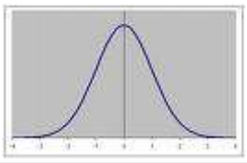
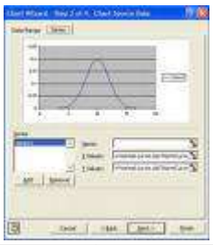
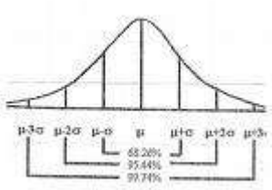
LECTURE NOTES
Course No: STCA-101
STATISTICS

TIRUPATI

Karl Pearson	**R. A. Fisher**
 Karl Pearson (né Carl Pearson)	 Sir Ronald Aylmer Fisher (1890-1962)
Born: 27 March 1857 Islington, London, England	**Born:** 17 February 1890 East Finchley, London , England

Prepared By

Dr. G. MOHAN NAIDU M.Sc., Ph.D.,
Assistant Professor & Head
Dept. of Statistics & Mathematics
S.V. Agricultural College
TIRUPATI

ACHARYA N.G. RANGA AGRICULTURAL UNIVERSITY

LECTURE OUTLINE

Course No. STCA-101 Credits: 2 (1+1)
Course Title: STATISTICS

THEORY

S. No.	Topic/Lesson
1	Introduction to Statistics, Definition, Advantages and Limitations.
2	Frequency distribution: Construction of Frequency Distribution table.
3	Measures of Central Tendency: Definition, Characteristics of Satisfactory average.
4	Arithmetic Mean, Median, Mode for grouped and ungrouped data – Merits and Demerits of Arithmetic Mean.
5	Measures of Dispersion: Definition, standard deviation, variance and coefficient of variation.
6	Normal Distribution and its properties. Introduction to Sampling: Random sampling, concept of standard error of Mean.
7	Tests of Significance: Introduction, Types of errors, Null hypothesis, level of significance and degrees of freedom, steps in testing of hypothesis.
8	Large sample tests: Test for Means – Z-test, One sample and Two samples with population S.D. known and Unknown.
9	Small sample tests: Test for Means – One sample t – test, Two samples t-test and Paired t-test.
10	Chi-Square test in 2x2 contingency table with Yate's correction, F-test.
11	Correlation: Definition, types, properties, Scatter diagram, calculation and testing.
12	Regression: Definition, Fitting of two lines Y on X and X on Y, Properties, inter relation between correlation and regression.
13	Introduction to Experimental Designs, Basic Principles, ANOVA its assumptions.
14	Completely Randomized Design: Layout, Analysis with equal and unequal replications.
15	Randomized Block Design: Layout and Analysis.
16	Latin Square Design: Layout and Analysis.

PRACTICALS

S .No.	Topic
1	Construction of Frequency Distribution tables
2	Computation of Arithmetic Mean for Grouped and Un-grouped data
3	Computation of Median for Grouped and Un-grouped data
4	Computation of Mode for Grouped and Un-grouped data
5	Computation of Standard Deviation and variance for grouped and ungrouped data
6	Computation of coefficient of variation for grouped and ungrouped data
7	SND (Z) test for single sample, Population SD known and Unknown
8	SND (Z) test for two samples, Population SD known and Unknown
9	Student's t-test for single and two samples
10	Paired t-test and F-test
11	Chi-square test – 2x2 contingency table with Yate's correction
12	Computation of correlation coefficient and its testing
13	Fitting of simple regression equations Y on X and X on Y
14	Completely Randomized Design: Analysis with equal and unequal replications
15	Randomized Block Design: Analysis
16	Latin Square Design: Analysis

S T A T I S T I C S

Statistics has been defined differently by different authors from time to time. One can find more than hundred definitions in the literature of statistics.

Statistics can be used either as plural or singular. When it is used as plural, it is a systematic presentation of facts and figures. It is in this context that majority of people use the word statistics. They only meant mere facts and figures. These figures may be with regard to production of food grains in different years, area under cereal crops in different years, per capita income in a particular state at different times etc., and these are generally published in trade journals, economics and statistics bulletins, news papers, etc.,

When statistics is used as singular, it is a science which deals with collection, classification, tabulation, analysis and interpretation of data.

The following are some important definition of statistics.

Statistics is the branch of science which deals with the collection, classification and tabulation of numerical facts as the basis for explanations, description and comparison of phenomenon - Lovitt

The science which deals with the collection, analysis and interpretation of numerical data - Corxton & Cowden

The science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates - King

Statistics may be called the science of counting or science of averages or statistics is the science of the measurement of social organism, regarded as whole in all its manifestations - Bowley

Statistics is a science of estimates and probabilities -Boddington

Statistics is a branch of science, which provides tools (techniques) for decision making in the face of uncertainty (probability) - Wallis and Roberts

This is the modern definition of statistics which covers the entire body of statistics

All definitions clearly point out the four aspects of statistics collection of data, analysis of data, presentation of data and interpretation of data.

