F_{cal} value is greater than F_{tab} at (t-1, N-t) d.f. and at given α, we conclude that the treatments are significantly different.

SE (mean of T_i) = $\sqrt{\frac{E}{r_i}}$

SE (d) = $\sqrt{E\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$ where r_i and r_j are replications of T_i and T_j. If $r_i = r_j = r$

SE (d) = $\sqrt{\frac{2E}{r}}$ we go for CD only when the treatment effects are significant

CD = SE x $t_{\text{error d.f. i .e. (N-t) d.f.}}$

If $\left|\overline{Y}_i - \overline{Y}_j\right| \geq$ C.D. we say T_i is significant different from T_j.

Example-34: Given below are the weight gain of baby chicks (gms) under 4 different feeds, analyze the data using CRD.

	Treatments				
Observations	**t_1**	**t_2**	**t_3**	**t_4**	
1	55	61	42	169	
2	49	112	97	137	
3	42	30	91	169	
4	21	89	95	85	
5	52	63	92	154	
Total	**219**	**355**	**407**	**714**	**1695**
Mean	**43.8**	**71**	**81.4**	**142.8**	

CF = $\frac{G^2}{N} = \frac{(1695)^2}{20} = 143651$ General Mean (GM) = $\frac{1695}{20} = 84.75$

Total SS = $\sum\sum Y_{ij}^2$ ï CF = $55^2 + 61^2 +$ é é .. 154^2 ï CF = 37794

SS due to treatments = $\frac{(219)^2 + (355)^2 + (407)^2 + (714)^2}{5} - \text{CF} = 26235.2$

Error SS = Total SS ï treat SS = 37794.0 ï 26235.2 = 11558.8

ANOVA

S.V.	d.f.	SS	MSS	F_{cal}
Treatment	3	26235.2	8745.1	12.1*
Error	16	11558.8	722.4	
Total	19	37794.0		

$F_{3,16(0.05)} = 3.25$

Since $F_{cal} \geq F_{tab}$ so treatment effects are significantly different

CD 5% = $\sqrt{\frac{2E}{r}}$ x $t_{16,\,0.025}$ = $\sqrt{\frac{2x722.4}{4}}$ x 2.12 = 40.3

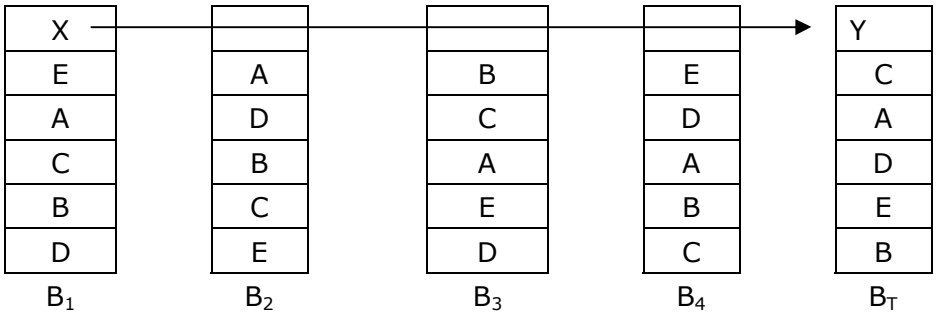
C.V. (Coefficient of Variation) = $\frac{\sqrt{E}}{G.M.}$ x 100 = $\frac{\sqrt{722}}{84.75}$ x 100 = 31.7%

Conclusion: t_1, t_2 and t_3 do not differ significantly in respect of weight gain but these treatments differ significantly from treatment t_4.

Randomized Block Design (R.B.D.)

If the whole experimental area is not homogeneous and the fertility gradient is in one direction only, then it is possible to divide the whole area into homogeneous blocks perpendicular to the direction of fertility gradient. The treatments are randomly allocated separately to each of these blocks, and the result is a randomized block design.

Layout of R.B.D.: The plots within block must be homogeneous as far as possible. Thus, if the direction called fertility gradient in which fertility changes are maximum is known, we proceed as follows.



Suppose the fertility of the field might be having a slope from X to Y, it would be advantageous to place the blocks one after another along the gradient XY

1. Whole experimental material is divided into blocks or groups such that each treatment occurs once and only once in each block. The number of blocks to be formed is equal to the number of replications.
2. Each of these groups is further divided into a number of experimental units (plots). The number of plots within a block should be equal to the number of treatments.
3. The number of treatments within each block are applied by a random procedure.

Why Randomized Block Design is used:

1. Sensitiveness: This design removes the variation between the blocks and from that within blocks which generally results in decrease of experimental error and thus sensitivity is increased. Cochran has shown that experimental error of a R.B.D. is 60 per cent of a C.R.D.

2. Flexibility: This design allows any number of treatments and replications and the only restriction is that number of replications is equal to the number of blocks.

3. Ease of analysis: The statistical analysis is easy even in the case of missing values.

Demerits:

i) It cannot control the variation in the experimental material from two sources and in such cases is not an efficient design.

ii) If the number of treatments is large then size of blocks will increase and thus heterogeneity within the blocks will increase.

Analysis:

For analysis we use the linear additive model

$Y_{ij} = \mu + t_i + b_j + e_{ij}$ where Y_{ij} is the value of the unit for the i^{th} treatment in the j^{th} block (i= 1, 2, é .., t; j = 1, 2,é ., r)

μ is the general mean effect, t_i is the effect due to i^{th} treatment, b_j is the effect due to j^{th} block, e_{ij} is random error which is assumed to be independently and normally distributed with mean zero and variance σ_e^2.

Let there be t treatments, each treatment being replicated r times (equal to number of blocks)

Let $T_i = \sum_j Y_{ij}$; $R_j = \sum_i Y_{ij}$

Treatments/Blocks	1	2	r	Totals
1	Y_{11}	Y_{12} - - - - - - - - - -	Y_{2r}	T_1
2 . . . t	Y_{21} Y_{t1}	Y_{22} - - - - - - - - - - Y_{t2} - - - - - - - - - -	Y_{22} Y_{tr}	T_2 T_t
	R_1	R_2 - - - - - - - - - -	R_r	G

$$\text{C.F.} = \frac{(GT)^2}{N} = \frac{G^2}{rt}$$

$$\text{Total S.S.} = \sum_i \sum_j Y_{ij}^2 - \text{C.F.} = S \text{ (Say)}$$

Sum of squares due to treatments = $\sum_i \frac{T_i^2}{r} - C.F. = S_1$

Sum of square due to blocks = $\sum_j \frac{R_j^2}{t} - C.F. = S_2$

S.S. due to error = Total S.S. ï S.S. due to treatments ï S.S. due to blocks

ANOVA

Source	d.f.	SS	MSS	F_{cal}
Blocks	(r-1)	S_1	B	$F_{(r-1)(r-1)(t-1)} = B/E$
Treatment	(t-1)	S_2	T	$F_{(t-1)(r-1)(t-1)} = T/E$
Error	(r-1)(t-1)	S_3	E	-
Total	rt-1	S		

$H_0 : \mu_1 = \mu_2$ é .. = μ_t

Against alternative H_1 that treatment means are not equal. If F_{cal} (treatments) come out to be significant at a specified Ŭ, then we compare the treatment means with the C.D.

$SE_m = \sqrt{\frac{E}{r}}$

$SE_d = \sqrt{\frac{2E}{r}}$

C.D. at Ŭ= 0.005 = SE(d) x t value at error d.f. at 5%; C.D. 1% = SE(d) x t value at error d.f. at 1%

Example-35: Five varieties of cotton A, B, C, D and E were tried in RBD with five replications and following yields were obtained

	B	E	C	A	D
Block-1	6.87	4.82	7.87	5.94	9.60
	E	D	B	C	A
Block-2	I6.66	8.46	8.91	6.69	6.84
	C	A	D	B	E
Block-3	6.65	8.02	6.78	8.44	5.32
	A	C	E	D	B
Block-4	6.65	8.02	6.78	8.44	5.32
	D	B	A	E	C
Block-5	5.27	8.95	6.12	4.46	5.98

Analyze the data and draw the conclusions.

Solution:

Replications / Treatments	R_1	R_2	R_3	R_4	R_5	Total	Mean
A	5.94	6.84	8.02	7.24	6.12	34.16	6.83
B	6.87	8.91	8.44	7.59	8.98	40.79	8.16
C	7.87	6.69	6.65	7.98	5.98	35.17	7.03
D	9.60	8.46	6.78	7.50	5.27	37.61	7.52
E	4.82	6.66	5.32	5.79	4.46	27.05	5.41
Total:	**35.20**	**37.56**	**35.21**	**36.10**	**30.81**	**174.88**	

G = 174.88, r = 5 and t = 5

C.F. = $(174.88)^2/25$ = 1223.06

Total sum of squares = $5.94^2 + 6.84^2 +$ é é é é é é é é é . 4.46^2 ï C.F. = 42.072

$$\text{Block S.S.} = \frac{R_1^2 + R_5^2 + \ldots\ldots\ldots + R_5^2}{5} - \text{C.F.}$$

$$= \frac{35.20^2 + 37.66^2 + \ldots\ldots\ldots + 30.81^2}{5} - 1223.06 = 4.135$$

$$\text{Treat S.S.} = \frac{T_1^2 + T_5^2 + \ldots\ldots\ldots + T_5^2}{5} - \text{C.F.}$$

$$= \frac{34.16^2 + 40.79^2 + \ldots\ldots\ldots + 27.05^2}{5} - 1223.06 = 21.547$$

Error S.S. = Total S.S. ï Block S.S. ï Treat S.S. = 16.380

ANOVA

Source	d.f.	SS	MSS	F_{cal}
Blocks	4	4.135		
Treatment	4	21.457	5.387	4.90
Error	16	16.380	1.026	
Total	**24**	**42.072**		

SE (m) = $\sqrt{\frac{1.026}{5}}$ = 0.453; SE (d) = $\sqrt{\frac{2\,x\,1.026}{5}}$ = 0.640

$CD_{5\%}$ = SE(d) x t_{16} at Ŭ= 0.05 = 0.640 x 2.120 = 1.357

$CD_{5\%}$ = SE(d) x t_{16} at Ŭ= 0.01 = 0.640 x 2.721 = 1.870

Bartlett's Test of Homogeneity of Variances:

Sometimes the question arises whether the two or more variances obtained from different samples differ significantly from one another or not. In case of two variances, the answer can be obtained by the F test. But in case of more than two variances, Bartlettôs test of homogeneity is adequate.

Let the number of samples be k and their variances are s_1^2, s_2^2,é .,s_k^2 and their corresponding d.f. is $ɜ_1, ɜ_2, ɜ_3$,é .., $ɜ_k$ ($ɜ_i = n_i$ ï 1) where n_i is the size of i^{th} sample.

Now the following steps are needed for the test:

i) Calculate the total degrees of freedom

$N = ɜ_1 + ɜ_2 + ɜ_3$ é .. + $ɜ_k$

ii) Calculate $\bar{s}^2$, weighted average of the variances:

$$\bar{s}^2 = \frac{ɜ_1 s_1^2 + ɜ_2 s_2^2 + \ldots\ldots + ɜ_k s_k^2}{ɜ_1 + ɜ_2 + \ldots\ldots + ɜ_k} = \frac{\sum_{i=1}^{k} \nu_i s_i^2}{N} \qquad N = \sum_{i=1}^{k} \nu_i$$

iii) Calculate $ɢ^{2'}$ and C, the correction factor:

$$ɢ^{2'} = N \log_e \bar{s}^2 - \sum_{i=1}^{k} ɜ_i \log_e s_i^2$$

$$\text{and } C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\left(\frac{1}{\nu_1}\right) - \frac{1}{N}\right]$$

iv) Test Statistic

Calculate $ɢ^2 = ɢ^{2N}/C$ which follows $ɢ^2$ distribution with (k-1) d.f.

Conclusion:

If $ɢ^2_{cal} \geq ɢ^2_{tab}$ with (k-1) d.f. at a specified Ŭ, then we reject H_0 and conclude that population variances are not homogeneous.

Example-36: The sample variances 1.27, 2.58 and 3.75 based on 9, 13 and 12 degrees of freedom are obtained from three different samples. Apply Bartlettôs test for testing the homogeneity of three population variances.

Sample	**Sample Variances s^2**	**Degree of Freedom ν**	$\frac{1}{\nu}$	**νs^2**	**$\log_e s^2 = 2.3026 \log_{10} s^2$**	**$\nu \log_e s^2$**
1	1.27	9	0.11111	11.43	0.2390	2.1510
2	2.58	13	0.07692	33.54	0.9478	12.3214
3	3.75	12	0.08333	45.00	1.3218	15.8616
Total	**7.60**	**N=34**	**0.27136**	**89.97**	**2.5086**	**30.3340**

Here $\bar{s}^2 = \frac{\sum \nu_i s_i^2}{N} = \frac{89.97}{34} = 2.646$ $\qquad N = \sum_{i=1}^{k} \nu_1$

And $\chi^{2Nj} = N \log_e s^{-2} - \sum \nu_i \log_e s_i^2$

$= 34 \log_e 2.646 - 30.3340 = 2.7514$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum\left(\frac{1}{\nu_1}\right) - \frac{1}{N}\right] = 1 + \frac{1}{3 \times 2}[0.27136 - 0.02941] = 1.04033$$

$$\therefore \text{Corrected } \chi^2_{cal} = \frac{\chi'^2}{C} = \frac{2.7514}{1.04033} = 2.645$$

χ^2_{tab} (with k - 1) = 2 df at Ŭ= 0.05 = 5.991

Since the $\chi^2_{cal} < \chi^2_{tab}$, hence we conclude that the population variances are homogenous.

Example-37: Certain gram variety tested on 64 plots gave an average yield as 985 kg/ha, and variance 1600 kg^2/ha. Test at 5% level of significance that the experiment agreed with the breeders claim that the average yield of the variety is 1000 kg/ha. Also construct 95% confidence interval for population mean.

Solution: Here n = 64, $\bar{X}$ = 985 kg/ha and s^2 = 1600 kg^2/ha or s = 40 kg/ha

$H_0 : \mu_0 = 1000$ kg/ha

$H_1 : \mu_0 \neq 1000$ kg/ha

Level of significance $\alpha = 0.05$

Population variance is unknown and sample is large so, Z-test is used

$$Z_{cal} = \frac{\overline{X} - \varepsilon_0}{s/\sqrt{n}} = \frac{985-1000}{40/\sqrt{64}} = \frac{985-1000}{40/8} = \frac{985-1000}{5}$$

$$= \frac{-15}{5} = -3 \text{ or } \quad |Z_{cal}| = 3.0$$

because $|Z_{cal}| > 1.96$ so, we reject the null hypothesis at 5% level of significance. Hence it can be concluded that experiment does not confirm breederôs claim that average yield of variety is 1000 kg/ha.

95% confidence interval for mean (μ) = $\overline{x} \pm z_{\ddot{y}/2} \frac{s}{\sqrt{n}} = 985 \pm \frac{40}{\sqrt{64}} \times 1.96$

$= 985 \pm 5 \text{ x } 1.96 = (975.2, 994.8)$

Summary Table for Various Tests of Hypotheses

H_o	Test Statistic	H_1	Critical Region
$\mu = \mu_o$	$Z = \dfrac{\overline{X} - \varepsilon_0}{\sigma/\sqrt{n}}$; σ known	$\mu \neq \mu_o$ $\mu > \mu_o$ $\mu < \mu_o$	$\lvert Z\rvert \geq z_{\alpha/2}$ $Z > z_{\alpha}$ $Z < -z_{\alpha}$
$\mu = \mu_o$	$t = \dfrac{\overline{X} - \varepsilon_0}{s/\sqrt{n}}$; $\nu = n - 1$, σ known	$\mu \neq \mu_o$ $\mu > \mu_o$ $\mu < \mu_o$	$\lvert t\rvert \geq t_{\alpha/2,\, n-1}$ $t > t_{\alpha,\, n-1}$ $t < -t_{\alpha,\, n-1}$
$\mu_1 - \mu_2 = d_o$	$Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - d_0}{\sqrt{(\sigma_1^2 n_1)} + \sqrt{(\sigma_2^2 n_2)}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 \neq d_o$ $\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$	$\lvert Z\rvert \geq z_{\alpha/2}$ $Z > z_{\alpha}$ $Z < -z_{\alpha}$
$\mu_1 - \mu_2 = d_o$	$t = \dfrac{(\overline{X}_1 - \overline{X}_2) - d_0}{s_p\sqrt{(1/n_1)} + \sqrt{(1/n_2)}}$; $\nu = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$;	$\mu_1 - \mu_2 \neq d_o$ $\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$	$\lvert t\rvert \geq t_{\alpha/2,\, n_1+n_2-2}$ $t < -t_{\alpha} n_1+n_2-2$ $t > t_{\alpha,\, n_1+n_2-2}$
$\mu_1 - \mu_2 = d_o$	$t' = \dfrac{(\overline{X}_1 - \overline{X}_2) - d_0}{\sqrt{(s_1^2 n_1)} + \sqrt{(s_2^2 n_2)}}$; $\nu = \dfrac{(s_1^2/n_1 + (s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$; $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 \neq d_o$ $\mu_1 - \mu_2 < d_o$ $\mu_1 - \mu_2 > d_o$	$\lvert t\rvert \geq t_{\alpha/2}$ $t > t_{\alpha}$ $t < -t_{\alpha}$
$\mu_D = d_o$	$t = \dfrac{\overline{D} - d_0}{sd/\sqrt{n}}$; $\nu = n - 1$, paired observations	$\mu_D \neq d_o$ $\mu_D > d_o$ $\mu_D < d_o$	$\lvert t\rvert \geq t_{\alpha/2,\, n-1}$ $t > t_{\alpha,\, n-1}$ $t < -t_{\alpha,\, n-1}$
$\sigma^2 = \sigma_0^2$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$; $\nu = n - 1$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi_{\alpha}^2$
$\sigma_1^2 = \sigma_2^2$	$F = \dfrac{s_1^2}{s_2^2}$; $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$	$\sigma_1^2 > \sigma_2^2$	$F > F_{\alpha(\nu_1, \nu_2)}$

EXERCISES

1. A sample of 400 male adultôs from Haryana is found to have a mean height of 171.38 cms. Can it be reasonably regarded as a sample from a large population of mean height 171.17 cms and standard deviation of 3.30 cms? [Hint: use one sample Z-test]

2. Ten specimens of copper wires drawn from a large lot have the breaking strengths (in kg. wt) equal to 578, 572, 570, 568, 512, 578, 570, 572, 569, 548. Test whether the mean breaking strength of the lot may be taken to be 578 kg wt. [Hint: use one sample t-test]

3. A manufacturer claimed that atleast 90% of the tools which he supplied were upto the standard quality. A random sample of 200 tools showed that only 164 were upto the standard. Test his claim at 1% level of significance. [Hint: use Z-test for single proportion]

4. You are working as a purchase manager for a company. The following information has been supplied to you by two manufacturers of electric bulbs

	Company A	**Company B**
Mean life (in hours)	1300	1250
Standard deviation (in hours)	82	93
Sample size	100	80

Is brand A of bulbs is superior in respect to higher mean life at a risk of 5%.

[Hint: Since sample size are large, therefore, use two samples Z-test].

5. A company is interested to know if there is any difference in the average salary received by the managers of two divisions. Accordingly samples of 12 managers in the first division and 10 in the second division were selected at random and results are given below:

	First Division	**Second Division**
Sample size	12	10
Average monthly salary (Rs.)	25000	22400
Standard deviation (Rs)	640	960

Apply two sample t-test to find out whether there is a significant difference in the average salary.

6. Given below is the contingency table for production in three shifts the number of defective goods turn over. Use chi-square test to test whether the number of defective goods depends on the shift run by the factory.

No. of Defective goods			
Shifts	**1st week**	**2nd week**	**3rd week**
1	15	5	20
2	20	10	20
3	25	15	

7. Ten individuals are chosen at random from a population and their heights are found to be in inches,

   64, 65, 65, 66, 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 (Hint: One sample t-test).

8. Two independent random samples were taken from two populations:

Sample I:	12	14	10	8	16	5	3	9	11	
Sample II:	21	18	14	20	11	19	8	12	13	15

   Assuming a normal distribution for the population, test significance of difference between the population means (Hint: Two sample t-test)

9. 10 women were given an injection to induce blood pressure. Their blood pressures before and after the injection were as follows:

S. No.	**Before Injection**	**After Injection**
1	70	87
2	86	93
3	84	94
4	88	90
5	96	95
6	70	72
7	99	102
8	94	97
9	72	89
10	98	101

(a) Do you think mean blood pressure before injection is the same as mean blood pressure after injection?

(b) Give a 95% confidence interval for the mean change in blood pressure.

10. A random sample of 16 values from a normal population should mean 41.5 cm, and sum of squares of deviation from mean is 135 cm^2. Construct a 95% confidence interval for population mean.

11. A random sample of 300 were taken from a population of 9000 buffaloes in a region. 90 buffaloes were found suffering from a disease. Construct 95% confidence interval for the (a) proportion and (b) total number of buffaloes suffering from disease in the whole region.

12. Random sample of 64 men from a population has mean height equal to 68.8 inches and standard deviation of height equal to 2.4 inches. Find the 90% confidence interval for μ the mean height of men in the population.

13. Three treatments A, B & C are compared in a completely randomized design with 6 replications for each. The layout and wheat yield in Kg./plot are given in the following table ï

A 17	B 19	A 29	C 33	B 23	B 21
B 15	A 25	A 17	C 35	C 29	B 23
A 34	C 25	B 19	C 37	A 23	C 27

Analyze the experimental yields and state your conclusions?

14. The plants of wheat of 6 varieties were selected at random and the heights of their shoots were measured in cms.

Varieties			**Height in cms**						
1	85	90	89	93	84	87	-	-	-
2	88	87	94	95	91	-	-	-	-
3	95	93	87	89	91	95	92	93	-
4	83	89	90	84	85	85	-	-	-
5	90	89	61	93	88	89	90	-	-
6	93	89	87	88	88	89	90	87	90

Do the data indicate that there is no significant difference between the mean height of the plants of the different varieties?

CHAPTER-V

CORRELATION ANALYSIS

So far we have studied descriptive and inferential techniques that involve a single quantitative variable. In this chapter, we shall study the inferential techniques that involve two variables that are studied simultaneously for each individual. If for every value of a variable X we have a corresponding value of another variable Y then the series of pairs of values (X, Y) is known as a bivariate distribution. For example, a series of pairs of the ages of husbands and their wives at the time of marriage form a bivariate distribution.

In a bivariate distribution if the change in one variable appears to be accompanied by a change in the other variable and vice-versa then the variables are said to be correlated. The term correlation is mutual relationship between two or more variables. If for an increase (decrease) in the variable X, variable Y also shows increase (decrease) accordingly, the correlation will be positive. If the increase (decrease) in X, is accompanied by a corresponding decrease (increase) in variable Y, then the correlation will be negative. If change in one variable does not show a change in the other variable, the two variables are called uncorrelated or independent.

5.1 Methods of Measuring Correlation:

Scatter diagram, Karl Pearson coefficient of correlation and Spearmanôs rank correlation coefficient are the frequently used methods in correlation studies.

5.1.1 Scatter Diagram Method:

Let (X_i, Y_i), i =1, 2,·····,n be the n pairs of observations. If these paired observations are plotted on a graph paper in the XY-plane such that each pair is represented by dot in the diagram. The diagram so obtained is known as scatter or dot diagram. By looking at the scatter diagram of the various points, we can form an idea as to whether the variables are correlated or not. If all the points lie on a straight line arising from the lower left hand corner to the upper right hand corner, correlation is said to be perfect positive i.e. r = + 1. On the other hand, if all the points lying on a straight line falling from the upper left hand corner to the lower right hand corner, then correlation is said to be perfectly negative. If the plotted points fall in a narrow band, there would be a high degree of correlation

between the variables. Correlation shall be positive if points show a rising tendency from lower left hand corner to the upper right hand corner and negative if the points show a declining tendency from the upper left hand corner to the lower right hand corner of the diagram. On the other hand, if the points are widely scattered over the diagram, then we expect no correlation or poor correlation between X and Y.

Limitations of Scatter Diagram:

Scatter diagram only tells about the nature of the relationship whether it is positive or negative and whether it is high or low. It does not provide the extent of the measure of relationship between the variables. It is subjective in nature and different individuals may have different interpretations from the same set of data.

5.1.2 Karl Pearson Coefficient of Correlation or Product Moment Correlation Coefficient or Simple Correlation Coefficient:

It is a mathematical method for measuring the degree or strength of linear relationship between two variables and was suggested by Karl Pearson (1901). The Pearson coefficient of correlation between X and Y is denoted by the symbol r_{xy} or simply ‘r’ and is defined as the ratio of covariance between X and Y to the product of the standard deviations of X and Y.

$$r = \frac{Cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2}\sqrt{\sum (Y_i - \bar{Y})^2}}$$

where σ_x and σ_Y are the standard deviations of X and Y. Covariance and standard deviations can be written in terms of sum of squares and sum of products also.

$$Cov(X, Y) = \sum (X_i - \bar{X})(Y_i - \bar{Y}) / n$$

$$\sigma_X = \sqrt{\sum (X_i - \bar{X})^2 / n} \text{ and } \sigma_Y = \sqrt{\sum (Y_i - \bar{Y})^2 / n}$$

Here $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y and n is the number of paired observations. A simplified formula for computation of correlation coefficient is:

$$r = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sqrt{[\sum X_i^2 - (\sum X_i)^2/n][\sum Y_i^2 - (\sum Y_i)^2/n]}}$$

Shortcut and Step Deviation Method:

Since coefficient of correlation is independent of change of origin and change of scale, so the step deviation or short cut method can be used.

Case-I: If both change of origin and change of scale are considered

$$\text{i.e. } U = \frac{X - a}{h} \text{ and } V = \frac{Y - b}{k}$$

$$\text{then } r = \frac{\sum U_i V_i - (\sum U_i)(\sum V_i)/n}{\sqrt{\left[\sum U_i^2 - (\sum U_i)^2/n\right]\left[\sum V_i^2 - (\sum V_i)^2/n\right]}}$$

Case-II: If only change of origin is considered

i.e. dx= X-a and dy = Y-b

then

$$r = \frac{\sum dxdy - (\sum dx)(\sum dy)/n}{\sqrt{\sum dx^2 - (\sum dx)^2/n}\sqrt{\sum dy^2 - (\sum dy)^2/n}}$$

Properties:

i) It is a unit less number i.e. independent of the units of measurement.

ii) The correlation coefficient always lies between -1 & +1 i.e. -1 Ò r Ò 1.

iii) The coefficient of correlation is independent of change of scale and shift of origin of the variables X and Y.

iv) If two variables are independent, their correlation coefficient is zero but the converse is not true. It is because correlation coefficient measures only linear type of relationship, thus even if it is zero, the variables may have a non-linear relationship.

v) The degree of relationship between two variables is symmetrical i.e. $r_{yx} = r_{xy}$

Correlation for a Bivariate Frequency Distribution:

If the number of observations is very large and are divided into classes, in such situations the pairs of observations are represented in the form of a bivariate frequency distribution. For a bivariate frequency distribution of X and Y the correlation coefficient is calculated by the following formula:

$$r_{xy} = \frac{\sum f_{xy} XY - (\sum f_x X)(\sum f_y Y)/N}{\sqrt{[\sum f_x X^2 . (\sum f_x X)^2/n][\sum f_y Y^2 . (\sum f_y Y)^2/n]}}$$

where $N = \Sigma f_x = \Sigma f_y = \Sigma\Sigma f_{xy}$ and X and Y are mid values.

Testing the Significance of Population Correlation Coefficient

Case-1: Testing of ρ = 0 i.e. to test whether the variables in the population are linearly uncorrelated.

Step-wise Procedure:

i) $H_0 : \rho = 0$

ii) $H_1 : \rho \neq 0$ (Two Tailed Test)

$H_1 : \rho > 0$ (Right Tailed Test)

$H_1 : \rho < 0$ (Left Tailed Test)

iii) Select level of significance α

iv) Test statistic under H_0

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ follows student's t-distribution with (n-2) d.f.

v) Decision:

t_{cal} is compared with its critical t-value at (n-2) d.f. and level of significance α

Two Tailed Test Reject H_0 if $|t_{cal}| > t_{\alpha/2(n-2)}$

Right Tailed Test Reject H_0 if $t_{cal} > t_{\alpha(n-2)}$

Left Tailed Test Reject H_0 if $t_{cal} < -t_{\alpha(n-2)}$

Example-1: A random sample of married couples shows the age of husbands (x) and their wives (y) in different years as:

x:	30	29	36	72	37	36	51	48	37	50	51	36
y:	27	20	34	67	35	37	50	46	36	42	46	35

Calculate the coefficient of simple correlation between age of husbands and their wives and test for its significance

Solution: n = 12, $\Sigma x = 513$, $\Sigma y = 475$

$\Sigma x^2 = 23557$, $\Sigma y^2 = 20385$ and $\Sigma xy = 21861$

$$r = \frac{\Sigma xy - (\Sigma x)(\Sigma y)/n}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}} = \frac{21861 - \frac{513 \times 475}{12}}{\sqrt{\left[23557 - \frac{513^2}{12}\right]\left[20385 - \frac{475^2}{12}\right]}}$$

$$= \frac{21861 - 20306}{\sqrt{[23557 - 21930.75][20385 - 18802.08]}}$$

$$= \frac{1555}{\sqrt{1626.25 \times 1586.92}} = \frac{1555}{1606.44} = 0.969$$

H_0: $\rho = 0$ (age of husbands and their wives are independent)

H_1: $\rho \neq 0$ (Two tailed test)

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.969 \times \sqrt{12-2}}{\sqrt{1-0.969^2}} = \frac{0.969 \times \sqrt{10}}{\sqrt{1-0.939}} = 12.4$$

Since $|t_{cal}|$ > table t-value (2.228) at 10 df and $\alpha = 0.05$, therefore, we conclude that age of husband has a high positive correlation with the age of wife.

Example-2: Compute the Karl Pearson correlation coefficient from the following data and test for its significance.

Solution: **X : 9 8 7 6 5 4 3 2 1**

Y : 15 16 14 13 11 12 10 8 9

X	**$x = (X - \bar{X})$**	**x^2**	**Y**	**$y = (Y - \bar{Y})$**	**y^2**	**xy**
9	4	16	15	3	9	12
8	3	9	16	4	16	12
7	2	4	14	2	4	4
6	1	1	13	1	1	1
5	0	0	11	-1	1	0
4	-1	1	12	0	0	0
3	-2	4	10	-2	4	4
2	-3	9	8	-4	16	12
1	-4	16	9	-3	9	12
$\times X=45$	$\times x=0$	$\times x^2=60$	$\times Y=108$	$\times y=0$	$\times y^2=60$	$\times xy=57$

$\overline{X}$ = × X/n = 45/9 = 5

$\overline{Y}$ = × Y/n = 108/9 = 12

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{57}{\sqrt{60 \times 60}} = 0.95$$

Test of Significance: $H_0 : ɟ = 0$ vs $H1 : ɟ Í 0$ (Two tailed test)

Let Ŭ= 0.05

Test statistic: $t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.95\sqrt{9-2}}{\sqrt{1-(0.95)^2}} = \frac{0.95\sqrt{7}}{\sqrt{1-0.9025}} = \frac{2.513}{\sqrt{.0975}} = 8.05$

Since $|t_{cal}| = 8.05 > t_{tab} = 2.36$ at 7 d.f., therefore we reject H_0 and conclude that there is significant correlation among the variables in the population.

Example-3: Calculate the coefficient of correlation between total cultivable area (X) and the area under wheat (Y) from the following bivariate distribution of data selected from 66 villages.

Area under Wheat (in hectare)	**Total cultivable Area (in hectare)**					
	0-200	**200-400**	**400-600**	**600-800**	**800-100**	**Total**
0-50	12	6	-	-	-	18
50-100	2	18	4	2	1	27
100-150	-	4	7	3	-	14
150-200	-	1	-	2	1	4
200-250	-	-	-	1	2	3
Total	**14**	**29**	**11**	**8**	**4**	**66**

Solution:

Mid Values	**X**	**100**	**300**	**500**	**700**	**900**				
Y	**u** **v**	**-2**	**-1**	**0**	**1**	**2**	**Total f_y**	**f_yv**	**f_yv^2**	**$f_{xy}uv$**
25	-2	12 (48)	6 (12)				18	-36	72	60
75	-1	2 (4)	18 (18)	4 (0)	2 (-2)	1 (-2)	27	-27	27	18
125	0	-	4 (0)	7 (0)	3 (0)	-	14	0	0	0
175	1	-	1 (-1)	-	2 (2)	1 (2)	4	4	4	3
200	2	-	-	-	1 (2)	2 (8)	3	6	12	10
	Total f_x	**14**	**29**	**11**	**8**	**4**	**66**	**-53**	**115**	**91**
	f_xu	**-28**	**-29**	**0**	**8**	**8**	**-41**			
	f_xu^2	**56**	**29**	**0**	**8**	**16**	**109**			
	f_xuv	**52**	**29**	**0**	**2**	**8**	**91**			

From the table we get

$$N = 66,\ \Sigma f_x u = -41,\ \Sigma f_x u^2 = 109,\ \Sigma f_y v = -53,\ \Sigma f_y v^2 = 115 \text{ and } \Sigma f_{xy}\, uv = 91$$

$$\text{Thus } r_{xy} = \frac{\Sigma f_{xy} uv - (\Sigma f_x u)(\Sigma f_y v)/N}{\sqrt{\left[\Sigma f_x u^2 - (\Sigma f_x u)^2/N\right]\left[\Sigma f_y v^2 - (\Sigma f_y v)^2/N\right]}}$$

$$= \frac{91 - (-41)(-53)/66}{\sqrt{\{109 - (-41)^2/66\}\{115 - (-53)^2/66\}}} = 0.749$$

Thus the correlation coefficient between total cultivable area and the area under wheat is +0.749.

Example-4: For a random sample of 18 paired observations from a bivariate normal population, the correlation coefficient is obtained as -0.6. Test the significance of correlation in the population at Ŭ= 0.05.

Solution: Set up the null and alternative hypothesis as:

i) $H_0 : ɟ = 0$

ii) $H_1 : \rho \neq 0$ (Two Tailed Test)

iii) $\alpha = 0.05$

iv) Compute the test statistic

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.6\sqrt{18-2}}{\sqrt{1-0.36}} = \frac{-0.6(4)}{\sqrt{0.64}}$$

$$= \frac{-2.4}{0.8} = -3.00 \text{ with } (18-2) = 16 \text{ d.f.}$$

$|t_{cal}| = 3.00$

t_{tab} for two tailed test at $\alpha = 0.05$ $(t_{0.025,\ 16}) = 2.12$

Since $|t_{cal}| > t_{tab}$, therefore, we reject H_0 and conclude that there is significant correlation between the variables in the population.

Case-II: Testing of $\rho = \rho_0$ $(\rho_0 \neq 0)$

For testing the significance of correlation coefficient for a bivariate normal population in which $\rho \neq 0$, i.e. in non-null case, Prof. R.A. Fisher proved that the sampling distribution of r is by no means student's t and suggested the use of Fisher Z-transformation.

$Z_r = \frac{1}{2}\log_e \frac{1+r}{1-r}$ and proved that even for small samples, the distribution of Z_r is approximately normal with mean $Z_\rho = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho}$ and variance 1/(n-3) and for large values of n (say n>50), the approximation is fairly good.

Step-wise Procedure:

1. $H_0 : \rho = \rho_0$ $(\rho_0 \neq 0)$
2. $H_1 : \rho \neq \rho_0$
3. Select α
4. Test statistic under H_0

$$Z_r = \frac{1}{2}\log_e \frac{1+r}{1-r} = 1.1513\log_{10} \frac{1+r}{1-r}$$

$$Z_\rho = \frac{1}{2}\log_e \frac{1+\rho_0}{1-\rho_0} = 1.1513 \log_{10} \frac{1+\rho_0}{1-\rho_0}$$

Thus $Z_r \sim N\left(Z_\rho, \frac{1}{n-3}\right)$

Or $Z = \frac{Z_r - Z_p}{\sqrt{\frac{1}{n-3}}} = (Z_r - Z_\rho)\sqrt{n-3}$

If $|Z_{cal}| \geq z_{tab}$, at a specified α, then we reject H_0 and conclude that correlation coefficient in the population is significantly different from ρ_0.

Example-5: A correlation coefficient of 0.5 is obtained from a sample of 19 pairs of observations. Can the sample be regarded as drawn from a bivariate normal population in which true correlation coefficient is 0.7?

Solution:

i) $H_0 : \rho = 0.7$ and

ii) $H_1: \rho \neq 0.7$ (Two tailed test)

iii) Choose level of significant $\alpha = 0.05$

Applying Fisher Z-transformation, we get:

$$Z_r = 1.1513 \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5}$$

$= 1.1513 \log_{10} 3 = 1.1513\ (0.4771) = 0.5492$

and Mean $Z_P = 1.1513 \log_{10} \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+0.7}{1-0.7}$

$= 1.1513 \log_{10} 5.67 = 1.1513\ (0.7536) = 0.8676$

Computing Z-test statistic, we get:

$$Z_{cal} = \frac{Z_r - Z_p}{1/\sqrt{n-3}} = (Z_r - Z_p)\sqrt{n-3}$$

$= (0.5492 - 0.8676)\sqrt{16} = -0.3184(4) = -1.27$

Since the $|Z_{cal}| = 1.27$ is less than the tabulated value $Z_{\alpha/2} = 1.96$ at 5% level of significance, therefore, we do not reject H_0 and conclude that the sample may be regarded as coming from a bivariate normal population with $\rho = 0.7$.

Testing the Significance of Difference between two Independent Correlation Coefficients:

The above case concerning single correlation coefficient can be generalized to test the significance of difference between two independent correlation coefficients. Let r_1 and r_2 be the sample correlation coefficients observed in two independent samples of size n_1 and n_2, respectively. Then

$$Z_1 = \log_e\left(\frac{1+r_1}{1-r_1}\right) \text{ and } Z_2 = \log_e\left(\frac{1+r_2}{1-r_2}\right)$$

Testing Procedure:

i) $H_0 : \rho_1 = \rho_2 = \rho$ (say) Correlation coefficients do not differ significantly i.e. the samples are drawn from the same bivariate normal population or from the different populations with the same correlation coefficient ρ.

ii) $H_1: \rho_1 \neq \rho_2$ (Two Tailed Test)

iii) Choose level of significance α

iv) Test Statistic (Under H_0)

$$Z = \frac{Z_1 - Z_2}{\sqrt{V(Z_1 - Z_2)}} = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0,1)$$

(v) Conclusion

If $|Z_{cal}| \geq z_{\alpha/2}$, then reject H_0.

Example-6: Two independent samples of 23 and 28 pairs of observations were analyzed and their correlation coefficients were found as 0.5 and 0.8, respectively. Do these values differ significantly?

Solution:

i) $H_0 : \rho_1 = \rho_2$, i.e. correlation coefficients do not differ significantly i.e. the samples are drawn from the same population.

ii) $H_1 : \rho_1 \neq \rho_2$ (Two tailed test)

iii) $\alpha = 0.05$

iv) Test Statistic:

$$Z_1 = 1.1513 \log_{10} \frac{1+r_1}{1-r_1} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5}$$

$= 1.1513 \log_{10} 3 = 0.55$

$$Z_2 = 1.1513 \log_{10} \frac{1+r_2}{1-r_2} = 1.1513 \log_{10} \frac{1+0.8}{1-0.8}$$

$= 1.1513 \log_{10} 9 = 1.10$

$$Z_{cal} = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0.55 - 1.10}{\sqrt{\frac{1}{20} + \frac{1}{25}}} = \frac{-0.55}{0.30} = -1.83$$

Since $|Z_{cal}| = 1.83$ is less than $Z_{tab} = 1.96$ at 5% level of significance (two tailed test), therefore we do not reject H_0. Thus we conclude that the correlation values do not differ significantly.

5.1.3 Rank Correlation Coefficient:

Sometimes we come across statistical series in which the variables are not capable of quantitative measurement but can be arranged in the serial order. This happens when we are dealing with qualitative data. In such cases, Charleôs Edward Spearman, a British psychologist developed a formula in 1904 which gives the correlation coefficient between the ranks of n individuals on the two attributes under study.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i denotes the difference between the ranks of i^{th} paired observation i = 1, 2, é .., n

Properties of Spearman's Rank Correlation:

i) Spearmanôs formula is the only formula to be used for finding the correlation coefficient if we are dealing with qualitative data.

ii) Spearmanôs correlation can be applied even when the data do not follow normal distribution. It is the distribution free measure as it does not make any assumption about the population from which the samples are drawn.

iii) Spearmanôs formula and Karl Pearson formula give the same value if they are applied on the same data, provided no item is repeated or all items of the series are different.

iv) The limits of Spearmanôs rank correlation coefficient are from -1 to +1.

Note: The Spearmanôs formula cannot be applied in case of bivariate frequency distribution.

Equal Ranks/Repeated Ranks: When equal ranks are assigned to some entries, an adjustment in the above formula for calculating the coefficient of rank correlation is made. The adjustment consists of adding $(m^3-m)/12$ in the value of $\times d_i^2$, where m stands for the number of individuals whose ranks are common. If there are more than one such groups of individuals with common ranks, then this value is added as many times as the number of such groups and formula can be written as:

$$r_s = 1 - \frac{6[\sum d_i^2 + (m^3 - m)/12 + ...]}{n(n^2 - 1)}$$

5.1.4 Partial and Multiple Correlations: Quite often there is inter relation between various variables recorded during any study and consequently value of one variable is simultaneously influenced by many other variables. When more than two variables are involved in a study, four major problems may arise:

i) We may be interested in studying the inter-dependence or correlation of two variables only when other variables included in study are kept constant. This is the problem of partial correlation.

ii) We may also be interested in studying the correlation between dependent variable and a number of independent variables. This is the problem of multiple correlation.

iii) We may wish to examine the effect of number of independent variables upon the dependent variable. This is the problem of multiple regression.

iv) We may be interested in studying the effect of one independent variable upon the dependent variable when the effect of all other independent variables are eliminated and this is the problem of partial regression.

Partial Correlation: The partial correlation may be defined as a statistical tool measuring the linear relationship between two variables when all other variables involved in the study are kept constant or when their linear effects are eliminated. We denote $r_{12.3}$

as the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant and is computed as under:

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r^2_{13}} \cdot \sqrt{1 - r^2_{23}}}$$

Partial correlation coefficients such as $r_{12.3}$, $r_{13.2}$ are often referred to as first order partial correlation coefficients since one variable has been held constant. Further, $r_{12.34}$, $r_{13.24}$ and $r_{14.23}$ etc. are called second order partial correlation coefficients since two variables are kept constant and these may be computed as:

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} \cdot r_{24.3}}{\sqrt{1 - r^2_{14.3}} \cdot \sqrt{1 - r^2_{24.3}}}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2} \cdot r_{34.2}}{\sqrt{1 - r^2_{14.3}} \sqrt{1 - r^2_{34.2}}}$$

First order partial correlations are tested using ά ô test

$$t = \frac{r_{ab.c} \sqrt{(n-3)}}{\sqrt{1 - r^2_{ab.c}}} \text{ with (n-3) degrees of freedom}$$

where, $r_{ab.c}$ is the first order partial correlation between a and b keeping character c constant.