Importance: Statistics plays an important role in our daily life, it is useful in almost all sciences – social as well as physical – such as biology, psychology, education, economics, business management, agricultural sciences etc., . The statistical methods can be and are

being followed by both educated and uneducated people. In many instances we use sample data to make inferences about the entire population.

1. Planning is indispensable for better use of nation's resources. Statistics are indispensable in planning and in taking decisions regarding export, import, and production etc., Statistics serves as foundation of the super structure of planning.
2. Statistics helps the business man in the formulation of polices with regard to business. Statistical methods are applied in market and production research, quality control of manufactured products
3. Statistics is indispensable in economics. Any branch of economics that require comparison, correlation requires statistical data for salvation of problems
4. State. Statistics is helpful in administration in fact statistics are regarded as eyes of administration. In collecting the information about population, military strength etc., Administration is largely depends on facts and figures thud it needs statistics
5. Bankers, stock exchange brokers, insurance companies all make extensive use of statistical data. Insurance companies make use of statistics of mortality and life premium rates etc., for bankers, statistics help in deciding the amount required to meet day to day demands.
6. Problems relating to poverty, unemployment, food storage, deaths due to diseases, due to shortage of food etc., cannot be fully weighted without the statistical balance. Thus statistics is helpful in promoting human welfare
7. Statistics are a very important part of political campaigns as they lead up to elections. Every time a scientific poll is taken, statistics are used to calculate and illustrate the results in percentages and to calculate the margin for error.

In agricultural research, Statistical tools have played a significant role in the analysis and interpretation of data.

1. In making data about dry and wet lands, lands under tanks, lands under irrigation projects, rainfed areas etc.,
2. In determining and estimating the irrigation required by a crop per day, per base period.
3. In determining the required doses of fertilizer for a particular crop and crop land.

4. In soil chemistry also statistics helps classifying the soils basing on their analysis results, which are analyzed with statistical methods.
5. In estimating the losses incurred by particular pest and the yield losses due to insect, bird, or rodent pests statistics is used in entomology.
6. Agricultural economists use forecasting procedures to determine the future demand and supply of food and also use regression analysis in the empirical estimation of function relationship between quantitative variables.
7. Animal scientists use statistical procedures to aid in analyzing data for decision purposes.
8. Agricultural engineers use statistical procedures in several areas, such as for irrigation research, modes of cultivation and design of harvesting and cultivating machinery and equipment.

Limitations of Statistics:

1. Statistics does not study qualitative phenomenon
2. Statistics does not study individuals
3. Statistics laws are not exact laws
4. Statistics does not reveal the entire information
5. Statistics is liable to be misused
6. Statistical conclusions are valid only on average base

Types of data: The data are of two types i) Primary Data and ii) Secondary Data

i) Primary Data: It is the data collected by the primary source of information i.e by the investigator himself.

ii) Secondary Data: It is the data collected from secondary sources of information, like news papers, trade journals and statistical bulletins, etc.,

Variables and Attributes:

Variability is a common characteristic in biological sciences. A quantitative or qualitative characteristic that varies from observation to observation in the same group is called a variable. In case of quantitative variables, observations are made using interval scales whereas in case of qualitative variables nominal scales are used. Conventionally, the quantitative variables are termed as variables and qualitative variables are termed as attributes. Thus, yield of a crop, available nitrogen in soil, daily temperature, number of leaves per plant and number of eggs laid by insects are all variables. The crop varieties, soil types, shape of seeds, seasons and sex of insects are attributes.

The variable itself can be classified as continuous variable and discrete variable. The variables for which fractional measurements are possible, at least conceptually, are called continuous variables. For example, in the range of 7 kg to 10 yield of a crop, yield might be 7.15 or 7.024kg. Hence, yield is a continuous variable. The variables for which such factional measurements are not possible are called discrete or discontinuous variables. For example, the number of grains per panicle of paddy can be counted in full numbers like 79, 80, 81 etc. Thus, number of grains per panicle is a discrete variable. The variables, discrete or continuous are dented by capital letters like X and Y.

Construction of Frequency Distribution Table:

In statistics, a **frequency distribution** is a tabulation of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way the table summarizes the distribution of values in the sample.