Second order partial correlation is also tested by ά ô test

$$t = \frac{r_{ab.cd} \sqrt{(n-4)}}{\sqrt{1 - r^2_{ab.cd}}} \text{ with (n-4) degrees of freedom}$$

where, $r_{ab.cd}$ is the second order partial correlation.

In general, k^{th} order partial correlation is tested by:

$$t_{cal} = \frac{r_{12.34....(k+2)} \sqrt{n-k-2}}{\sqrt{1 - r^2_{12.34.....k+2}}} \text{ with (n-k-2) degrees of freedom}$$

If $|t_{cal}|$ is > t_{tab} at a specified Ŭ, then we reject the null hypothesis.

Multiple Correlation: Multiple correlation is an extension of the technique of simple correlation to the problems which involve two or more independent variables. Multiple correlation may be defined as a statistical tool designed to measure the degree of

relationship of one dependent variable with three or more independent variables. It is denoted by a symbol R. In a trivariate distribution, in which each of the variable X_1, X_2 and X_3 has n observations, the multiple correlation coefficient of X_1 with X_2 and X_3 is usually denoted by $R_{1.\ 23}$. The subscripts to the left of the dot stand for the dependent variable while the subscripts to the right of the dot represent the independent variables. The coefficient of multiple correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows:

$$R_{1.23} = \sqrt{(r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23})/(1 - r^2_{23})}$$

$$R_{2.13} = \sqrt{(r^2_{12} + r^2_{23} - 2r_{12}r_{23}r_{13})/(1 - r^2_{13})}$$

$$R_{3.12} = \sqrt{(r^2_{31} + r^2_{32} - 2r_{31}r_{32}r_{12})/(1 - r^2_{12})}$$

where r_{12}, r_{13}, r_{23} stands for zero order or simple correlation coefficients. Multiple correlation coefficient can never be negative. It is necessarily positive or zero. By squaring $R_{1.23}$, we obtain the coefficient of determination which indeed is the per cent variability explained in dependent variable because of the influence of independent variables.

Testing the Significance of an observed Multiple Correlation Coefficient:

If R is the observed multiple correlation coefficient of a variate with ókô other variables in a random sample of size ónô from (k+1) variates population, then under the null hypothesis (H_o) that the multivariate correlation coefficient (R) in the population is zero, the statistic:

$$F = \frac{R^2}{1 - R^2} \quad \frac{n - k - 1}{k}$$ follows F distribution with (k, n-k-1) degrees of freedom.

If $F_{cal} \geq F_{Ŭ}(k, n ï\ k ï\ 1)$, then we reject H_0.

Example-7: In a poetry recitation competition, ten participants were ranked by two judges as:

Participant No.:	1	2	3	4	5	6	7	8	9	10
Judge x :	7	9	6	5	3	4	8	10	2	1
Judge y:	10	9	3	5	6	7	2	8	1	4

Measure the strength of relationship in the ranking behaviour of the two judges.

Solution: Let R_x and R_y denote the ranks given by the judges X and Y respectively.

Participant No.	$\mathbf{R_x}$	$\mathbf{R_y}$	$\mathbf{d = R_x - R_y}$	$\mathbf{d^2}$
1	7	10	-3	9
2	9	9	0	0
3	6	3	3	9
4	5	5	0	0
5	3	6	-3	9
6	4	7	-3	9
7	8	2	6	36
8	10	8	2	4
9	2	1	1	1
10	1	4	-3	9
				$\mathbf{\Sigma d^2 = 86}$

Here, n = 10 and $\Sigma d^2 = 86$

Using Spearmanôs rank correlation formula, we have

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 86}{10(10^2-1)} = 1 - \frac{6 \times 86}{10 \times 99} = 0.48$$

Testing significance of rank correlation coefficient: For n > 20 the distribution of Spearmanôs coefficient r_s tends to be normal with var $(r_s) = \frac{1}{n-1}$. However, for lower values of n (say 10) the statistic $t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$ follows (approximately) Studentôs t-distribution with (n-2) degrees of freedom.

Testing $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$

$$t_{cal} = \frac{r_s}{\sqrt{1-r_s^2}}\sqrt{n-2} = \frac{0.48}{\sqrt{1-0.48^2}}\sqrt{10-2} = 1.54$$

Table t-value at $\alpha = 0.05$ (Two-tailed test) for 8 df is 2.31

Since, $|t_{cal}| < t_{tab}$, hence the marks obtained by students of this group are independent.

Example-8: The marks of eight students in two papers economics and statistics are given below. Compute the rank correlation coefficient:

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Marks in Economics (X):	25	30	38	22	50	70	30	90
Marks in statistics (Y) :	50	40	60	40	30	20	40	70
Ranks assigned to X :	2	3.5	5	1	6	7	3.5	8
Ranks assigned to Y :	6	4	7	4	2	1	4	8
Difference in ranks (d) :	-4	-0.5	-2	-3	4	6	-0.5	0
d^2 :	16	0.25	4	9	16	36	0.25	0

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2-1)}$$

Here, $\Sigma d^2 = 81.5$, the item 30 is repeated twice hence $m_1=2$, item 40 is repeated thrice hence $m_2 = 3$.

$$r_s = 1 - \frac{6\left[81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right]}{8(8^2-1)}$$

$$= 1 - \frac{6[81.5 + 0.5 + 2]}{504} = 1 - 1 = 0$$

Hence, there is no rank correlation between the marks obtained in the two subjects.

Example-9: Calculate $r_{12.3}$, $r_{23.1}$ given r_{23}= -0.36, r_{31} = 0.34 and r_{12}= 0.70

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r^2_{13}}.\sqrt{1 - r^2_{23}}} = \frac{0.7 - (.34)(-.36)}{\sqrt{1 - (0.34)^2}.\sqrt{1 - (0.36)^2}} = 0.94$$

$$r_{23.1} = \frac{r_{23} - r_{12}.r_{13}}{\sqrt{1 - r^2_{12}}.\sqrt{1 - r^2_{13}}} = \frac{-0.36 - (0.7)(0.34)}{\sqrt{1 - (0.7)^2}.\sqrt{1 - (0.34)^2}} = -0.89$$

Example-10: Given the values n = 20, $r_{12.3}$ =0.7738, $r_{14.3}$ = 0.7243 and $r_{24.3}$ = 0.5262, calculate $r_{12.34}$ and test for its significance.

Solution: $r_{12.34} = \dfrac{r_{12.3} - r_{14.3}.r_{24.3}}{\sqrt{1-r^2_{14.3}}.\sqrt{1-r^2_{24.3}}}$

substituting the values, we get

$$r_{12.34} = \frac{.7738 - (.7243)(.5262)}{\sqrt{1-(.7243)^2}.\sqrt{1-(.5262)^2}} = 0.67$$

Test of significance: $H_0 : \rho_{12.34} = 0$ vs $H_1 : \rho_{12.34} \neq 0$ Let $\alpha = 0.05$

$$t_{cal} = \frac{r_{12.34}}{\sqrt{1-r^2_{12.34}}}\sqrt{n-4} = \frac{0.67\sqrt{20-4}}{\sqrt{1-0.67^2}} = \frac{0.67 \text{ x } 4}{0.74} = 3.62$$

Since, $|t_{cal}| = 3.62 > t_{0.05,16} = 2.12$, therefore, we reject H_0 and conclude that there is significant partial correlation between the variables.

Example-11: Calculate $R_{1.23}$ for the following data and test its significance

$n = 25$, $r_{12} = 0.60$, $r_{13} = 0.70$ and $r_{23} = 0.65$

Solution: We have $R_{1.23} = \sqrt{(r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23})/(1-r^2_{23})}$

$$= \sqrt{\{(0.6)^2 + (0.7)^2 - 2(0.6)(0.7)(0.65)\}/\{(1-(0.65)^2\}} = 0.725$$

Test of significance: H_o: There is no multiple correlation in the population

H_1: Multiple correlation in the population is greater than zero

Let Ŭ= 0.05

We know that:

$F = \dfrac{R^2}{1-R^2} \quad \dfrac{n-k-1}{k}$ follows F distribution with (k, n-k-1) degrees of freedom

So, $F_{cal} = \dfrac{(0.725)^2}{[1-(0.725)^2]} \text{ x } \dfrac{25-2-1}{2} = 12.19$

The tabulated value of F with (2, 22) d.f. at 5% level of significance is 3.44. Here the calculated value of F statistic is greater than the tabulated value. So we reject the H_0 and conclude that observed multiple correlation coefficient is significantly greater than zero.

EXERCISES

1. Obtain the coefficient of rank correlation between share and debenture prices.

Share Prices (X)	50	55	65	50	55	60	50	65	70	75
Debenture Prices (Y)	110	110	115	125	140	115	130	120	115	160

2. Calculate coefficient of simple correlation from the following data:

X:	11	8	4	2	6	7	8	6	8	4	10	8
Y:	10	3	1	2	3	5	2	10	11	3	2	5

3. Calculate the coefficient of rank correlation for the following data:

Students	1	2	3	4	5	6
Marks in math	75	40	52	65	60	80
Marks in statistics	30	45	35	50	48	42

4. The following table gives the frequency distribution according to age groups of marks obtained by 52 students in an intelligence test. Calculate the coefficient of correlation between age and intelligence.

Marks	**Age in years**				
	16-18	**18-20**	**20-22**	**22-24**	**Total**
10-20	2	1	1	-	4
20-30	3	2	3	2	10
30-40	3	4	5	6	18
40-50	2	2	3	4	11
50-60	-	1	2	2	5
60-70	-	1	2	1	4
Total	**10**	**11**	**16**	**15**	**52**

CHAPTER-VI

REGRESSION ANALYSIS

The term ñregression ñliterally means ñstepping back towards the averageò. It was first used by a British biometrician, Sir Francis Galton (1822-1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to ñregressò or ñstepbackò to the average population height.

By regression we mean average relationship between two or more variables. One of these variables is called the dependent variable or response variable, and the other variable is the independent or the explanatory variable. If the explanatory variables are two or more then it is called the multiple regression analysis. Regression analysis can be further divided into linear and non-linear. In the linear regression, the dependent variable varies at a constant rate with a given change in the independent variable, the constant rate of change can be in absolute terms or in terms of percentage. In the non linear regression, the dependent variable changes at varying rates with a given change in the explanatory variable.

6.1 Linear Regression:

Linear regression line is one which gives the best estimate of one variable (Y) for any given value of the other variable (X). It should be noted that the two regression lines i.e. one of Y on X and another of X on Y cut each other at the point of average of X and Y. The regression equation of Y on X can be expressed as $Y = a + b X + e$, where Y is dependent variable and X is an independent variable, óaô is the intercept which the regression line makes with the y-axis, óbô is the slope of the line and e_i are the random error i.e. the effect of some unknown factors. The values of a and b are obtained by the method of least squares by which we find a and b such that the errors sum of squares $\Sigma e_i^2 = \Sigma(y_i - a-bx_i)^2$ is minimized. Mathematically it is minimized by differentiating it partially w.r.t. parameters to be estimated and putting them equal to zero. The two normal equations will be obtained as under:

$$\times Y = n a + b \times X \quad \text{---- (i)} , \qquad \times XY = a\times X + b\times X^2 \quad \text{---- (ii)}$$

Solving these two normal equations we can get the estimated values of a and b as

$$a = \bar{Y} - b\bar{X} \quad \text{and}$$

$$b_{yx} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n}$$

The regression equation Y on X becomes

$$Y = a + b_{yx} X$$

$$= Y - b_{yx}\bar{X} + b_{yx} X = Y + b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - \bar{Y} = b_{yx}(X - \bar{X}) \text{ where } b_{yx} \text{ is the regression coefficient of Y on X}$$

Similarly the regression equation of X on Y can be expressed as $X = a + b_{xy} Y$ and is obtained by interchanging X and Y, the normal equations will be:

$$\times X = n\,a + b \times Y \text{ ---- (i)}, \qquad \times XY = a \times Y + b \times Y^2 \text{ ---- (ii)}$$

and can again be solved for getting the estimated values of a and b, i.e.

$$a = \bar{X} - b\bar{Y} \quad \text{and}$$

$$b_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum Y_i^2 - (\sum Y_i)^2/n}$$

Similarly regression equation of X on Y can directly be written as:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

b_{xy} is the regression coefficient of X on Y.

Shortcut Method:

Since regression coefficients are independent of change of origin and not of change of scale, therefore, short cut method may be used as follows:

Consider $dx = X - A$ and $dy = Y - B$

then

$$b_{yx} = \frac{\sum dxdy - \sum dx \sum dy/n}{\sum dx^2 - (\sum dx)^2/n}$$

and $$b_{xy} = \frac{\sum dxdy - \sum dx \sum dy/n}{\sum dy^2 - (\sum dy)^2/n}$$

Regression Equation in terms of Correlation Coefficient:

The regression coefficients can be written in terms of correlation coefficient as follows:

$b_{yx} = r\sigma_y/\sigma_x$ and $b_{xy} = r\sigma_x/\sigma_y$

and the regression equations of Y on X and X on Y can be expressed as

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

Testing the Significance of an observed Regression Coefficient:

Consider the null hypothesis Ho: $b = b_0$

and the alternative hypothesis H_1: $b \neq b_0$

Choose a suitable level of significance, say $\alpha = 0.05$

Test statistic is $t = \frac{b - b_0}{SE(b)}$ follows t-distribution with (n-2) d.f.

Where $SE(b) = \frac{s^2}{s_{xx}}$ and $s^2 = \frac{SSE}{n-2} = \frac{s_{yy} - b\,s_{xy}}{n-2}$

If $|t_{cal}| \geq t_{\alpha/2,\, n-2}$ then we reject H_0.

Properties of Regression Coefficients:

i) Correlation coefficient is the geometric mean between two regression coefficients, i.e. $r = \pm\sqrt{b_{yx}\, b_{xy}}$

ii) If one of the regression coefficients is greater than one, the other must be less than one.

iii) Arithmetic mean of the regression coefficients is greater than the correlation coefficient.

iv) Regression coefficients are independent of change of origin but not of scale.

v) Both the regression coefficients are of the same sign and this sign depends on the sign of covariance.

Difference between Correlation and Regression:

- The coefficient of correlation measures the degree of linear relationship between two variables where as the regression coefficient gives the average change in dependent variable corresponding to a unit change in independent variable.
- The coefficient of correlation lies from -1 to 1. This can never exceed unity while the regression coefficient can exceed unity.
- The coefficient of correlation is always symmetrical for any two variables (r_{xy}= r_{yx}) but for the regression coefficient it is not so i.e. b_{yx} is not equal to b_{xy}.
- The coefficient of correlation is independent of change of scale and shift of origin but the regression coefficient is independent of the shift of origin only but not of scale.
- In regression analysis, the variables have cause and effect relation which is not required in correlation analysis.
- Correlation analysis is confined to study of the linear relationship between the variables and thus have limitations, while the regression analysis has much wider applications as it can study linear and non linear relationship between the two variables.
- Correlation coefficient has no unit while regression coefficient has the unit of dependent variable. It indicates the amount of change in dependent variable as per unit change in independent variable.

6.2 Non-linear Regression:

Sometimes it may happen that the original data is not in a linear form but can be reduced to linear form by some simple transformation of variables. This will be illustrated by considering the following curves:

Fitting of a Power Curve: $Y= aX^b$ to a set of n points $(x_1, y_1), (x_2, y_2),$ é .., (x_n, y_n)

Taking logarithm of both sides, we get

$$\log Y = \log a + b \log X$$

$\Rightarrow$ $U= A+ bV$, where $U= \log Y$, $A= \log a$ and $V =\log X$.

This is a linear equation in U and V.

Fitting of Exponential Curves: (i)$Y = ab^x$, (ii) $Y = ae^{bX}$ to a set of n points

i) $Y= ab^x$

Taking logarithm of both sides, we get

$\log Y = \log a + X \log b$

$\Rightarrow$ $U = A + BX$, where $U= \log Y$, $A = \log a$ and $B = \log b$

After solving the normal equations for estimating A and B, we finally get

a = antilog (A) and b= antilog (B)

ii) $Y= ae^{bX}$

$\log Y= \log a + bX \log e = \log a + (b \log e) X$

$\Rightarrow$ $U = A+ BX$, where $U= \log Y$, $A = \log a$ and $B = b \log e$.

After solving the normal equations for A and B, we have

a = antilog (A) and b= B/log e

Example-1: Using the following data, obtain the regression equation of Y on X

X: Additional Expenditure (000 Rs); Y = Sales (Crore Rs)

X:	14	19	24	21	26	22	15	20	19
Y:	31	36	48	37	50	45	33	41	39

Also estimate the sale for additional expenditure of Rs. 25000/-

Solution:

X	**Y**	**$x = (X - \bar{X})$**	**$y = (Y - \bar{Y})$**	**xy**	**x^2**
14	31	-6	-9	54	36
19	36	-1	-4	4	1
24	48	4	8	32	16
21	37	1	-3	-3	1
26	50	6	10	60	36
22	45	2	5	10	4
15	33	-5	-7	35	25
20	41	0	1	0	0
19	39	-1	-1	1	1
$\sum X=180$	**$\sum Y=360$**	**$\sum x=0$**	**$\sum y=0$**	**$\sum xy=193$**	**$\sum x^2=120$**

$\bar{X} = 180/9=20, \quad \bar{Y} = 360/9=40$

$$b_{yx} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = 193/120 = 1.608$$

Regression equation of Y on X is:

$$\hat{Y} - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\hat{Y} - 40 = 1.608 (X - 20)$$

$$\hat{Y} = 1.605 X + 7.84$$

$$\hat{Y}_{25} = 1.605(25) + 7.84 = 40.125$$

Thus estimated sales for additional expenditure of Rs. 25000 is Rs. 40.125 Crores

Example-2: The ages and blood pressures of 9 men are given below:

Age (X)	55	41	36	47	49	42	60	72	63
BP (Y)	142	124	117	127	144	138	154	157	148

i) Find the correlation coefficient between X and Y and test for its significance

ii) Find the regression equation of Y and X

iii) Estimate the blood pressure of a man of 40 years

Solution:

Age	BP	(X-50)		(Y-139)		
X	Y	dx	dx^2	dy	dy^2	dxdy
55	142	5	25	3	9	15
41	124	-9	81	-15	225	135
36	117	-14	196	-22	484	308
47	129	-3	9	-10	100	30
49	144	-1	1	5	25	-5
42	138	-8	64	-1	1	8
60	154	10	100	15	225	150
72	157	22	484	18	324	396
63	148	13	169	9	81	117
Total 465	**1253**	**15**	**1129**	**2**	**1474**	**1154**

i) Coefficient of correlation is given by

$$r = \frac{n\sum dxdy - \sum dx \sum dy}{\sqrt{n\sum dx^2 - (\sum dx)^2}\sqrt{n\sum dy^2 - (\sum dy)^2}}$$

$$= \frac{9(1154) - (15)(2)}{\sqrt{9(1129) - (15)^2}\sqrt{9(1474) - (2)^2}}$$

$$= \frac{10386 - 30}{\sqrt{10161 - 225}\sqrt{13266 - 4}} = \frac{10356}{11479} = 0.902$$

Test of Significance

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.902\sqrt{7}}{\sqrt{1-(0.902)^2}} = 5.53$$

Since $|t_{cal}| > t_{0.025,\ 7} = 2.305$. Hence there is positive and significant correlation in the population.

ii) The regression equation of Y on X is

$$\hat{Y} = \overline{Y} + \hat{b}_{yx}(X - \overline{X})$$

$$\overline{X} = \frac{465}{9} = 51.7\ ;\ \overline{Y} = \frac{1253}{9} = 139.2\ ;$$

$$\hat{b}_{yx} = \frac{n\Sigma dxdy - \Sigma dx \Sigma dy}{n\Sigma dx^2 - (\Sigma dx)^2} = 10356/9936 = 1.04$$

Hence $\hat{Y}$ =139.2+1.04(X-51.7) = 85.31 + 1.04 X

iii) For X = 40 $\hat{Y}$ = 85.31 + 1.04 (40) = 126.91

Hence, the estimated blood pressure for a man of 40 years is 127.

Example-3: For the following results

Variance of X = 9

Regression equations 8X ï 10Y + 66 = 0; 40X ï 18Y = 214, Find

i) The mean value of X and Y

ii) Coefficient of correlation between X and Y, and

iii) Standard deviation of Y

Solution:

i) Since regression lines pass through $(\overline{X}, \overline{Y})$, therefore, we have

$8\overline{X}$ ï $10\overline{Y}$ = -66 (1)

$40\overline{X}$ ï $18\overline{Y}$ = 214 (2)

Multiplying equation (1) by 5

$40\overline{X}$ -50$\overline{Y}$ = -330

$40\overline{X}$ -18$\overline{Y}$ = 214

-32$\overline{Y}$ = -544 hence $\overline{Y}$ = 17

Putting the value of $\overline{Y}$ in equation (1)

$8\overline{X}$ -10(17) = -66 we get $\overline{X}$ = 13

ii) Coefficient of correlation between X and Y

Let (i) is the regression equation of X and Y

8X = 10Y ï 66

$$X = \frac{10}{8}Y - \frac{66}{8} \text{ or } b_{xy} = \frac{10}{8}$$

From equation (2) ï 18Y = 214 ï 40X

$$Y = -\frac{214}{18} + \frac{40}{18}X \text{ or } b_{xy} = \frac{40}{18}$$

Since both regression coefficients are greater than 1, our assumption is wrong. Hence equation (1) is regression equation of Y on X.

-10Y = -66 ï 8X

$$Y = \frac{66}{10} + \frac{8}{10}X \qquad \text{or} \qquad b_{yx} = \frac{8}{10}$$

From equation (2) 40X = 214 + 18Y

$$X = \frac{214}{40} + \frac{18}{40}Y \qquad \text{or} \qquad b_{xy} = \frac{18}{40}$$

$$\text{Thus } r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{18}{40} \times \frac{8}{10}} = \sqrt{0.36} = 0.6$$

iii) The standard deviation of Y can be determined from any regression coefficient

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

Substituting the values

$$\frac{18}{40} = 0.6\frac{3}{\sigma_y}, \text{ we get } \sigma_y = 4$$

Example-4: The following data relate to marks obtained by 250 students in Economics and Statistics in an examination:

Subject	**Arithmetic Mean**	**Standard Deviation**
Economics	48	4
Statistics	55	5

Coefficient of correlation between marks in economics and statistics is +0.8. Draw the two lines of regression and estimate the marks obtained by a student in statistics who secured 50 marks in economics.

Solution: Let marks in economics be denoted by X and in statistics by Y.

Regression equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$\bar{X} = 48, \bar{Y} = 55, \ \sigma_x = 4; \sigma_y = 5, r = 0.8$$

$$X - 48 = 0.8\ (4/5)\ (Y-55)$$

$$X - 48 = 0.64\ (Y-55)$$

$$X = 0.64Y + 12.8$$

Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$\hat{Y}$ ï 55 = 0.8(5/4) (X ï 48)

$\hat{Y}$ ï 55 = (X ï 48)

$\hat{Y}$ = 7 + X

For X = 50 $\hat{Y}$ = 7 + 50 = 57

Thus for marks in economics equal to 50, estimated marks in statistics shall be 57.

Example-5: Following Statistics were obtained in a study conducted for examination of relationship between yield of wheat and annual rainfall.

	Yield (kg/acre)	**Annual Rainfall (inches)**
Mean :	985.0	12.8
SD :	70.1	1.6

r = 0.52

Assuming a linear relationship between yield and rainfall obtain yield of wheat/acre when the rainfall of 9.2 inches.

Solution: Let the rainfall be denoted by X and yield by Y. The required yield can be obtained from the regression equation of Y on X, which is

$$\hat{Y} - \overline{Y} = r\frac{\hat{u}_y}{\hat{u}_X}(X - \overline{X})$$

$$\hat{Y} - 985 = 0.52\frac{70.1}{1.6}(X - 12.8) = 22.78\ (X - 12.8)$$

or $\hat{Y} - 985 = = 22.78\ X - 291.58$

or $\hat{Y} = 693.42 + 22.78X$

when X = 9.2, $\hat{Y}$ = 693.42 + 22.78 x 9.2 = 903 kg/acre

Example-6: Given

X :	36	28	38	42	44	46	30	34	32	40
Y :	128	156	142	135	177	184	149	191	163	170

i) Calculate the Karl-Pearson correlation coefficient between X and Y and interpret the results.

ii) Obtain the equations for two regression lines and estimate the values of Y for X= 30 and also the value of X for Y = 149

Solution: From the given data we obtain

n = 10, $\Sigma X = 370$, $\Sigma Y = 1595$

$\Sigma X^2 = 14020$, $\Sigma Y^2 = 258445$ and $\Sigma XY = 59274$

$$S_{xx} = \sum X^2 - \frac{(\sum X)^2}{n} = 14020 - \frac{(370)^2}{10} = 330$$

$$S_{yy} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 258445 - \frac{(1595)^2}{10} = 4042.5$$

and $$S_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 59274 - \frac{370 \text{ x } 1595}{10} = 259$$

i) $$r = \frac{S_{XY}}{\sqrt{S_{XX}\ S_{YY}}} = \frac{259}{\sqrt{330 \text{ x } 4042.5}} = \frac{259}{1155} = 0.224$$

Test of significance:

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.224}{\sqrt{1-0.224^2}}\sqrt{10-2} = 0.65$$

Let $\alpha = 0.05$ then $t_{0.025,\,8} = 2.31$

Since $t_{cal} < t_{0.025,\,8}$, so correlation between X and Y is statistically non-significant at 5% level.

ii) Let the equation of regression line of Y and X be.

$$Y = a + b_{yx} X$$

Then $$b_{yx} = \frac{S_{xy}}{S_{xx}} = \frac{259}{330} = 0.785$$

$$a = \overline{Y} - b_{yx}\,\overline{X} = 159.5 - 0.785 \text{ x } 37 = 130.46$$

Thus $$Y = 130.46 + 0.785\,X$$

When X = 30, $\hat{Y}$ = 130.46 + 0.785 x 30 = 154.01

Also, let the equation of the line of regression of X and Y be

$\hat{X} = a + b_{xy} Y$

Then $b_{xy} = \frac{S_{xy}}{S_{yy}} = \frac{259}{4042.5} = 0.064$

and $a = \bar{X} - b_{xy} \bar{Y}$ = 37 – 0.064 x 159.5 = 26.79

Then, $\hat{X} = 26.79 + 0.064Y$

When Y = 149, $\hat{X}$ = 26.79 + 0.064 x 149 = 36.33

Example-7: The length of panicles (x) in cm and the number of grains per panicles (y) for 15 plants from a field of paddy are given below. Fit a regression line of y on x and estimate the number of grains per panicle when panicle length in 25 cm.

x:	22.4	23.3	24.1	24.3	23.5	22.3	23.9	24	24.9	20	19.8	22	24.5	23.6	21.1
y:	95	109	133	132	136	116	126	124	137	90	107	108	143	127	92

Solution: From the given data, we have

n = 15, Σx = 343.7, Σy = 1775

Σx^2 = 7911.17, Σy^2 = 214247 and Σxy = 40998.6

$\bar{x}$ = 22.91 and $\bar{y}$ = 118.33

Also, $s_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 214247 - \frac{(1775)^2}{15} = 4205.33$

$s_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 7911.17 - \frac{(343.7)^2}{15} = 35.86$

$s_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 40998.6 - \frac{343.7 \text{ x } 1775}{15} = 327.43$

$b = \frac{s_{xy}}{s_{xx}} = \frac{327.43}{35.86} = 9.13$

$a = \bar{y} - b\,\bar{x} = 118.33 - 9.13 \text{ x } 22.91 = 118.33 - 209.17 = -90.84$

$\hat{y} = -90.84 + 9.13\,x$, $\hat{y}$ when x = 25 is 137.41

Example-8: From 18 pairs of observations on height of plant (X) and their respective produce (Y) following quantities were obtained.

$\overline{X} = 60$, $\overline{Y} = 10$, $\Sigma(X - \overline{X})^2 = 1500$, $\Sigma(Y - \overline{Y})^2 = 60$ and

$\Sigma(X - \overline{X})\ (Y - \overline{Y}) = 180$

Find the two regression coefficients and hence calculate the correlation coefficient between X and Y and test its significance

Solution: $b_{xy} = \dfrac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - \overline{X})^2} = \dfrac{180}{1500} = 0.12$

$b_{yx} = \dfrac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(Y - \overline{Y})^2} = \dfrac{180}{60} = 3.0$

$r_{xy} = \sqrt{b_{yx}\ b_{xy}} = \sqrt{0.12 \times 3} = \sqrt{0.36} = 0.6$

testing the hypothesis H_0 : $\rho = 0$ against H_1 : $\rho \neq 0$

$t_{cal} = \dfrac{r}{\sqrt{1 - r^2}}\sqrt{n-2} = \dfrac{0.6}{\sqrt{1-0.36}}\sqrt{18-2} = \dfrac{0.6 \times 4}{\sqrt{0.65}} = \dfrac{2.4}{0.8} = 3$

table $t_{0.025,\ 16} = 2.12$ $|t_{cal}| > t_{tab}$, therefore we reject H_0

Example-9: For 12 wheat earheads, the characters, length of earhead (x) and grain per earhead (y) were recorded and the following quantities were obtained:

$\overline{x} = 20,\ \overline{y} = 30$

$s_{xx} = 800$, $s_{yy} = 996$ and $s_{xy} = 760$

Fit a regression line and estimate the grains per earhead of length 10 units. Also test the significance of the regression coefficient.

Solution: The equation of the regression line of Y on X given by $\hat{y} = \hat{b}_0 + \hat{b}_1 x$

Given $n = 12,\ \overline{x} = 20,\ \overline{y} = 30$

$s_{xx} = 800,\ s_{yy} = 996$ and $s_{xy} = 760$

Hence, $\hat{b}_1 = \dfrac{s_{xy}}{s_{xx}} = \dfrac{760}{800} = 0.95$

and $\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x} = 30 - 0.95 \text{ x } 20 = 11$

The regression equation of y on x is

$$\hat{y} = 11 + 0.95x$$

when x = 10, $\hat{y}$ = 20.5. Thus we expect 20.5 grains per earhead of length 10 units on an average.

The regression coefficient $\hat{b}_1$ = 0.95, which indicates that for every unit increase of the length of the earhead, we can expect an increase of 0.95 grains per earhead on the average. For testing significance of the regression coefficient, we formulate and test the hypothesis.

H_0 : $b_1 = 0$ against H_1 : $b_1 \neq 0$

To test H_0, we need to obtain an estimate s^2 of σ^2, the variance of the error term. Hence

The sum of squares due to error (SSE) = s_{yy} - $\hat{b}_1$ s_{xy}

= 996 ï 0.95 x 760 = 274

and $s^2 = \dfrac{SSE}{n-2} = \dfrac{274}{10} = 27.4$

The test statistic, t, is calculated as

$$t_{cal} = \frac{\hat{b}_1}{SE(\hat{b}_1)} = \frac{\hat{b}_1}{\sqrt{s^2/S_{xx}}} = \frac{0.95}{\sqrt{27.4/800}} = \frac{0.95}{0.185} = 5.135$$

$t_{0.05,\,10}$ = 2.228

Since $|t_{cal}| > t_{tab}$, therefore, we reject the null hypothesis and conclude that the number of grains per earhead increases significantly with the increase in length of the earhead.

EXERCISES

1. Find the regression equations of X on Y and Y on X from the data given below

   Husbandôs age (X) 26 28 30 31 35

   Wifeôs age (Y) 20 27 28 30 25

   Also calculate the coefficient of correlation.

2. Obtain the two lines of regression and estimate the value of Y if X is 70 and that of X if Y is 90 from the following given data:

	X-series	Y-series
Arithmetic Mean	18	100
Standard deviation	14	20

   Coefficient of correlation between X and Y series = 0.8

3. From 18 pairs of observation on height of papaya plants (X) in inches and their produce (Y) in kg, the following quantities were calculated.

   $\overline{X} = 60$, $\overline{Y} = 10$, $\Sigma(X_i - \overline{X})^2 = 1500$

   $\Sigma (Y_i - \overline{Y})^2 = 60$ $\Sigma (X_i - \overline{X})(Y_i - \overline{Y}) = 180$

   Calculate the regression of yield on the height of plant and test its significance. Also calculate the coefficient of correlation between the two variates and test its significance.

Chapter-VII

NON-PARAMETRIC TESTS

Many common statistical tests like t, Z and F, used to test the hypothesis concerning population parameters like, population mean (μ), population variance$(\sigma)^2$ etc., are based on certain assumptions about the population or the parameters of the population and hence they are called parametric tests. The most important assumption in the application of these tests is that the populations from which samples are drawn follow normal distribution. Also the usual t-test for testing the difference of two means requires that samples be drawn from populations which are normally distributed and should have equal variances. For small samples, if the population is not normally distributed, the conclusion based on parametric tests will no longer be valid. For such situations, statisticians have devised alternate procedures, which do not require the above assumptions about the population distribution (except some mild or weaker assumptions such as observations are independent and continuous are to be satisfied) and are known as non-parametric or distribution free tests. These tests are applicable even when the data are measured on nominal and ordinal scales and utilize some simple aspects of sample data such as sign of measurements or rank statistic or category frequencies.

Advantages of Non-Parametric Tests:

i) No assumption is made about the distribution of the parent population from which the sample is drawn.

ii) These tests are readily comprehensible and easy to apply.

iii) Since the data in social sciences, in general, do not follow normal distribution, hence non-parametric tests are frequently applied by social scientists.

iv) Non-parametric methods are applicable to data which are given in ranks or grades like A^+, A, B, B^-, C etc.

v) If the sample size is as small as 6, there is no alternative except to use a non-parametric test unless the nature of the population distribution is exactly known.

Disadvantages:

i) Non parametric tests can result in loss of much of the information contained within the data.

ii) They are less efficient and less powerful than their counter parametric tests.

iii) The drawback of ranking, on which many non-parametric tests are based, is that the magnitude of the observations is lost and so the important information within the data is not properly utilized.

From the above discussion, it is clear that whenever the assumption of parametric tests are met, then the non-parametric tests should not be used, as these tests are not as sensitive and powerful as the classical parametric tests, in dealing with the data. When we are unable to apply parametric tests only then we resort to non-parametric tests. In this chapter, various non-parametric tests applicable under different situations are discussed and illustrated with examples.

7.1 One Sample Tests:

These tests lead us to decide whether the population follows a known distribution or the sample has come from a particular population. We can also test whether the median of the population is equal to a known value. A test is also given to test the randomness of a sample drawn from a population as it is the crucial assumption in most of the testing procedures.

7.1.1 Runs Test for Randomness:

One of the most important aspects of all types of statistical testing is that the sample selected must be representative of the population as far as possible since decisions about the population are to be made on the basis of results obtained from samples. Thus the requirement of the randomness of the sample is mandatory. However, in many situations, it is difficult to decide whether the assumption of randomness is fulfilled or not. The assumption of randomness can be tested by runs test, which is based on the theory of runs.

Run and Run Length: A run is defined as a sequence of identical letters (symbols) which are followed and preceded by different letters or no letter at all and number of letters in a run is called run length.

Suppose that after tossing a coin say 20 times, following sequence of heads (H) and tails (T) occur:

HH	TTT	HHH	T	HH	TT	HH	T	HH	TT
1	2	3	4	5	6	7	8	9	10

The first sequence of HH is considered as a run of length 2. Similarly occurrence of TTT is considered as another run of length 3 and so on. So counting the runs in similar way, the total number of runs occurred in above sequence is 10 i.e. R = 10.

The total number of runs (R) in any given sample indicates whether the sample is random are not. Too many or too small number of runs creates doubt about the randomness of sample.

Test Procedure:

Generalizing the problem, let one kind of elements be denoted by plus (+) sign and second kind of elements be denoted by minus (-) sign. The concept of ó+ô and ó-ó provides the direction of change from an established pattern. In the above example if óHô is denoted by ó+ô and óTô denoted by ó-ó then we have

n_1 = number of ó+ô signs = 11

n_2 = number of ó-ô signs = 9

and total sample size $n = n_1 + n_2 = 20$

while the number of runs (R) = 10

Now if sample size is small so that n_1 and n_2 are less than or equal to 20 each, then to test the hypothesis:

H_0: Observations in the sample are random against

H_1: Observations in the sample are not random

Compare the observed number of runs (R) in the sample with two critical values of (R) for given values of n_1 and n_2 at a predetermined level of significance (critical values of óRô for Runs test are available in the Appendix of various statistics books)

Decision Rule:

Reject H_0 if $R \leq c_1$ or $R \geq c_2$, otherwise accept H_0 where c_1 and c_2, are two critical values and may be obtained from standard tables.

For the above example the critical value c_1 from less than table for $n_1 = 11$ and $n_2 = 9$ is ó6ô and from more than table the critical value c_2 is 16. So our acceptance region is

$6 < r < 16$. Since the observed value of 'R' is 10, so H_0 is accepted, hence we conclude that our sample observations are drawn at random.

Large sample runs test:

If the sample size is large so that n_1 or n_2 is more than 20, the sampling distribution of 'R' can be closely approximated by the normal distribution with mean and variance.

$$E(R) = \varepsilon_r = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$V(R) = \sigma^2_r = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

$$\text{Thus } Z = \frac{R - \varepsilon_r}{\mathring{u}_r} \sim N(0,1)$$

So, if $|Z_{cal}| \geq 1.96$ then reject H_0 at 5% level of significance and conclude that the sample is not random

< 1.96 then accept H_0 at 5% level of significance

Example-1: A researcher wants to know whether there is any pattern in arrival at the entrance of the shopping mall in terms of males and females or simply such arrivals are random. One day he stationed himself at the main entrance and recorded the arrival of Men (M) and Women (W) of first 40 shoppers and noted the following sequence.

M WW MMM W MM W M W M WWW MMM W MM WWW MMMMMM WWW MMMMMM

Test randomness at 5% level of significance

Solution: H_0 : Arrival of men and women is random

H_1 : Arrival is not random

Here Number of men (n_1) = 25

Number of women (n_2) = 15

Number of runs (R) = 17

Since here $n_1 = 25$ is > 20 so sampling distribution of 'R' is approximated by normal distribution with mean:

$$\varepsilon_r = \frac{2n_1n_2}{n_1+n_2}+1 = \frac{2 \text{ x } 25 \text{ x } 15}{25+15}+1 = \frac{750}{40}+1 = 19.75$$

$$\text{and } \mathring{u}_r = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2 \text{ x } 25\text{x } 15\,(2 \text{ x } 25 \text{ x } 15 - 25 - 15}{(25+15)^2\,(25+15-1)}}$$

$$\mathring{u}_r = \sqrt{\frac{750 \text{ x } 710}{1600 \text{ x } 39}} = 2.39$$

$$\text{Thus, } Z_{cal} = \frac{R-\varepsilon_r}{\mathring{u}_r} = \frac{17-19.75}{2.39} = -1.15$$

$\therefore$ Since $|Z_{cal}| < 1.96$ so H_0 is accepted i.e. sample is considered as random.

Note: To test the randomness of a sample of n observations, runs can also be generated by considering positive and negative signs of the deviations of observations from the median of the sample.

Let x_1, x_2, é .., x_n be a sample of size n drawn in that order and we wish to test:

H_0 : Observations in the sample are drawn at random

H_1 : Observations in the sample are not random

Let M be the sample median. For generating runs of positive and negative signs, we compute the deviations x_1 ï M, x_2 ï M,·····,x_n ï M and consider only the signs of these deviations.

Let n_1 and n_2 be the number of +ve and -ve signs so that $n_1 + n_2 \leq n$ (ignoring the zero deviation) and R be the total number of runs of +ve and ï ve signs in the sample.

Decision Rule:

Follow the usual procedure of runs test as given above.

Example-2: Following measurements were recorded in a sample of 20 earheads in a Wheat variety V_1: 8.9, 8.4, 10.3, 11.1, 7.8, 9.3, 9.9, 8.2, 10.9, 10.3, 10.8, 8.6, 9.4, 8.9, 9.4, 8.9, 9.5, 9.9, 9.6, 9.7, 9.2 and 10.0. Test the randomness of the sample using runs test.

Solution:

H_0 : Sample is drawn at random

H_1 : Sample is not drawn at random

Let Ŭ= 0.05

Test Statistic: Median is found to be 9.55, therefore, generating runs of the +ve and ï ve signs, we have.

8.9	8.4	10.3	11.1	7.8	9.3	9.9	8.2	10.9	10.3
-	–	+	+	–	–	+	–	+	+
10.8	8.6	9.4	8.9	9.5	9.9	9.6	9.7	9.2	10.0
+	–	–	–	–	+	+	+	–	+

Find the median which is the arithmetic mean of two middle observations 9.5 and 9.6 come out to be equal to 9.55

Consider the signs of the deviation x_i ï 9.55, i = 1, 2,·····,20 and count the number of runs R as the test statistic:

- - + + - - + - + + + - - - - + + + + - +

Here number of plus signs (n_1) = 10

Number of minus signs (n_2) = 10

Number of runs (R) = 10

Conclusion:

Critical value c_1 for n_1 = 10 and n_2 = 10 (at Ŭ= 0.05) = 6

Critical value c_2 for n_1 = 10 and n_2 = 10 (at Ŭ= 0.05) = 16

The test statistic R (=10) lies in the acceptance region 6 < R < 16. Hence H_0 is not rejected and we conclude that the sample is considered as drawn at random.

Non-Parametric Alternative to One Sample t-test:

7.1.2 Sign Test:

The sign test is used for testing the median rather than mean as location parameter of a population i.e. whether the sample has been drawn from a population with the specified median M_o. It is a substitute of one sample t-test when the normality assumption of the parent population is not satisfied. This test is the simplest of all the non-parametric tests and its name comes from the fact that it is based upon the signs of the differences and not on their numerical magnitude.

Let x_1, x_2,·····,x_n be a random sample of size ónô from a population with unknown median M.

It is required to test the null hypothesis H_0: M = M_0 (some pre-specified value) against a two-sided (H_1: M ≠ M_0) or one sided (H^*_1: M > M_0 or H^{**}_1 : M < M_0) alternative.

If the sample comes from the distribution with median M_0, then on the average, half of the observations will be greater than M_0 and half will be smaller than Mo. Compute the differences (x_i - M_0) for i = 1, 2,·····,n and consider their signs and ignore the sample values equal to M_0. Let the number of plus and minus signs be r and s respectively, with r + s ≤ n. For test statistic, we consider only the plus signs (r).

The distribution of ór ô for given n is a binomial distribution with p = P (X > M_0) = 0.5. Thus the null hypothesis becomes:

H_0: p = 0.5 vs $H_{1:}$ p ≠ 0.5 or H^*_1: p > 0.5 or H^{**}_1 : p < 0.5. In case there is pre-information that sample has an excess number of plus signs we may use one tailed test (i.e. H^*_1: p > 0.5). Similarly, alternative hypothesis H^{**}_1 : p < 0.5 will be chosen when it is expected that the sample will have few plus signs. The only difference between one tailed and two-tailed tests is that of critical values for a prefixed α.

Test criterion:

For a two tailed test, reject H_0 if r ≥ $r_{\alpha/2}$ or ≤ $r'_{\alpha/2}$ where $r_{\alpha/2}$ and $r'_{\alpha/2}$ are the critical values at the significance level α. Here $r_{\alpha/2}$ is the smallest integer such that

$$\sum_{r=r_{\alpha/2}}^{n} \binom{n}{r}\left(\frac{1}{2}\right)^n \le \alpha/2$$

and $r'_{\alpha/2}$ is the largest integer such that

$$\sum_{r=0}^{r'_{\alpha/2}} \binom{n}{r}\left(\frac{1}{2}\right)^n \le \alpha/2$$

for one tailed alternative $H^*_1 - p > 1/2$, reject H_0 if r≥ r_α where r_α is the smallest integer such that

$$\sum_{r=r_{\check{U}}}^{n} \binom{n}{r}\left(\frac{1}{2}\right)^n \le \check{U}$$

and for $H^{**}_1 - p > 1/2$, reject H0 if r ≤ r'_α where r'_α is the largest integer such that

$$\sum_{r=0}^{r'_\alpha}\binom{n}{r}\left(\frac{1}{2}\right)^n \le \alpha$$

Large Sample Approximation:

If r + s > 25, then normal approximation to the binomial may be used. In that case test statistic Z is computed as under:

$$Z = \frac{r - (r+s)/2}{\sqrt{(r+s)/4}} = \frac{r-s}{\sqrt{r+s}}$$

If $|Z_{cal}| \ge Z_{tab}$ at a specified Ŭ, then we reject H_0.

Example-3: Following measurements were recorded for length (cms) of 20 randomly selected earheads of a wheat variety V. 8.9, 8.4, 10.3, 11.1, 7.8, 9.3, 9.9, 8.2, 10.9, 10.3, 10.8, 8.6, 9.4, 8.9, 9.5, 9.9, 9.6, 9.7, 9.2 and 10.0. Test the hypothesis that the median length of earheads is (i) equal to 9.5 cm (ii) more than 9.5 cm (iii) less than 9.5 cm at α = 0.05.

Solution: Obtain the signs of the differences (x_i ï 9.5), which are pasted below the measurements. We find that there are 10 plus, 9 minus and one zero differences. Ignore the zero difference so that the sample size n reduced to 19. In this case n = 19 and r (number of plus signs = 10).

8.9	8.4	10.3	11.1	7.8	9.3	9.9	8.2	10.9	10.3
−	−	+	+	−	−	+	−	+	+
10.8	8.6	9.4	8.9	9.5	9.9	9.6	9.7	9.2	10.0
+	−	−	−	Ignored	+	+	+	−	+

Conclusion:

i) The critical region for a two sided alternative (H_1 : M Í 9.5 cms) at α = 0.05 is given by $r \ge r_{\alpha/2} = 14$ or $r \le r_{\alpha/2} = 4$ Since here r = 10 which is neither greater than equal to 14 nor less than equal to 4 thus the null hypothesis is not rejected. Thus we conclude that the median length of ear heads is not significantly different from 9.5 cm.

ii) The critical region for a right alternative $(H_1^* : M > 9.5\,\text{cms})$ is given by $r \ge r_\alpha$. We note from the table $r_{.05} = 14$, hence H_0 is not rejected.

iii) Similarly the critical region for left alternative $(H_1^* : M > 9.5\,\text{cms})$ is given by $r \le r_\alpha^*$. We note from the table $r'_{0.05} = 5$, hence H_0 is not rejected.

Example-4: The following data relate to the daily production of cement (in metric tonnes) in a large plant for 30 days.

11.5	10.0	11.2	10.0	12.3	11.1	10.2	9.6	8.7	9.3
9.3	10.7	11.3	10.4	11.4	12.3	11.4	10.2	11.6	9.5
10.8	11.9	12.4	9.6	10.5	11.6	8.3	9.3	10.4	11.5

Use sign test to test the null hypothesis that the plant average daily production (μ) of cements is 11.2 metric tonnes against alternative hypothesis $\mu < 11.2$ metric tonnes at the 0.05 level of significance.

Solution: H_0: Average production of cement is = 11.2 metric tonnes

H_1: Average production of cement is < 11.2 metric tonnes

Let Ŭ= 0.05

Take the +ve and ï ve signs according as the deviation from 11.2 is +ve or ï ve

+	-	0 (Ignored)	-	+	-	-	-	-	-
-	-	+	-	+	+	+	-	+	-
-	+	+	-	-	+	-	-	-	+

Number of plus signs(r) 11

Number of minus signs (s) 18

Number of zeros 1

So the Sample Size reduced to 29 and r = 11

Substituting the values in the formula:

$$Z = \frac{r - (r+s)/2}{\sqrt{(r+s)/4}} = \frac{r-s}{\sqrt{r+s}} = \frac{11-18}{\sqrt{11+18}} = \frac{-7}{\sqrt{29}} \simeq -1.3$$

Since $|Z_{cal}|$ is > ï $Z_{0.05}$ = -1.645, the null hypothesis is not rejected and we conclude that the plantôs average production of cement is equal to 11.2 m tonnes.

7.1.3 Wilcoxon Signed Ranks Test:

Wilcoxon signed ranks test (Wilcoxon, 1945, 1949) is similar to sign test as it is used to test the same hypothesis about the median of the population. The sign test is based only on the signs of differences but Wilcoxon Signed Rank test takes into consideration not only the signs of differences such as positive or negative but also the

size of the magnitude of these differences. So this test is more sensitive and powerful than the sign test provided the distribution of population is continuous and symmetric.

Let $x_1, x_2, \ldots, x_n$ denote a random sample of size 'n' drawn from a continuous and symmetric population with unknown median M. We are required to test the hypothesis about the median (M) that is:

$H_0 : M = M_0$ against the alternative hypothesis $H_1 : M \neq M_0$

Choose $\alpha = 0.05$

Take the differences of sample values from M_0 i.e. $d_i = x_i - M_0$, i = 1, 2, ..., n and ignore the zero differences. Then assign the respective ranks to the absolute differences in ascending order of magnitude (after ignoring the signs of these differences) so that the lowest absolute value of the differences get the rank '1' second lowest value will get rank '2' and so on. For equal values of absolute differences, average value of ranks would be given.

After assigning the ranks to these differences assign the sign of original differences to these ranks. These signed ranks are then separated into positive and negative categories and a sum of ranks of each category is obtained. Let T^+ denote the sum of ranks of the positive d_i's and T^- denote the sum of ranks of negative differences.

Then clearly $T^+ + T^- = \dfrac{n(n+1)}{2}$, n being the number of non-zero d_i's.

For Small Samples:

Let T = Minimum (T^+, T^-) is taken as test statistic. If $T_{cal} \leq$ critical values of T at specified values of n and α, then H_0 is rejected.

For Large Samples:

For n > 25, T is approximately normally distributed under H_0 with $\mu_T = n(n+1)/4$

and $\sigma_T = \sqrt{n(n+1)(2n+1)/24}$ and $Z_{cal} = \dfrac{T - \varepsilon_T}{\sigma_T}$ is to be compared with z_{tab} for testing H_0 against H_1.

Example-5: Following measurements were recorded for length (in cms) of 20 randomly selected ear-heads of a wheat variety. Test the hypothesis whether median length is equal to 9.5 cm.

8.9	8.4	10.3	11.1	7.8	9.3	9.9	8.2	10.9	10.3
10.8	8.6	9.4	8.9	9.5	9.9	9.6	9.7	9.2	10.0

Solution:

H_0 : M = 9.5 cm

H_1 : M Í 9.5 cms

Let Ŭ= 0.05

The signed deviations of the observations from 9.5 cm with the ranks of their absolute values in parenthesis are:

- 0.6 (9.5), -1.1 (14), 0.8 (11.5), 1.6 (18), -1.7 (19), -0.2 (3.5), 0.4 (6.5),

-1.3(15.5), 1.4(17), 0.8 (11.5), 1.3 (15.5), -0.9(13), - 0.1(1.5), - 0.6 (9.5),

0 (ignored), 0.4 (6.5), 0.1 (1.5), 0.2 (3.5), -0.3(5), 0.5(8)

Here T^+ = 99.5 and T^- = 90.5 so that T= 90.5.

The table value T_α (two tailed test) for n=19 (number of non-zero deviations) at α = 0.05 is 46. Since $T_{cal} > T_{Ŭ}$, therefore H_0 is not rejected and we conclude that median length of the earheads is equal to 9.5 cms.

Using large sample approximation:

$$\mu_T = n(n+1)/4 = 19(20)/4 = 95$$

and $$\mathring{u}_T = \sqrt{n\,(n+1)\,(2n+1)/24} = \sqrt{19\,(20)\,(39)/24} = 24.8$$

$$Z_{cal} = \frac{T - \varepsilon_T}{\mathring{u}_T} = \frac{90.5 - 95}{24.8} = -0.18$$

Since $|Z_{cal}|$ = 0.18 is less than Z_{tab} (at Ŭ = 0.05) = 1.96, therefore, H_0 is not rejected.

7.2 Two Sample Tests for Dependent Samples (Non-Parametric Alternative to paired t-test):

7.2.1 Paired Sample Sign Test:

This test is a non-parametric alternative to paired t-test and is used for the comparison of two dependent samples. Here it is desired to test the null hypothesis that the two samples are drawn from the populations having the same median i.e.

$H_0 : M_1 = M_2$ vs $H_1 : M_1$ Í M_2.

The procedure of single sample sign test explained in section 7.1.2 can be applied to paired sample data.

Here we draw a random sample of nô pairs and observations (x_1, y_1), (x_2, y_2) é .. (x_n, y_n) giving ónô differences

$$d_i = x_i ï y_i \quad \text{for } i = 1, 2,\cdots\cdots,n$$

It is assumed that the distribution of differences is continuous in the vicinity of its median ôMô i.e. $P[d > M] = P[d < M] = 1/2$.

All the procedure of one-sample sign test will remain valid for the paired sample sign test, with d_i displaying the role of x_i in one sample sign test.

7.2.2 Paired Sample Wilcoxon Signed Ranks Test:

This test (Wilcoxon, 1945, 1949) deals with the same problem as the paired sample sign test and is an extension of one sample Wilcoxon signed ranks test (Section 7.1.3). But this test is more powerful than paired sample sign test since it takes into account the sign as well as the magnitude of the difference between paired observations.

The observed differences $d_i = x_i - y_i$ are ranked in the increasing order of absolute magnitude and then the ranks are given the signs of the corresponding differences. The null and alternative hypotheses are the same as in paired sample sign test i.e.

$$H_0 : M_1 = M_2 \text{ vs } H_1 : M_1 \neq M_2$$

If H_0 is true, then we expect the sum of the positive ranks to be approximately equal to the absolute value of the sum of negative ranks. The whole procedure of this test is the same as that of one sample Wilcoxon signed ranks test with the only difference that in this test $d_i$ôs are given by

$$d_i = x_i ï y_i, \qquad i = 1, 2, é .., n$$

Example-6: The weights of ten men before and after change of diet after six months are given below. Test whether there has been any significant reduction in weight as a result of change of diet at 5% level of significance.

Solution: The difference between before and after weights in the sample alongwith their signed ranks are given below:

Sr. No.	Weight (kgs.)		Difference (d_i)	Rank
	Before (x_i)	After (y_i)		
1	74	61	13	10
2	80	69	11	9
3	69	61	8	5
4	82	72	10	7.5
5	64	71	-7	-4
6	85	79	6	3
7	71	75	-4	-2
8	91	81	10	7.5
9	84	75	9	6
10	64	62	2	1

H_0 : There is no effect of diet in reducing the weight

H_1 : Diet is effective in reducing the weight

Summing up positive and negative ranks we have

$T^+ = 49$, $T^- = 6$ so that T = Minimum (49, 6) = 6

The critical value of the Wilcoxon statistic (T_α) at $\alpha = 0.05$ (one tailed test) for n = 10 is equal to ($T_{0.05}$) = 10. Since $T_{cal} < T_\alpha$, therefore, H_0 is rejected. Thus we conclude that there is significant reduction in weight as a result of change of diet.

Example-7: Two types of package programmes were offered to 30 farmers in an investigation and were used to award scores for each programme on its merits and scores are given below. Test whether there is any significant difference between two types of programmes at $\breve{U} = 0.05$ by (i) Paired sample Wilcoxon signed ranks test (ii) Paired Sign test.

Solution: The differences in the sample values alongwith their signed ranks are given below:

Farmer Sr. No.	Type-I (x_i)	Type II (y_i)	$d_i = x_i - y_i$	Rank	Farmer Sr. No.	Type I (x_i)	Type II (y_i)	$di = x_i - y_i$	Rank
1	64	68	-4	-12	16	39	50	-11	-25
2	70	72	-2	-4.5	17	47	40	7	19
3	65	60	5	14.5	18	35	35	0	ignored
4	72	69	3	8.5	19	57	50	7	19
5	35	42	-7	-19	20	68	52	16	28
6	52	49	3	8.5	21	70	68	2	4.5
7	45	45	0	ignored	22	43	51	-8	-21.5
8	76	73	3	8.5	23	59	55	4	12
9	60	58	2	4.5	24	38	39	-1	-1.5
10	48	54	-6	-16.5	25	59	51	8	21.5
11	39	42	-3	-8.5	26	45	50	-5	-14.5
12	67	54	13	26	27	62	68	-6	-16.5
13	50	65	-15	-27	28	72	63	9	23
14	76	75	1	1.5	29	78	68	10	24
15	42	40	2	4.5	30	48	52	-4	-12

Solution: Here, H_0 : Two package programmes are equally meritorious

H_1 : Two package programmes are not equally meritorious

Counting the sum of positive ranks (T^+) and negative ranks (T^-), we get

$T^+ = 227.5,\ T^- = 178.5$

$T = \text{Minimum}\ (T^+, T^-) = 178.5$

Since the sample size is large, therefore, the normal approximation is used. Hence we have

$$Z_{cal} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{178.5 - \frac{30(31)}{4}}{\sqrt{\frac{30(31)(61)}{24}}} = -1.11$$

Since $|Z_{cal}| = 1.11 < 1.96$, we, therefore, do not reject H_0 and conclude that two package programmes are equally meritorious i.e. there is no significant difference between the merits of two programmes.

ii) Now we will apply paired sign test to solve the above problem. We note that

Number of plus signs (r) = 16

Number of minus signs (s) = 12

Number of zeros = 2, hence sample size reduced to 28.

Since r + s = 28, therefore, normal approximation to the binomial distribution may be used. Hence the value of test statistic is given by

$$Z_{cal} = \frac{r - s}{\sqrt{r + s}} \qquad = \frac{16 - 12}{\sqrt{16 + 12}} = 0.76$$

Since $|Z_{cal}| < 1.96$, therefore, we do not reject H_0.

7.3 Two Sample Tests for the Independent Samples:

Non-Parametric Alternative to two sample t-test:

7.3.1 Mann-Whitney U-test:

This test was developed by Mann and Whitney (1974) and is a non-parametric alternative to the usual two-sample t-test. To apply this test, the two samples that are to be compared should be independent and the variable under consideration should have a continuous distribution. The null hypothesis under this test is that the two population distributions from which the samples have been drawn are identical. Under experimental situation, the null hypothesis may be that two treatments are identical i.e. two treatment effects do not differ significantly.

Procedure: Let the two independent samples $x_1, x_2, ..., x_{n1}$ and $y_1, y_2, ..., y_{n2}$ be drawn from two populations having distribution functions F(x) and F(y) which are assumed to be continuous. The hypothesis to be tested is

$$H_0 : F(x) = F(y) \qquad \text{vs} \qquad H_1 : F(x) \neq F(y)$$

For small samples:

When n_1 and n_2 both are less than 8, the test is described below:

The $(n_1 + n_2)$ observations of both the samples are combined and arranged in ascending order of their magnitude alongwith the identity to the sample of which each observation belong.

Find out U_1 by counting how many scores of Xôs precede (are lower than) each score of Yôs and add them.

Find out U_2 by counting how many scores of Yôs precede (are lower than each score of Xôs and add them.

We reject H_0 for small values of $U = \min (U_1, U_2)$ i.e. if the calculated value of U is less than or equal to the tabulated value for given n_1 and n_2 from the table of critical values for Mann-Whitney U-statistic, then the null hypothesis is rejected.

For moderately large samples:

When size of any one sample lies between 9 and 20, we have following steps:

Combine the $(n_1 + n_2)$ observations belonging to both the samples and rank them into a single group in ascending order of magnitude. For tied score give the average rank. The ranks of the Xôs (first sample) and Yôs (second sample) are summed up as R_1 and R_2 respectively.

A more convenient way of finding U_1 and U_2 is given below:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

or $\quad U_2 = n_1 n_2 - U_1$

Tables of critical values for moderate sample sizes at a specified Ŭ are available and test criterion is the same as explained for small samples.

For large samples:

When one or both the samples are larger then 20, the tables of critical values of U-statistic are not much useful. With the increase in sample size, the sampling distribution of U takes the shape of a normal distribution and the corresponding standard normal variate is given by:

$$Z = \frac{U - n_1 n_2 / 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

Conclusion:

i) For two tailed test if $|Z_{cal}| \geq 1.96$, then H_0 is rejected at $\alpha = 0.05$

ii) For right tailed test if $Z_{cal} \geq 1.64$, then H_0 is rejected at $\alpha = 0.05$

iii) For left tailed test if $Z_{cal} \leq -1.64$, then H_0 is rejected at $\alpha = 0.05$

Example-8: An experiment was conducted for comparing two types of grasses on nine plots of size 5x 2 m each. The yields of grasses per plot (kgs) at the time of harvesting are as given below:

	1	**2**	**3**	**4**	**5**	**6**
Grass I (X)	1.96	2.12	1.64	1.78	1.99	
Grass II (Y)	2.13	2.10	2.24	2.08	2.20	2.32

Test whether:

i) There is significant difference between the two types of grasses in respect of yield;

ii) Is grass I is lower yielder than grass II

Solution: (i) $H_0 : F(x) = F(y)$ i.e. there is no significant difference in the yields of two types of grasses

$H_1 : F(x) \neq F(y)$

Combining the data of both the samples and ranking them in ascending order of magnitude, we have

Yield	1.64	1.78	1.96	1.99	2.08	2.10	2.12	2.13	2.20	2.24	2.32
Rank	1	2	3	4	5	6	7	8	9	10	11
Sample	X	X	X	X	Y	Y	X	Y	Y	Y	Y

Calculating the values of U_1 and U_2 by counting the number of Xôs preceding (are lower than) each Y and adding up and vice versa.

$U_1 = 4 + 4 + 5 + 5 + 5 + 5 = 28$

and $U_2 = 0 + 0 + 0 + 0 + 2 = 2$

The critical value of U-statistic at $\alpha = 0.05$ ($n_1 = 5$, size of smaller sample and $n_2 = 6$, size of large sample) for a two tailed test = 3

Since $U = \text{Min}(U_1, U_2) = 2$ which is less than critical value, therefore, H_0 is rejected. Thus we conclude that there is significant difference in the yields of two types of grasses.

ii) In this case $H_1 : F(x) < F(y)$

Critical value of U at $\alpha = 0.05$ for one tailed test = 5

Since calculated $U_2 = 2$ is less than the critical value, therefore, H_0 is rejected. Thus it is concluded that grass I is lower yielder than grass-II.

Example-9: A random sample of 15 students was taken from private school in city and another random sample of 12 students was taken from public school in the same city and was administered the same test in English. The scores of students from both schools were recorded as follows:

Sr. No.	School (A)	Ranks	School (B)	Ranks
1	73	14	70	11
2	75	16	78	19
3	83	23.5	79	20
4	77	18	81	22
5	72	13	65	7
6	69	10	63	5
7	56	2	74	15
8	80	21	83	23.5
9	68	9	67	8
10	60	3	76	17
11	84	25	88	27
12	61	4	48	1
13	64	6		Sum (R_2)=176
14	71	12		
15	86	26		
		Sum (R_1) = 202		

Test whether the quality of education in English in private schools is similar to the one in public schools.

Solution: After mixing both the samples and arranging in ascending order of their magnitude and assigning ranks we have

48	1	69	10	78	19
56	2	70	11	79	20
60	3	71	12	80	21
61	4	72	13	81	22
63	5	73	14	83	23
64	6	74	15	83	24
65	7	75	16	84	25
67	8	76	17	86	26
68	9	77	18	88	27

Sum of ranks for private school (R_1) = 202

Sum of ranks for public school (R_2) = 176

$n_1 = 15$ and $n_2 = 12$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (12)(15) + \frac{15(15+1)}{2} - 202$$

= 180 + 120 - 202 = 98

$$U_2 = (12)(15) + \frac{12(12+1)}{2} - R_2$$

= 180 + 78 ï 176 = 82

$\therefore$ U = minimum (U_1, U_2) = 82

Critical value of U-statistic for $n_1 = 12$ and $n_2 = 15$ at Ŭ= 0.05 (for two tailed test) = 49. Since $U_{cal} = 82$ is more than the critical value, therefore, H_0 is accepted.

Although the sample sizes n_1= 12 and n_2= 15 are not large, but for the sake of illustration, we solve the example through normal approximation.

Thus assuming sample sizes large (for illustration) we have:

$$E(U) = \frac{12 \text{ x } 15}{2} = 90$$

$$\mathring{u}_u = \sqrt{\frac{12 \text{ x } 15(12+15+1)}{12}} = \sqrt{\frac{180(28)}{12}} = 20.5$$

$$\therefore Z_{cal} = \frac{U - E(U)}{\mathring{u}_u} = \frac{82 - 90}{205} = -0.39$$

Since |z| = 0.39 < 1.96

We accept H_0 i.e. evidence does not suggest that there is any significant difference in the quality of education in English in public and private schools.

7.3.2 Siegel-Tukey Test for Equal Variability:

This test (Siegel and Tukey, 1960) is used to test the null hypothesis whether two samples have been drawn from the two populations having the same variance i.e. whether the two populations have the equal variability. Siegel-Tukey test is the non-parametric alternative to the corresponding parametric F-test. The only assumptions underlying this test are that the populations have continuous distribution and sample sizes are not too small i.e. $n_1 + n_2 > 20$. So our hypothesis is

H_0 : $\sigma_1^2 = \sigma_2^2$ i.e. two populations have equal variance

H_1 : $\sigma_1^2 \neq \sigma_2^2$

Procedure:

Draw random samples of sizes n_1 and n_2 from the two populations. Then the observations of the two samples are combined and arranged in order of their increasing size. Ranks are allocated according to the following scheme:

- The lowest value is ranked 1.
- The highest two values are ranked 2 and 3 (the largest value is given rank 2)
- The lowest two unranked values are ranked 4 and 5 (the smallest value is given the rank 4).
- The highest two unranked values are ranked 6 and 7 (the largest value is given the rank 6)

This procedure continues, working from the end towards the centre, until no more than one unranked value remains. That is to say, if the number of values is odd, the middle value has no rank assigned to it.

Let n_1 and n_2 denote the sizes of the two samples and let $n_1 \leq n_2$. Let R_x and R_y denote the rank sums of two series X and Y and let R be the minimum of the rank sums of two series. Then the value of test statistic is given by

$$Z = \frac{R - n_1(n_1 + n_2 + 1)/2 + \frac{1}{2}}{\sqrt{n_1 n_2 (n_1 + n_2 + 1/12}} \sim N(0, 1)$$

If $|Z_{cal}| > 1.96$ reject H_0 at Ŭ= 0.05

and if $|Z_{cal}|$ Ò1.96 accept H_0

Example-10: An agronomist wants to know if a particular variety of paddy gives the same variability or spread in the yield under two different agronomic practices X and Y. Thus he laid the experiment in 10 plots for each practice and the yields (q ha^{-1}) are given below:

x	:	57.3	56.1	52.4	58.8	58.5	60.1	60.1	59.8	62.6	59.4
y	:	63.1	52.9	53.6	65.3	66.5	53.6	54.2	61.7	57.3	54.9

Apply Siegal-Tukey rank sum dispersion test to test the assertion.

Solution: Rank assignment to the combined data of two samples X and Y

Sample	x	y	y	y	y	y	x	x	y	x
Value	52.4	52.9	53.6	53.6	54.2	54.9	56.1	57.3	57.3	58.5
Rank	1	4	5	8	9	12	13	16	17	20
Sample	x	x	x	x	x	y	x	y	y	y
Value	58.8	59.4	59.8	60.1	60.1	61.7	62.6	63.1	65.3	66.5
Rank	19	18	15	14	11	10	7	6	3	2

Here $n_1 = n_2 = 10$

$R_x = 1 + 13 + 16 + 20 + 19 + 18 + 15 + 14 + 11 + 7 = 134$

$R_y = 4 + 5 + 8 + 9 + 12 + 17 + 10 + 6 + 3 + 2 = 76$

Hence R Minimum (134, 76) = 76

$$Z = \frac{76 - 10\,(10+10+1)/2 + \frac{1}{2}}{\sqrt{(10 \times 10)\,(10+10+1)/12}} = \frac{-28.5}{\sqrt{175}} = \frac{-28.5}{13.23} = -2.15$$

Here $|Z_{cal}| = 2.15 > 1.96$, thus we reject the null hypothesis and conclude that there is significant difference in variability in the yields of paddy under two different practices.

7.4 k-Sample Tests ($k \geq 3$)

1. Median Test:

The median test is used to test the null hypothesis whether two or more populations have the same median or not. To test this assertion, random samples of sizes $n_1, n_2, \cdots, n_k$ are drawn from ókô populations. The observations obtained in each sample are combined and median of the combined sample is determined. This is called grand median say (M). Let O_{1i} (i = 1, 2, ·····,k) denote the number of observations in the i^{th} sample that exceed the grand median ómô and let O_{2i} denote the number of observations in the i^{th} sample, which are less than or equal to the grand median. Arrange these frequencies into 2 x k contingency table as follows:

	1	**2**	**3**	**……**	**k**	**Total**
> Median	O_{11}	O_{12}	O_{13}	é é é é é .	O_{1k}	N_1
≤ Median	O_{21}	O_{22}	O_{23}	é é é é ..é	O_{2k}	N_2
	n_1	n_2	n_3	é é é é é ..	n_k	N

Thus N_1 denotes the number of observations above the grand median and N_2 the number of observations less than or equal to the grand median in all samples then $N = N_1 + N_2$ are the total number of observations.

Assumptions:

1. Samples are drawn randomly and independent of each other.
2. The measurement scale is ordinal

Test Procedure:

H_0 : Median of all ðkô populations is same, against the alternative.

H_1 : At least two populations have different median.

The rest of the procedure is the same as that of chi-square test for independence.

Test statistic: Let E_{ij} denote expected frequency of (i, j)th cell in 2 x k contingency table (i = 1, 2 and j = 1, 2,·····,k). Now under null hypothesis, the expected frequencies are given by:

$$E_{1i} = \frac{N_1}{N} n_i \text{ and } E_{2i} = \frac{N_2}{N} n_i$$

Thus test statistic for 2 k categories is given by:

$$T = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

which approximately follows a chi-square distribution with ðk-1ô degree of freedom

Test Criteria:

If the value of $\left(\frac{N-1}{N}\right)T$ is greater than or equal to tabulated value of χ^2 for (k-1) d.f. reject H_0, otherwise H_0 is not rejected.

Note: For k = 2, i.e. for the comparison of two populations the contingency table is given below:

	Sample –I	**Sample – II**	**Total**
> Median	a	b	R_1
< Median	c	d	R_2
Total	C_1	C_2	

$$\chi^2_{cal} = \frac{N(ad - bc)^2}{R_1 R_2 C_1 C_2}$$

If N < 50 or any cell frequency is less than 5, then Yatesô correction of continuity is applied and we get:

$$\chi^2_{cal} = \frac{N[|ad - bc| - N/2]^2}{R_1 R_2 C_1 C_2}$$

If $\chi^2_{cal} \geq \chi^2_{1,\alpha}$, the critical value of χ^2 distribution with 1 df at a given α, then reject H_0.

Example-11: The scores of two groups of students are given below:

Group –I: 61, 64, 59, 53, 56, 52, 57, 54, 51, 57, 60, 62

Group-II: 60, 58, 57, 64, 67, 70, 63, 62, 52, 53, 55, 56

Test by applying median test if the two groups differ significantly

Solution: After arranging the combined data of two groups in ascending order of magnitude, we get:

51, 52, 52, 53, 53, 54, 55, 56, 56, 57, 57, 57, 58, 59, 60, 60, 61, 62, 62, 63, 64, 64, 67, 70

Combined median = Average 12^{th} and 13^{th} terms in the ordered data = (57 + 58)/2 = 57.5

The values above and below the median in the two groups are determined and put in 2 x 2 table as:

	Group-I	**Group-II**	**Total**
No. of scores > Median	5	7	12
No. of Scores < Median	7	5	12
	12	12	24

$$\chi^2_{cal} = \frac{N(ad - bc)^2}{R_1 R_2 C_1 C_2} = \frac{24[25 - 49]^2}{12 \times 12 \times 12 \times 12} = \frac{24 \times 24 \times 24}{12 \times 12 \times 12 \times 12} = 0.67$$

Since χ^2_{cal} is less than tabulated value of χ^2 distribution with 1 d.f. at 5% level (3.84), therefore, H_0 is not rejected and we conclude that the samples are drawn from the populations with the same median.

2. Kruskal-Wallis H-Test:

To test whether several independent samples have come from identical populations, analysis of variance is the usual procedure provided the assumptions underlying are fulfilled. But if the assumptions are not fulfilled then Kruskalï Wallis test (Kruskal and Wallis, 1952) is used for one way classification data (i.e. completely randomized design). It is an extension of Mann Whitney U-test in which only two independent samples are considered.

Kruskal Wallis test is an improvement over the median test. In the median test the magnitude of various observations was compared with median value only. But in this method the magnitude of observations is compared with every other observation by considering their ranks.

So, if $n_1, n_{2,}\cdots\cdots,n_k$ denote the sample sizes of ðkô samples selected from k-populations and $N = n_1+n_{2,}\cdots\cdots,n_k$ denote the total number of observations in ðkô samples, then combined sample of ðNô observations in increasing order of magnitude will have ranks from 1 to N. Clearly the sum of ranks will be equal to N (N+1)/2. So under the null hypothesis H_0 : All population are identical vs H_1: Populations are not identical.

Let $R_1, R_{2,}\cdots\cdots,R_k$ denote the sum of ranks of the observations of 1st, 2nd, kth sample respectively. Then Kruskal-Wallis H- statistic is given as under:

$$H = \left(\frac{12}{N(N+1)}\right)\left[\frac{R_1^2}{n_1}+\frac{R_2^2}{n_2}+.....+\frac{R_k^2}{n_k}\right]-3(N+1)$$

$$= \left(\frac{12}{N(N+1)}\right)\sum_{i=1}^{k}\frac{R_1^2}{n_i}-3(N+1)$$

Under H_0, the sampling distribution of ðHô statistic is approximately chi-square with ðk-1ô degrees of freedom. Hence reject H_0 if calculated value of H is greater than or equal to critical value of χ^2 with (k-1) df and at a given level ðŬô otherwise H_0 is not rejected.

Example-12: Sample of three brands of cigarettes were tested for nicotine content. The observations (in mg) are:

Brand 1	7.7, 7.9, 8.5, 8.0, 8.4, 9.1
Brand 2	9.2, 8.6, 9.5, 8.7
Brand 3	8.9, 8.6, 7.8, 10.1, 10.0

Use Kruskal-Wallis Test to test if nicotine contents in three brands are equal.

Solution:

H_0 : The nicotine contents in the three brands are equal

H_1 : The nicotine contents in the three brands are not equal

Here $n_1 = 6$, $n_2 = 4$, $n_3 = 5$ so that N = 15

The ranks of corresponding observation are

Brand 1 1, 3, 6, 4, 5, 11

Brand 2 12, 7.5, 13, 9

Brand 3 10, 7.5, 2, 15, 14

Note: Two identical values (8.6) have been given the average rank (7 + 8)/2 = 7.5. The rank totals are $R_1 = 30$, $R_2 = 41.5$ and $R_3 = 48.5$

The value of test statistic H is given by:

$$H_{cal} = \frac{12}{15 \text{ x } 16}\left(\frac{30^2}{6} + \frac{41.5^2}{4} + \frac{48.5^2}{5}\right) - 3 \text{ x } 16$$

= 0.05 (150 + 430.06 + 470.45) ï 48 = 4.56

The tabulated value of χ^2 with 2 df at 5% level is 5.99

Since $H_{cal} < \chi^2_{tab}$, therefore, we do not reject H_0 and conclude that nicotine contents in three brands are equal.

Example-13: To test whether four types of treatments differ or not, the treatments were applied to 25 pods randomly. The green pod yield (kg) under four treatments was as under:

Treatments	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I	3.20	3.35	3.56	2.87	3.89	4.20	3.65
II	3.44	2.88	2.95	3.26	3.98	3.87	3.22
III	3.15	2.66	3.06	2.75	3.45		
IV	2.42	2.30	2.55	2.85	3.05	2.42	

Test the hypothesis that there is no significant difference among the four treatments by (i) Median test (ii) Kruskal Wallis test.

Solution: i) H_0 : Medians of all the treatments are equal.

H_1 : Medians of atleast two treatments are significantly different.

The sequence of combined samples in order of magnitude is

$$\underset{1}{[2.30]} \quad \underset{2.5}{[2.42]} \quad \underset{2.5}{[2.42]} \quad \underset{4}{[2.55]} \quad \underset{5}{(2.66)} \quad \underset{6}{(2.75)} \quad \underset{7}{[2.85]}$$

$$\underset{8}{2.87} \quad \underline{\underset{9}{2.88}} \quad \underline{\underset{10}{2.95}} \quad \underset{11}{[3.05]} \quad \underset{12}{(3.06)} \quad \underset{13}{(3.15)} \quad \underset{14}{3.20} \quad \underline{\underset{15}{3.22}}$$

$$\underline{\underset{16}{3.26}} \quad \underset{17}{3.35} \quad \underline{\underset{18}{3.44}} \quad \underset{19}{(3.45)} \quad \underset{20}{3.56} \quad \underset{21}{3.65} \quad \underline{\underset{22}{3.87}} \quad \underset{23}{3.89}$$

$$\underline{\underset{24}{3.98}} \quad \underset{25}{4.20}$$

Here k = 4, $n_1 = 7$, $n_2 = 7$, $n_3 = 5$, $n_4 = 6$ and N = 25

Median of the combined sample (M) = 3.15

Treatment	**Observed frequency**		**Total**
	<M	**≥ M**	
I	1	6	7
II	2	5	7
III	3	2	5
IV	6	0	6
Total:	**12**	**13**	**25**

Calculation of Expected Frequencies:

Treatment	Category ï I (< M)		Category-II (ÓM)	
	Observed	Expected	Observed	Expected
I	1	$\frac{12}{25} \times 7 = 3.4$	6	$\frac{13}{25} \times 7 = 3.6$
II	2	$\frac{12}{25} \times 7 = 3.4$	5	$\frac{13}{25} \times 7 = 3.6$
III	3	$\frac{12}{25} \times 5 = 2.4$	2	$\frac{13}{25} \times 5 = 2.6$
IV	6	$\frac{12}{25} \times 6 = 2.8$	0	$\frac{13}{25} \times 6 = 3.2$

$$T = \sum_i \sum_j \frac{(Oi_j - E_{ij})^2}{E_{ij}} = \frac{(1-3.4)^2}{3.4} + \frac{(2-3.4)^2}{3.4} + \frac{(3-2.4)^2}{2.4} + \frac{(6-2.8)^2}{2.8} + \frac{(6-3.6)^2}{3.6}$$

$$+ \frac{(5-3.6)^2}{3.6} + \frac{(2-2.6)^2}{2.6} + \frac{(0-3.2)^2}{3.2} = 11.56$$

Test Statistic: $\frac{(N-1)}{N} T = \left(\frac{25-1}{25}\right)(11.56) = 11.10$ (Calculated value)

Since calculated value > 7.815 (tabulated value), therefore, null hypothesis is rejected. Thus median effect of various treatments is significantly different.

ii) The sum of ranks of each individual treatment is

For treatment I R_1 = 8 + 14 + 17 + 20 + 21 + 23 + 25 = 128

For treatment II R_2 = 9 +10 + 15 + 16 + 18 + 22 + 24 = 114

For treatment III R_3 = 5 + 6 + 12 + 13 + 19 = 55

For treatment IV R_4 = 1 + 2.5 + 2.5 + 4 + 7 + 11 = 28

Here $n_1 = 7,\ n_2 = 7,\ n_3 = 5,\ n_4 = 6$

Therefore $N = n_1 + n_2 + n_3 + n_4$

$= 7 + 7 + 5 + 6 = 25$

Applying Kruskal Wallis test, we have:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

$$H = \frac{12}{25 \text{ x } 26}\left[\frac{128^2}{7} + \frac{114^2}{7} + \frac{55^2}{5} + \frac{28^2}{6}\right] - 3 \text{ x } 26$$

$$= \frac{12}{25 \text{ x } 26} \text{ x } 4932.81 - 3 \text{ x } 26$$

= 91.067 ï 78 = 13.067

The tabulated value of chi-square $\chi^2_{3,0.05} = 7.815$

Since $H_{cal} > 7.815$, therefore, null hypothesis is rejected. Thus effect of various treatments is significantly different.

3. Friedman's Test:

This is a non-parametric test for k-related samples of equal size say n, parallel to two-way analysis of variance (Randomized Block Design). Here we have k-related samples of size n arranged in n blocks and k columns in a two-way table as given below:

Blocks	**Samples (Treatments)**							**Block Totals**
	1	2	3	..	..	..	k	
1	R_{11}	R_{12}	R_{13}	..	..	..	R_{1k}	k(k+1)/2
2	R_{21}	R_{22}	R_{23}	..	..	..	R_{2k}	k(k+1)/2
3	R_{31}	R_{32}	R_{33}	..	..	..	R_{3k}	k(k+1)/2
:								
:								
n	R_{n1}	R_{n2}	R_{n3}	..	..	..	R_{nk}	k(k+1)/2
Column totals	**R_1**	**R_2**	**..**	**..**	**..**	**..**	**R_k**	**nk(k+1)/2**

Where R_{ij} is the rank of the observation belonging to sample j in the i^{th} block for j = 1,2,·····,k and i = 1, 2,·····,n. This is the same situation in which there are k treatments and each treatment is replicated n times. Here it should be carefully noted that the observations in a block receive ranks from 1 to k. The smallest observation receives rank 1 at its place and the largest observations receive rank k at its place. Similarly observations receive ranks accordingly in other blocks. Hence the block total are constant and equal to k(k+1)/2, the sum of k integers.

The null hypothesis H_o to be tested is that all the k samples have come from identical populations. In case of experiment design, the null hypothesis H_o is that there is no difference between k treatments. The alternative hypothesis H_1 is that at least two samples (treatments) differ from each other.

Under H_o, the test statistic is:

$$T = \frac{12}{nk\,(k+1)} \sum_{j=1}^{k} R_j^2 - 3n\,(k+1)$$

The statistic T is distributed as χ^2 with (k-1) d.f. Reject H_o if $T_{cal} \geq \chi^2_{(k-1)}$, otherwise H_o is not rejected.

Example-14: The iron determination (ppm) in five pea-leaf samples, each under three treatments were as tabulated below:

Sample No. Blocks	**Treatments**			**Block Totals**
	1	2	3	
1	591 (1)	682 (2)	727 (3)	6
2	818 (2)	591 (1)	863 (3)	6
3	682 (2)	636 (1)	773 (3)	6
4	499 (1)	625 (2)	909 (3)	6
5	648 (1)	863 (3)	818 (2)	6
Column Totals	**7**	**9**	**14**	**30**

Test the hypothesis that iron content in leaves under three treatments is same.

H_o : That the iron content in leaves under three treatments is the same

H_1: That at least two of them have different effect

Friedmanôs test statistic is:

$$T_{cal} = \frac{12}{nk\,(k+1)} \sum_{j=1}^{k} R_j^2 - 3n\,(k+1)$$

$$= \frac{12}{5 \text{ x } 3 \text{ x } 4}(7^2 + 9^2 + 14^2) - 3 \text{ x } 5 \text{ x } 4$$

= 65.2 ï 60 = 5.2

For Ŭ= 0.05, the table value of $\chi^2_{0.05,\,2} = 5.99$

Since the calculated T-value is less than 5.99, H_o is not rejected. This means that there is no significant difference in the iron content of pea leaves due to treatments.

7.2 Kendall's rank correlation coefficient:

Kendallôs rank correlation coefficient (Kendall, 1938) τ is a non-parametric measure of correlation and is based upon the ranks of the observations.

Consider each possible pair of individuals (i, j) and the order of this pair in the two rankings. If the pair appears in the same order in both rankings, we allot it a score of +1, and if it appears in reverse orders, a score of -1. The score is thus obtained for each of the $\binom{n}{2} = \frac{n\,(n-1)}{2}$ possible pairs. We then define a rank correlation coefficient τ as

$$\tau = \frac{\text{total score}}{n\,(n-1)/2} \qquad (1)$$

Obviously, τ = +1 for perfect agreement, because the score for each pair, each being in the same order in both rankings, is +1 ; and τ = -1 for perfect disagreement, because the score for each pair, being in reverse orders in the two rankings, is now -1.

Suppose the ranking in one series is in the natural order, viz. the order 1, 2,·····,n. Let us consider the corresponding ranking in the other series. Suppose out of $\binom{n}{2}$ pairs for the second series, P pairs have ranks in the natural order and Q pairs have ranks in the reverse order. Obviously, the P pairs will receive a score of +1 each, while the Q pairs will receive a score of -1 each. Thus, according to (1)

Ű= $\frac{P-Q}{\binom{n}{2}} = 1 - \frac{2Q}{\binom{n}{2}} = \frac{2P}{\binom{n}{2}} - 1$ since P + Q =total number of pairs = $\binom{n}{2}$

Test of Significance:

If the absolute value of τ or S = P ï Q ≥ corresponding value in the table of critical values for Kendall Tau for a given n, then H_0 is rejected.

Example-15: Ten hand writings were ranked by two judges in a competition. The rankings are given below. Calculate Kendallôs τ coefficient and test for its significance.

	Hand-writing									
	A	B	C	D	E	F	G	H	I	J
Judge 1	3	8	5	4	7	10	1	2	6	9
Judge 2	6	4	7	5	10	3	2	1	9	8

Solution: To calculate τ, it is convenient to rearrange one set of ranking so as to put it in the natural order : 1, 2,·····,n. If we do so far the ranking by Judge 1, the corresponding ranking by Judge 2 becomes:

2, 1, 6, 5, 7, 9, 10, 4, 8 and 3

The score obtained by considering the first member 2, in conjunction with the others is 8-1 = 7, because only 1 is smaller than 2. Similarly, the score involving the member 1 is 8, the score involving the member 6 is 4-3=1, and so on. The total score is

S = 7 + 8 + 1 + 2 + 1 ï 2 ï 3 + 0 ï 1= 19-6 = 13

On the other hand, the maximum possible score is (10 x 9)/2 = 45.

Thus τ = 13/45 = 0.289

Test of Significance:

The critical values of τ and S for n = 10 at 5% level of significance are 0.511 and 23 respectively. Since the calculated value of τ and S are less than the critical values, therefore, H_0 is not rejected.

Example-16: Two supervisors ranked 12 workers working under them in order of their efficiency as given below. Calculate Kendallôs τ coefficient between the two rankings and test for its significance.

Worker	1	2	3	4	5	6	7	8	9	10	11	12
Supervisor (X)	5	6	1	2	3	8.5	8.5	4	7	11	10	12
Supervisor (Y)	5.5	5.5	2	2	2	9	7	4	8	10.5	12	10.5

Solution: We rearrange the ranking of supervisor X in the natural order, and then we have the two sets of ranks as follows:

Supervisor (X)	1	2	3	4	5	6	7	8.5	8.5	10	11	12
Supervisor (Y)	2	2	2	4	5.5	5.5	8	9	7	12	10.5	10.5

The total score in this case is

$$S = 9 + 9 + 9 + 8 + 6 + 6 + 3 + 3 + 3 - 2 \text{ ï } 0 = 54$$

Maximum possible score = n (n-1)/2 = 12(11)/2=66

$$\text{Hence } \tau = \frac{54}{66} = 0.818$$

The critical values of τ and S (for n = 12 at α = 0.05) are 0.455 and 30 respectively. Since calculated values of τ and S are more than critical values, hence H_0 is rejected and we conclude that there is significant correlation between ranking of two supervisors.

Kendall's Coefficient of Concordance:

Kendall developed coefficient of concordance (K_c) for measuring the relationship between the k variates. It measures the extent of relationship (or the degree of association) between ók ô variates based on ón ô rankings for each variate. This coefficient avoids the requirement of computing several Spearman rank correlation coefficients pairwise. The two way table of ranks for the variates and the observations are below:

(Ranks)

Observation / **Variate**	**1**	**2**	**. . . .**	**n**
1	6	(n ï 1)	. . . .	2
2	3	5	. . . .	n-2
.	.	.		.
.	.	.		.
.	.	.		.
k	(n - 5)			1
Total	C_1	C_2		C_n

The following of concordance is given by:

$$K_c = \frac{\sum_{j=1}^{n} C_j^{\,2} - \frac{(\sum C_j)^2}{n}}{\frac{1}{12} k^2 n (n^2 - 1)}$$

where C_j denote the j-th column total for j = 1, 2,·····,n and n be the number of observations in each variate and k be the number of variates. K_c always lies between 0 and 1 i.e. $0 \le K_c \le 1$.

Test of significance (for large samples i.e. n > 7):

If the number of observations for each of the variate is greater than 7, Chi-square approximation is used for testing the significance of coefficient of concordance in the population.

To test the hypothesis H_0: There exists no correlation between k variates based on ranks, the test statistic χ^2 is given by:

$$\chi^2 = k (n-1) K_c$$

Conclusion: If $\chi^2_{cal} \ge \chi^2_{tab}$ with (n-1) d.f. at chosen Ũ, the null hypothesis is rejected. Otherwise, the null hypothesis is accepted.

Example-17: In a certain cattle judging competition, 10 cows were ranked by 4 judges (A, B, C and D) and ranks are given below. Compute the Kendallôs coefficient of concordance between the rankings of 4 judges and test for its significance.

	Cow									
Judge	1	2	3	4	5	6	7	8	9	10
A	5	3	4	1	8	7	6	9	10	2
B	4	6	3	2	7	5	10	8	9	1
C	4	7	5	3	6	9	8	10	2	1
D	3	5	1	4	9	10	7	6	8	2
Total (e_j)	16	21	13	10	30	31	31	33	29	6

Therefore, $\sum C_j = 220 \quad \sum C^2_j = 5754$

$$K_c = \frac{\sum_{j=1}^{n} C_j^{\,2} - \frac{(\sum C_j)^2}{n}}{\frac{1}{12} k^2 n (n^2 - 1)} = \frac{5754 - \frac{(220)^2}{10}}{\frac{1}{12}(4)^2 10 (100 - 1)} = 0.69$$

Test of Significance:

To test the null hypothesis H_0 : There is no correlation between the ranking of 4 judges.

The value of test statistic is given by

$$\chi^2_{cal} = k (n - 1) K_c = 4 (10 - 1) (0.69) = 24.84$$

Conclusion: Here χ^2_{cal} is more than critical value of χ^2 (= 16.92) with 9 d.f. at 5% level of significance. Hence, there exists significant correlation between the rankings of 4 judges.

EXERCISES

1. A manufacturer of electric bulbs claims that he has developed a new production process which will increase the mean lifetime (00 hours) of bulbs from the present value 11.03. The results obtained from 15 bulbs taken at random from the new process are given below:

   11.29, 12.15, 10.69, 13.25, 13.47, 11.76, 14.05, 14.38, 11.08, 12.25, 10.93, 11.02, 12.87, 12.00, 13.56.

   Can it be concluded that mean life time of bulbs has increased by applying (i) Sign test (ii) Wilcoxon Signed Ranks Test.

2. Following is the arrangement of 25 men (m) and 15 women (w) lined up to purchase ticket for a picture show:

   m ww mmm w m w m w m w m ww

   mmm w mm ww mmmmm w m www mmmmmm

   Test randomness at 5% level of significance.

3. Nine animals were tested under control and experimental conditions and the following values were observed. Test whether there is significant increase in the values under experimental conditions by (i) Paired Sign Test (ii) Paired Sample Wilcoxon Signed Rank Test

Animal No.	1	2	3	4	5	6	7	8	9
Control	21	24	26	32	55	82	46	55	88
Experimental	18	27	35	42	82	99	52	30	62

4. Six students went on a diet in an attempt to loose weight with the following results

Student	A	B	C	D	E	F
Weight before (lbs)	174	191	188	182	201	188
Weight after (lbs)	165	186	183	178	203	181

   Is the diet an effective means of loosing weight?

5. A sociologist asked twenty couples in the age group of 25-30 years regarding the number of children they planned to have. The answers of male and females were as follows:

Couple	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Male	3	2	1	2	3	4	3	0	2	3	5	6	4	2	1	2	3	5	6	3
Female	2	1	1	2	2	1	0	0	4	2	4	6	4	1	2	2	3	6	2	0

Apply sign test to conclude whether the attitude of male and female differ significantly on this issue.

6. A professor of sociology developed an intelligence test which he gave to two groups of 20 years old and 50 years old persons. The scores obtained were recorded in following table.

20 years old	**50 years old**
130	120
130	125
147	130
138	140
140	129
152	140
142	155
137	127
150	
140	

Perform the Mann-Whitney U-test to test the null hypothesis that average intelligence of both age groups is equal at 5% level.

7. Following are the test scores of 15 students of section I and 15 students of section II taught by two different professors. Test whether there is any significant difference in teaching method of two teachers by (i) Mann Whitney U-test (ii) Median test.

Roll No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Scores of Section I	81	83	72	78	70	85	62	95	86	71	69	80	91	81	77
Scores of Section II	76	80	76	86	77	92	65	85	82	75	64	80	95	86	66

8. Yields from two varieties of wheat V_1 and V_2 sown on 6 and 7 identical plots are given below:

Variety	**Yield (q/ha)**						
V_1	40	35	52	60	46	55	
V_2	47	56	42	57	50	57	50

Use Mann-Whitney U test to test whether the two varieties have identical yields.

9. In order to compare the two varieties of maize, the following yields (kg) were recorded for ten identical plots under each variety:

Variety A	32.1	2.6	17.8	28.4	19.6	21.4	19.9	3.1	7.9	25.7
Variety B	19.8	27.6	23.7	9.9	3.8	27.6	34.1	18.7	16.9	17.9

i) Use Mann-Whitney U test to test whether the two varieties gave equal yields.

ii) Use Siegal-Tukey test to test whether the two maize varieties have the equal variability in yield.

10. To test the I.Q. of young and old people, an intelligence test was administered to two groups of 25 year old and 50 year old persons. The scores obtained by both groups of persons were obtained as given below:

	Scores									
25 years old	130	135	148	139	141	153	142	138	151	141
50 years old	122	126	131	142	130	141	156	128		

Using Mann-Whitney U-test, test whether the average performance of both age groups is same or not?

Chapter-VIII

INTRODUCTION TO MULTIVARIATE ANALYSIS

Multivariate statistical analysis is appropriate whenever several responses are measured on each object or experimental unit. Univariate analysis applied separately to each response leads to incorrect conclusions, since responses measured on the same object are generally correlated. Multivariate analysis can be simply defined as the application of statistical methods that deal with reasonably large number of characteristics or variables recorded on each object in one or more samples simultaneously. It provides statistical tools for the study of joint relationships of variables in data that contains intercorrelations. In other words, multivariate analysis differs from univariate and bivariate analysis in that it directs attention away from the analysis of the mean and variance of a single variable or from the pairwise relationship between two variables, to the analysis of the co-variances or correlations which reflect the extent of relationship among three or more variables. For example, a biometrician concerned with developing a taxonomy for classifying species of fowl on the basis of anatomical measurements may collect information on skull length, skull width, humerus length and tibia length.

Remarks:

1. The term objects in multivariate analysis refer to things, persons, individuals, events or in general entities on which the measurements are recorded. And the measurements relate to characteristics or attributes of the objects that are being recorded and in general are called variables.
2. Multivariate analysis investigates the dependency not only amongst the variables but also among the individuals on which observations are made.

8.1 Data Cube and Data Matrices:

In multivariate analysis, a researcher generally deals with data collected from ónô individuals, on ópô characters recorded over ólô locations or periods or groups. Thus, the basic input can be considered in terms of a data cube denoted by x_{ijk}, where

i = 1, 2, é .., n refers to objects

j = 1, 2, é .., p refers to characteristics/variables or attributes

k =1, 2, é .., L refers to location/periods

The data cube can be given a two dimensional representation by writing matrices within matrices as follows:

		Characteristics			
Location/Period	**Objects**	X_1	X_2	…..	X_p
I	O_1	x_{111}	x_{121}	…..	x_1p_1
	O_2	x_{211}	x_{321}	…..	x_2p_1
	: :				
	O_n	x_{n11}	x_{n21}	…..	x_np_1
II	O_1	x_{112}	x_{122}	…..	x_1p_2
	O_2	x_{212}	x_{222}	…..	x_2p_2
	:				
	O_n	x_{n12}	x_{n22}	…..	x_np_2
	…..	…..	…..	…..	…..
L^{th}	O_1	x_{11L}	x_{12L}	…..	x_1p_L
	O_2	x_{21L}	x_{22L}	…..	x_2p_L
	:	:	:		:
	O_n	x_{n1L}	x_{n2L}	…..	x_{npL}

In many applications, the analyst considers only two coordinates of the data cube, so that the basic input becomes a data matrix or rectangular array of numerical entities. This may result from collecting information only on one occasion/location or groups. In such situations, the data matrix has ónô rows and ópô columns and can be described in terms of the elements x_{ij}, where i refers to objects and j refers to attributes.

A data matrix as ónô individuals recorded for ópô characters/variable can be described as:

Objects	X_1	X_2	…	X_p
1	x_{11}	x_{12}	é	x_1p
2	x_{21}	x_{22}	é	x_2p
: :	: :	: :	: :	: :
n	x_{n1}	x_{n2}	é	x_{np}

or simply

$$x = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

8.2 Descriptive Measures in Multivariate Analysis:

Let X be a random vector of p-components and denoted by:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad \text{or} \quad X' = [x_1, x_2, ..., x_p]$$

The mean vector (μ) covariance matrix ($\mathbf{\Sigma}$) and correlation matrix ($\boldsymbol{\rho}$) are given by:

$$\boldsymbol{\mu} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \qquad \sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

$$\text{and } \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

where $\mu_i = E(X_i)$, $\mathring{u}_i^2 = V(X_i)$, $\mathring{u}_{ij} = Cov(X_i, X_j)$ and $\rho_{ij} = corr(X_i, X_j) = \dfrac{Cov(x_i, x_j)}{\sqrt{v(x_i)\, v(x_j)}}$

The mean vector ($\boldsymbol{\mu}$), covariance matrix ($\mathbf{\Sigma}$) and the correlation matrix ($\boldsymbol{\rho}$) given above respectively represent the measures of central tendency, dispersion and linear association for the p-dimensional multivariate population. The sample estimates for these measures may be obtained as follows:

Let **x** denote an n x p data matrix, where n is the number of observations and p is the number of variables. Then sample mean vector denoted by $\bar{\mathbf{x}}$ is given by:

$$\bar{\mathbf{x}}' = \frac{1}{n} 1' X = \frac{1}{n}(1,1,.....,1)\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$= \frac{1}{n}\left[\sum x_{i1} \quad \sum x_{i2} \quad ..... \quad \sum x_{ip}\right]$$

$$= \left[\bar{x}_1 \quad \bar{x}_2 \quad ..... \quad \bar{x}_p\right]$$

The sample covariance matrix: $\mathbf{S} = \frac{1}{n-1}\mathbf{X_d'X_d}$ where $\mathbf{X_d} = \mathbf{X - 1}\,\bar{X}'$ is the matrix of mean corrected scores and the matrix $x_d' x_d$ is often called the corrected sum of squares and product matrix.

The sample correlation matrix is usually denoted by $\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$ can be obtained by $\mathbf{R = D'SD}$ where **D** denote the diagonal matrix whose entries along the main diagonal are the reciprocals of the standard deviation of the variables in x, i.e.

$$\mathbf{D} = \begin{bmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{s_p} \end{bmatrix} \text{ where } s_j^{\,2} = \frac{1}{n-1}\sum (x_{ij} - \bar{x}_j)^2$$

Example-1: Suppose the data matrix $\mathbf{x} = \begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix}$ be a multivariate sample of size n=4 from a trivariate population.

Then sample mean vector:

$$\bar{\mathbf{x}}' = \frac{1}{4}[1,1,1,1]\begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix} = \frac{1}{4}[12 \;\; 16 \;\; 20] = [3 \;\; 4 \;\; 5]$$

Now the mean corrected score matrix $\mathbf{x}_d$ is computed as:

$$\mathbf{X}_d = \mathbf{X} - \mathbf{1}\bar{\mathbf{X}}' = \begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [3 \;\; 4 \;\; 5]$$

$$= \begin{bmatrix} 2 & 3 & 3 \\ 3 & 2 & 6 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix} - \begin{bmatrix} 3 & 4 & 5 \\ 3 & 4 & 5 \\ 3 & 4 & 5 \\ 3 & 4 & 5 \end{bmatrix} = \begin{bmatrix} -1 & -1 & -2 \\ 0 & -2 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\therefore \mathbf{S} = \frac{1}{n-1}\mathbf{X}_d'\mathbf{X}_d = \frac{1}{4-1}\begin{bmatrix} -1 & 0 & 1 & 0 \\ -1 & -2 & 2 & 1 \\ -2 & 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & -1 & -2 \\ 0 & -2 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix} 2 & 3 & 2 \\ 3 & 10 & 1 \\ 2 & 1 & 6 \end{bmatrix} = \begin{bmatrix} 2/3 & 1 & 2/3 \\ 1 & 10/3 & 1/3 \\ 2/3 & 1/3 & 2 \end{bmatrix} = \begin{bmatrix} 0.67 & 1 & 0.67 \\ 1 & 3.33 & 0.33 \\ 0.67 & 0.33 & 2 \end{bmatrix}$$

$$\text{Now } \mathbf{D} = \begin{bmatrix} 1/s_1 & 0 & 0 \\ 0 & 1/s_2 & 0 \\ 0 & 0 & 1/s_3 \end{bmatrix} = \begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix}$$

$$\therefore \mathbf{R} = \mathbf{D}'\mathbf{S}\mathbf{D} = \begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix}\begin{bmatrix} 2/3 & 1 & 2/3 \\ 1 & 10/3 & 1/3 \\ 2/3 & 1/3 & 2 \end{bmatrix}\begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{2/3} & \sqrt{3/2} & \sqrt{2/3} \\ \sqrt{3/10} & \sqrt{10/3} & \frac{1}{\sqrt{30}} \\ \sqrt{2/3} & \frac{1}{\sqrt{18}} & \sqrt{2} \end{bmatrix}\begin{bmatrix} \sqrt{3/2} & 0 & 0 \\ 0 & \sqrt{3/10} & 0 \\ 0 & 0 & \sqrt{1/2} \end{bmatrix} = \begin{bmatrix} 1 & 3/\sqrt{20} & 1/\sqrt{3} \\ 3/\sqrt{20} & 1 & 1/\sqrt{60} \\ 1/\sqrt{3} & 1/\sqrt{60} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.67 & 0.58 \\ 0.67 & 1 & 0.13 \\ 0.58 & 0.13 & 1 \end{bmatrix}$$

Important Multivariate Methods:

Multivariate data recorded on a large number of interrelated variables is often difficult to interpret. Therefore, there is a need to condense and sum up the essential features of the data through dimension reduction or some appropriate summary statistics for better interpretation. As a broad classification, the multivariate techniques may be classified as dependence methods and interdependence methods.

The methods in which one or more variables are dependent and others are independent are called dependant techniques. Multivariate regression, multivariate analysis of variance, discriminant analysis and canonical correlation analysis are the notable dependence techniques.

If interest centres on the mutual association across all the variables with no distinction made among the variable types, then such techniques are called interdependence techniques. Principal component analysis, factor analysis, cluster analysis and multi-dimensional scaling are the important interdependence techniques.

Multivariate Regression: It is concerned with the study of the dependence of one or more variables on a set of other variables called independent variables with the objective to estimate or predict the mean values of the dependent variables on the basis of the known values of the independent variables. If there is only one dependent variable and many independent variables, then it is known as multiple regression.

Multivariate Analysis of Variance: It is simply a generalization of univariate analysis of variance, where the primary objective is on testing for significant differences on a set of variables due to changes in one or more of the controlled (experimental) variables.

Discriminant Analysis: It is used to find linear combinations of the variables that separate the groups. Given a vector of p observed scores, known to belong to one of two or more groups, the basic problem is to find some function of the p scores (i.e. a linear combination) which can accurately assign individual with a given score into one of the groups.

Principal Component Analysis: It is a dimension reduction technique where the primary goal is to construct orthogonal linear combinations of the original variables that account for as much of the total variation as possible. The successive linear combinations

are extracted in such a way that they are uncorrelated with each other and account for successively smaller amounts of total variation.

Cluster Analysis: The purpose of cluster analysis is to reduce a large data set to meaningful subgroups of individuals or objects. The division is accomplished on the basis of similarity of the objects across a set of specified characteristics. The individuals of a particular subgroup or cluster are, in some sense, more similar to each other than to elements belonging to other groups.

Canonical Correlation Analysis: The most flexible of the multivariate technique, canonical correlation simultaneously correlates several explanatory variables and several dependent variables. In usual sense, it determines the linear association between a set of dependent variables and a set of explanatory variables. In canonical analysis, we find two linear combinations, one for the predictor set of variables and one for the set of explanatory variables, such that their product moment correlation is maximum.

8.4 Testing Significance of Mean Vector (One Sample Case): This is useful for multivariate populations where it is required to test whether the population mean vector (μ) is equal to a specified mean vector (μ_0).

Let $\mathbf{x}_1, \mathbf{x}_2$, é .., $\mathbf{x}_n$ be a random sample of size n from a p-dimensional multivariate population having mean vector μ and covariance matrix $\mathbf{\Sigma}$,

H_0: $\mu = \mu_0$

H_1: μ Í μ_0

If covariance matrix Ɇ is known or sample is large then a χ^2 test is used to test the above hypothesis. If $\overline{\mathbf{X}}$ denotes the sample mean vector then the statistic $\chi^2 = n\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \sum^{-1} \left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)$ follows a chi-square distribution with p degrees of freedom and we reject H_0 if $\chi^2_{cal} > \chi^2(p)$ at desired Ŭ level of significance.

If the covariance matrix $\mathbf{\Sigma}$ is not known and sample size is small, then Hotelling T^2 (defined below) is used for testing H_0.

$T^2 = n\left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)' S^{-1} \left(\overline{\mathbf{X}} - \boldsymbol{\mu}_0\right)$, where $\mathbf{S}$ is the sample covariance matrix. The sampling distribution of Hotelling T^2 is given as:

$\frac{(n-p)}{(n-1)p}T^2 \approx F_{(p,\ n-p)}$ distribution or $T^2 \approx \frac{(n-1)p}{n-p}F_{(p,\ n-p)}$ distribution.

Therefore, we reject H_0 if $T^2_{cal} > \frac{(n-1)p}{n-p}F_{(p,\ n-p)}$ at Ŭlevel of significance.

8.5 Testing Equality of Two Mean Vectors (Two sample case): Consider two independent samples of sizes n_1 and n_2 from two p-variate normal populations with mean vectors μ_1 and μ_2 and having same but known covariance matrix $\mathbf{\Sigma}$.

H_0: $\mu_1 = \mu_2$

H_1: μ_1 Í μ_2

Test statistic is:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)' \mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) \text{follows}$$

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F(p, n_1 + n_2 - p - 1) \text{ distribution.}$$

where $\overline{\mathbf{X}}_1$ *and* $\overline{\mathbf{X}}_2$ are the sample mean vectors and S_p is the pooled covariance matrix defined by $\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$, $\mathbf{S}_1$ and $\mathbf{S}_2$ being the individual sample covariance matrices.

Example-2: Mean vector and covariance matrix for a sample of size 20 from a trivariate population are found to be:

$$\overline{\mathbf{X}} = \begin{bmatrix} 4.64 \\ 45.4 \\ 9.94 \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} 2.88 & 10.01 & -1.81 \\ 10.01 & 199.79 & -5.64 \\ -1.81 & -5.64 & 3.63 \end{bmatrix}$$

We wish to test the hypothesis that the null hypothesis H_0 : $\mu = \mu_0$ where $\mu_0 = \begin{bmatrix} 4 \\ 50 \\ 10 \end{bmatrix}$ against H_1 : μ Í μ_0.

$$T^2 = n(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}_0)$$

$$= 20\begin{bmatrix}4.64-4 & 45.4-50 & 9.97-10\end{bmatrix}\begin{bmatrix}2.88 & 10.01 & -1.81\\ 10.01 & 199.79 & -5.64\\ -1.81 & -5.64 & 3.63\end{bmatrix}^{-1}\begin{bmatrix}4.64-4\\ 45.4-50\\ 9.97-10\end{bmatrix} \simeq 9.7$$

For p = 3 and n = 20 and $\alpha = 0.05, \frac{(n-1)p}{n-p} F_{\alpha(p,n-p)} = 10.7$

Since T^2_{cal} is less than 10.7, we do not reject H_0.

Example-3: Consider the two samples of size n_1 = 45 and n_2 = 55 with $\overline{\mathbf{X}}_1 = \begin{bmatrix}204.4\\ 556.6\end{bmatrix}$, $\overline{\mathrm{X}}_2 = \begin{bmatrix}130\\ 355\end{bmatrix}$

$\mathbf{S}_1 = \begin{bmatrix}13825.3 & 23823.4\\ 23823.4 & 73107.4\end{bmatrix}$ and $\mathbf{S}_2 = \begin{bmatrix}8632 & 19616.7\\ 19616.7 & 55964.5\end{bmatrix}$

Test H_0: $\mu_1 = \mu_2$ against H_1: μ_1 Í μ_2

After simplification, we get:

$$\mathbf{S}_p = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2} = \begin{bmatrix}10963.7 & 21505.5\\ 21505.3 & 63661.3\end{bmatrix}$$

and $T^2_{cal} = \frac{n_1 n_2}{n_1+n_2}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)' \mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) = 16.2$

Comparing the calculated T^2 value with the critical value at Ŭ= 0.05 that is

$$\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{(p,n_1+n_2-p-1)} = \frac{98 \text{ x } 2}{97} F_{0.05}(2,97) = 6.26$$

Since $T^2_{cal} > 6.26$, therefore, we reject H_0.

Note:

1. Mahalanobis D^2 defined below can also be used for testing the above hypothesis:

$$D^2 = \left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)' \mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) = \frac{n_1 n_2}{n_1+n_2} T^2$$

2. If the two populations do not have same covariance matrices, then above test cannot be used. However, for large samples or when dispersion matrices are known, χ^2 test can be used.

Annexure-I

Model Test Paper for Mid-Term Examination
Stat-511 (Statistical Methods for Applied Sciences)

Time: 1½ Hrs. **M. Marks: 30**

Note: Attempt all questions.

Q.1 (a) Define any five:

(i) Random Variable (ii) Ogives (iii) Poisson Distribution (iv) Standard Error (v) Skewness (vi) Coefficient of Variation (vii) Conditional Probability

(b) A problem is given to two students A and B. The odds in favour of A solving the problem are 6 to 9 and against B solving the problem are 12 to 10. What is the probability that the problem will be solved if they both try independently?

(c) What is Box-Plot technique for description of data? (5+3+2)

Q.2 (a) Define Negative-Binomial Distribution alongwith its properties.

(b) Explain Normal Distribution and Standard Normal Distribution. How it is so important in statistics?

(c) If on an average 20% of the bolts manufactured by a company are defective, then find the probability that in a random sample of 400 bolts, the number of defective bolts are (i) between 76 and 82 (ii) atleast 75. (3+3+4)

Q.3 (a) Find the 95% confidence interval for the normal population mean if a random sample of 16 is drawn whose mean comes to 12.5 and sum of squares of deviations from mean is equal to 93.75.

(b) Find the Binomial Distribution for which mean = 4 and variance = 3.

(c) Given below are the weights of 60 apples in a box:

Weight (gms)	150-180	180-210	210-240	240-270	270-300
No. of Apples	5	10	20	15	10

Find the mean and variance of the distribution by shortcut method. (4+2+4)

Model Test Paper for Final (Theory) Examination
Stat-511 (Statistical Methods for Applied Sciences)

Time: 2½ Hrs. **M. Marks: 45**

Note: Attempt any five questions.

Q.1 (a) Explain the following:

(i) Steps for testing of hypothesis

(ii) Level of significance and Critical Region

(iii) Properties of Regression Coefficient

(b) Give the procedure of Fisher Z-transformation for testing the null hypothesis $H_0: \rho = \rho_0 (\rho_0 \neq 0)$ (6+3)

Q.2 (a) How will you test the significance of difference between two population means when population variances are known?

(b) Given below the increases in weights (lbs) of 10 and 12 cattle fed on diets A and B, respectively. Test at 5% level of significance that there is no difference between two diets: (4+5)

	Increase in Weight (lbs)											
Diet A	10	6	16	17	13	12	8	14	15	9		
Diet B	7	13	22	15	12	14	18	8	21	23	10	17

Q.3 (a) How will you test the significance of observed simple, partial and multiple correlation coefficients?

(b) Given the sample correlation coefficients $r_{12} = 0.7$ $r_{23} = 0.5$ and $r_{13} = 0.4$. Find the value of partial correlation coefficient $r_{32.1}$. (5+4)

Q.4 (a) Explain least squares method of curve fitting and use the technique to fit the quadratic curve $y = a + bx + cx^2$

(b) Construct 95% confidence interval for population mean μ given that $n = 20$, $\sum x_i = 400$, $\sum (x_i - \bar{x})^2 = 76$ (5+4)

Q.5 Write short notes on any three of the following:

(i) Box Plot (ii) Kendall Coefficient of Concordance (iii) Run Test for Randomness (iv) Advantages and Disadvantages of Non-Parametric Tests (3+3+3)

Q.6 The quantity of serum albumin (gms) per 100 ml in lepers under four different drugs was recorded as follows:

	Treatments			
Sample No. (Block)	Drug-I	Drug-II	Drug-III	Drug-IV
1	4.30	3.65	3.05	3.90
2	4.00	3.60	4.10	3.10
3	4.10	2.70	4.20	3.20
4	3.80	3.15	3.70	4.20
5	3.30	3.75	3.60	3.00
6	4.50	2.95	4.80	3.40

Apply Friedmanôs test to confirm whether the contents of serum albumin is same in different treatments. (9)

Model Test Paper for Final (Practical) Examination
Stat-511 (Statistical Methods for Applied Sciences)

Time: 1 Hrs. **M. Marks: 20**

Note: Attempt any two questions.

Q.1 The grain pod yield (kg) under four different treatments in an experiment is given below:

	Treatments			
Pod No.	T_1	T_2	T_3	T_4
1	4.18	4.45	4.16	3.49
2	4.41	3.89	3.70	3.38
3	4.51	3.98	4.11	3.59
4	3.88	4.28	3.81	3.85
5	4.89	4.95	4.46	4.01
6	5.01	4.88		3.49
7	4.61	4.26		

Test the hypothesis that there is no significant difference among the treatments affecting the yield by Kruskal-Wallis Test? (10)

Q.2 The ages and blood pressures of nine men are given below:

Age (x)	55	41	36	47	49	42	60	72	63
B.P. (y)	142	124	117	127	144	138	154	157	148

Find the regression equation of y on x and estimate the blood pressure of a man of 40 years. (10)

Q.3 (a) Find the Spearmanôs coefficient of rank correlation between performance in Economics and Statistics:

Student	1	2	3	4	5	6	7	8
Marks in Economics	25	30	38	22	50	70	30	90
Marks in Statistics	50	40	60	40	30	20	40	70

(b) From the following table regarding eye colour of mother and son, test if the colour of sonôs eye is associated with that of mother:

Mother's eye colour	**Son's eye colour**	
	Light Blue	**Not light blue**
Light Blue	45	16
Not Light Blue	4	35

(5+5)

Revision Exercises for Mid-Term Examination
Stat-511 (Statistical Methods for Applied Sciences)

Q.1 **Define the following:**

(i) Coefficient of Variation (ii) Conditional Probability (iii) Skewness (iv) Random Variable (v) Poisson Distribution (vi) Standard Error (vii) Ogives (viii) Pooled Variance (ix) Mean Deviation (x) Harmonic Mean

Q.2 (a) One problem is given to two students A and B. The odds in favour of A solving the problem are 6 to 9 and against B solving the problem are 12 to 10. What is the probability that the problem will be solved if they both try independently?

(b) What is Box-Plot technique for description of data?

Q.3 (a) Given below are the weights of 60 apples in a box:

Weight (gms)	150-180	180-210	210-240	240-270	270-300
No. of Apples	5	10	20	15	10

Find the AM, median, mode and variance of the distribution by shortcut method.

(b) Find the Binomial distribution for which mean = 4 and variance = 3.

Q.4 (a) Explain Normal Distribution and Standard Normal Distribution. How it is important in statistics?

(b) If on an average 20% of the bolts manufactured by a company are defective, then find the probability that in a random sample of 400 bolts, the number of defective bolts are (i) between 76 and 82 (ii) at least 75.

(c) Define Negative-Binomial distribution alongwith its properties.

Q.5 (a) What are raw and central moments? Give the relationship between first four raw and central moments.

(b) What is mathematical expectation? Find the expected number of boys in a family of 3 children.

Q.6 (a) What is Sampling Distribution? Derive the sampling distribution of (i) sample mean (ii) difference between two means (iii) sample proportion (iv) difference between two proportions.

(b) Calculate the mean deviation about mean of the following frequency distribution:

x	1	2	3	4	5	6	7
f	4	2	1	2	4	8	9

Q.7 (a) Explain point and interval estimation. Construct the confidence interval for the population mean when (i) population variance ($\mathring{u}^2$) is known (ii) when $\mathring{u}^2$ is unknown.

(b) Find the 95% confidence interval for the normal population mean if a random sample of 16 is drawn whose mean comes to 12.5 and the sum of squares of deviations from mean is equal to 93.75.

Q.8 (a) Give the probability function of Binomial, Poisson and Normal Distributions. When binomial distribution tends to Poisson Distribution and Normal Distribution.

(b) An aptitude test for selecting the bank officers was conducted on 1000 candidates. The average score was 42 and the standard deviation of scores was 24. Assuming normal distribution for scores, find the expected number of candidates whose score exceed (i) 58 (ii) lies between 40 and 50

Q.9 Assuming plant heights to be normally distributed with average height 72 cms and variance equal to 81 cm^2. Find (i) The minimum height of the selected plant if 20% largest plants are to be selected (ii) The maximum height of the selected plant if 15% shortest plants are to be selected.

Q.10 (a) Find the mean and variance of the following probability distribution:

x	-3	-1	0	2	4
p(x)	$\frac{1}{10}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{5}$

(b) If X, Y and Z are independent random variables, then find the mean and variance of the linear combination (4X ï 2Y + 3Z) if E(X) = 3, E(Y) = -4, E(Z) = 2, V(X) = 2, V(Y) = 1, V(Z) = 4

Q.11 (a) A cyclist covers his first five kilometers at an average speed of 10 k/hr and another three km at 8 km/hr and last two km at 5 km/hr. Find the average speed of the cyclist using harmonic mean.

(b) A test was given to 400 school children of whom 150 were boys and 250 girls. The results were as follows:

$n_1 = 150$	$n_2 = 250$
$\overline{X}_1 = 72$	$\overline{X}_2 = 73$
$\mathring{u}_1 = 7.0$	$\mathring{u}_2 = 6.4$

Find the mean and standard deviation of the combined group.

Q.12 (a) The probability that a man will alive next 25 years is 2/3 and the probability that his wife will alive is 3/5. Find the probability that (i) at least one will alive (ii) exactly one will alive (iii) Both will alive.

(b) In a single throw of a pair of die, find the probability of getting a doublet or the sum of points as 8.

(c) A bag contains 10 red, 5 white and 9 blue balls. If 4 balls are drawn at random, then find the probability that (i) all 4 are red balls (ii) 1 red and 2 white balls.

Revision Exercises for Final Examination
Stat-511 (Statistical Methods for Applied Sciences)

Q.1 Define null hypothesis, critical region, level of significance, one tailed and two tailed test used in testing of hypothesis. Explain the various steps involved in testing of hypothesis.

Q.2 What are the underlying assumptions of (i) Two sample Z-test and (ii) Two sample t-test for testing the equality of two population means? Explain the complete procedure of any one of them.

Q.3 What is difference between two-sample t-test and paired t-test? Explain the procedure of paired t-test.

Q.4 The length of tillers in a wheat variety V_1 for a random sample of 10 from a field gave the following data:

Tiller Length in Inches: 4.2, 4.6, 3.9, 4.1, 5.2, 3.8, 3.9, 4.3, 4.4, 5.6

Can we accept the hypothesis that the average tiller length of the variety is equal to 4.0 inches?

Q.5 Given below the increases in weights (lbs) of 10 and 12 cattle fed on diets A and B, respectively. Test at 5% level of significance that there is no difference between two diets:

	Increase in Weight (lbs)											
Diet A	10	6	16	17	13	12	8	14	15	9		
Diet B	7	13	22	15	12	14	18	8	21	23	10	17

Q.6 The heights of six randomly chosen sailors are 64, 66, 69, 70, 72, 73 inches and those of 10 randomly selected soldiers are: 61, 62, 65, 66, 69, 69, 70, 71, 72, 73. Discuss the suggestion that the sailors on an average are as tall as soldiers.

Q.7 Given below are the increase in weight (lbs) of 8 pigs when given two foods A and B on two different occasions. Test at 5% level whether (i) there is any difference between two foods; (ii) food B is superior to food A.

Food A	49	53	51	52	47	50	52	53
Food B	52	55	52	53	50	54	54	53

Q.8 An I.Q. test was administered to 8 persons before and after they were trained. The results are given below:

Persons	1	2	3	4	5	6	7	8
I.Q. before Training	110	120	123	132	120	115	118	122
I.Q. after Training	120	118	125	140	124	121	119	128

Test whether there is any change in I.Q. after training programme.

Q.9 Write short notes on the following:

(i) Box Plot (ii) Bartlett Test (iii) Fisher Z-transformation (iv) Kendall Coefficient of Concordance (v) Runs Test for randomness

Q.10 Give assumptions, situations and method of testing the differences between two population means when:

(i) Population variances are known (ii) Population variances are unknown

Q.11 What are the uses of chi-square test? How the independence of two attributes is tested in m x n contingency table. When and how YatesôCorrection is used?

Q.12 In a genetic crosses on 400 seeds (RrYy x RrYy) the following data on observed frequencies were obtained:

Phenotype	Ry	Ry	ry	ry
Observed Frequency	215	78	81	26

Test whether the data agrees with the ratio 9:3 3:1

Q.13 The following figure shows the distribution of digits in numbers chose at random from a telephone directory:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	180	200	190	230	210	160	250	220	210	150

Test whether digit may be taken to occur equally frequently in the directory.

Q.14 Explain the following:

(i) Steps for testing of statistical hypothesis (ii) Type-I and type-II errors (iii) Sampling distribution (iv) Mean Deviation about Mode (v) Standard Error (vi) Coefficient of determination (vii) Scatter Diagram (viii) Critical Region

Q.15 Show that the conditions at home has a bearing on the condition of the child on the basis of the following table:

	Condition at Home	
Condition of Child	Clean	Not Clean
Clean	75	40
Fairy Clean	35	15
Dirty	25	45

Q.16 From the following table regarding eye colour of mother and son, test if the colour of sonôs eye is associated with that of mother:

Mother's eye colour	**Son's eye colour**	
	Light Blue	**Not light blue**
Light Blue	45	16
Not Light Blue	4	35

Q.17 Discuss the F-test for testing the equality of two population variances. How the equality of more than two populations be tested?

Q.18 Test the assumption for equality of two population variances in Q5.

Q.19 Two independent samples of sizes 13 and 16 gave the sum of squares of deviations about their means as 240 and 600, respectively. Test at 5% level of significance the equality of population variances.

Q.20 Explain the principle of least squares method of curve fitting and use this technique to fit (i) the parabola $y = a + bx + cx^2$ (ii) $y = ax^b$ (iii) $y = ab^x$

Q.21 What is simple correlation? What are different methods of measuring it? How will you test the null hypothesis regarding population correlation coefficient ρ in the following cases:

(i) Null hypothesis $\rho = 0$

(ii) Null hypothesis $\rho = \rho_0$ ($\rho_0 \neq 0$)

Q.22 What is the difference between correlation and regression? What is the utility of two regression lines? Derive the relationship between coefficient of correlation and regression coefficients.

Q.23 From the following data on partially destroyed laboratory records, regression equations are: 8x ï 10y + 66 = 0 and 40x ï 18y ï 214 = 0 Variance (x) = 9. Find (i) the mean values of x and y; (ii) Coefficient of correlation between x and y.

Q.24 How the regression coefficient of y on x is tested? Also give the properties of correlation coefficient and regression coefficient.

Q.25 Explain simple, partial and multiple correlation coefficients. Give the test statistic and its distribution for testing each of them.

Q.26 Find the Spearmanôs coefficient of rank correlation and Kendallôs Ű between performance in Economics and Statistics:

Student	1	2	3	4	5	6	7	8
Marks in Economics(x)	25	30	38	22	50	70	30	90
Marks in Statistics (y)	50	40	60	40	30	20	40	70

Q.27 Calculate the multiple correlation coefficients $R_{1.23}$ and $R_{3.12}$ from the following data: $r_{12} = 0.60$; $r_{13} = 0.70$ and $r_{23} = 0.65$, n = 25 and test their significance.

Q.28 Calculate the partial correlation coefficients $r_{32.1}$, $r_{12.3}$ and $r_{13.2}$ from the following data $r_{12} = 0.7$, $r_{23} = 0.5$ and $r_{13} = 0.4$ and test for its significance.

Q.29 What are non-parametric tests? Discuss their advantages and disadvantages as compared to parametric tests.

Q.30 How Wilcoxon signed rank test is an improvement over sign test. Explain the complete procedure of both.

Q.31 To compare the effectiveness of three types of weight reducing diets, a homogeneous group of 22 women was divided into 3 sub-groups and each subgroup followed one of these diet plans for a period of two months. The weight reductions (kg) were recorded and are given below:

Diet Plan-I	4.3	3.2	2.7	6.2	5.0	3.9			
Diet Plan-II	5.3	7.4	8.3	5.5	6.7	7.2	8.5		
Diet Plan-III	1.4	2.1	2.7	3.1	1.5	0.7	4.3	3.5	0.3

Use Kruskal-Wallis test to test the hypothesis that the three weight reducing diet plans are equally effective at 5% level of significance.

Q.32 The grain pod yield (kg) under four different treatments in an experiment is given below:

	Treatments			
Pod No.	T_1	T_2	T_3	T_4
1	4.18	4.45	4.16	3.49
2	4.41	3.89	3.70	3.38
3	4.51	3.98	4.11	3.59
4	3.88	4.28	3.81	3.85
5	4.89	4.95	4.46	4.01
6	5.01	4.88		3.49
7	4.61	4.26		

Test at Ŭ = 0.05 that there is no significant difference among the treatments affecting the yield by Kruskal-Wallis Test?

Q.33 The quantity of serum albumin (gms) per 100 ml in lepers under four different drugs were recorded as follows:

	Treatments			
Sample No. (Block)	Drug-I	Drug-II	Drug-III	Drug-IV
1	4.30	3.65	3.05	3.90
2	4.00	3.60	4.10	3.10
3	4.10	2.70	4.20	3.20
4	3.80	3.15	3.70	4.20
5	3.30	3.75	3.60	3.00
6	4.50	2.95	4.80	3.40

Apply Friedmanôs test to confirm that the contents of serum albumin are same in different drugs.

Q.34 Two models of machine are under consideration for purchase. An organization demonstrates the output capacity in the trial period with a team of 15 operators separately. The output is given below:

Operator No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Model-I	78	75	53	68	82	64	95	86	64	71	51	80	51	70	75
Model-II	60	58	46	71	80	59	73	78	37	75	60	79	38	51	69

Use Wilcoxon signed rank test to infer if there is any significant difference between the output capacities of the two machine models at 5% level of significance.