The following steps are used for construction of frequency table

Step-1: The number of classes are to be decided

The appropriate number of classes may be decided by Yule's formula, which is as follows:

Number of classes = $2.5 \times n^{1/4}$. where 'n' is the total number of observations

Step-2: The class interval is to be determined. It is obtain by using the relationship

$$C.I = \frac{\text{Maximum value in the given data} - \text{Minimum value in the given data}}{\text{Number of classes}}$$

Step-3: The frequencies are counted by using Tally marks

Step-4: The frequency table can be made by two methods

a) Exclusive method

b) Inclusive method

a) **Exclusive method:** In this method, the upper limit of any class interval is kept the same as the lower limit of the just higher class or there is no gap between upper limit of one class and lower limit of another class. It is continuous distribution

Ex:

C.I.	Tally marks	Frequency (f)
0-10		
10-20		
20-30		

b) **Inclusive method:** There will be a gap between the upper limit of any class and the lower limit of the just higher class. It is discontinuous distribution

Ex:

C.I.	Tally marks	Frequency (f)
0-9		
10-19		
20-29		

To convert discontinuous distribution to continuous distribution by subtracting 0.5 from lower limit and by adding 0.5 to upper limit

Note: The arrangement of data into groups such that each group will have some numbers. These groups are called class and number of observations against these groups are called frequencies.

Each class interval has two limits 1. Lower limit and 2. Upper limit

The difference between upper limit and lower limit is called length of class interval. Length of class interval should be same for all the classes. The average of these two limits is called mid value of the class.

Example: Construct a frequency distribution table for the following data

25, 32, 45, 8, 24, 42, 22, 12, 9, 15, 26, 35, 23, 41, 47, 18, 44, 37, 27, 46, 38, 24, 43, 46, 10, 21, 36, 45, 22, 18.

Solution: Number of observations (n) = 30

Number of classes $= 2.5 \times n^{1/4}$

$= 2.5 \times 30^{1/4}$

$= 2.5 \times 2.3$

$= 5.8 \cong 6.0$

$$\text{Class interval} = \frac{Max.value - Min.value}{No.of\ .classes}$$

$$= \frac{46-8}{6}$$

$$= \frac{38}{6} = 6.3 \cong 6.0$$

Inclusive method:

C.I.	Tally marks	Frequency (f)
8-14	||||	4
15-21	||||	4
22-28	$\cancel{\|\|\|\|}$ |||	8
29-35	||	2
36-42	$\cancel{\|\|\|\|}$	5
42-49	$\cancel{\|\|\|\|}$ ||	7
Total		30

Exclusive method:

C.I.	Tally marks	Frequency (f)
7.5-14.5	||||	4
14.5-21.5	||||	4
21.5-28.5	$\cancel{\|\|\|\|}$ |||	8
28.5-35.5	||	2
35.5-42.5	$\cancel{\|\|\|\|}$	5
42.5-49.5	$\cancel{\|\|\|\|}$ ||	7
Total		30

MEASURES OF CENTRAL TENDENCY

One of the most important aspects of describing a distribution is the central value around which the observations are distributed. Any mathematical measure which is intended to represent the center or central value of a set of observations is known as measure of central tendency (or)

The single value, which represents the group of values, is termed as a ‘measure of central tendency’ or a measure of location or an average.

Characteristics of a Satisfactory Average:

1. It should be rigidly defined
2. It should be easy to understand and easy to calculate
3. It should be based on all the observations
4. It should be least affected by fluctuations in sampling
5. It should be capable of further algebraic treatment
6. It should not be affected much by the extreme values
7. It should be located easily

Types of average:

1. Arithmetic Mean
2. Median
3. Mode
4. Geometric Mean
5. Harmonic Mean

Arithmetic Mean (A.M): It is defined as the sum of the given observations divided by the number of observations. A.M. is measured with the same units as that of the observations.

Ungrouped data:

Direct Method: Let $x_1, x_2, \ldots\ldots, x_n$ be 'n' observations then the A.M is computed from the formula:

$$\text{A.M.} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $\sum_{i=1}^{n} x_i$ = sum of the given observations

n = Number of observations

Linear Transformation Method or Deviation Method: When the variable constitutes large values of observations, computation of arithmetic mean involves more calculations. To overcome this difficulty, Linear Transformation Method is used. The value x_i is transformed to d_i.

and $$\text{A.M.} = A + \frac{\sum_{i=1}^{n} d_i}{n}$$

where A = Assumed mean which is generally taken as class mid point of middle class or the class where frequency is large.

$d_i = x_i - A$ = deviations of the i^{th} value of the variable taken from an assumed mean and n = number of observations

Grouped Data:

Let $f_1, f_2, \ldots\ldots\ldots, f_n$ be 'n' frequencies corresponding to the mid values of the class intervals $x_1, x_2, \ldots\ldots\ldots x_n$ then

$$\text{A.M.} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots\ldots\ldots + f_nx_n}{f_1 + f_2 + \ldots\ldots\ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{N} \qquad \text{(direct method)}$$

and

$$\text{A.M.} = A + \left(\frac{f_1d_1 + f_2d_2 + \ldots\ldots\ldots\ldots + f_nd_n}{f_1 + f_2 + \ldots\ldots\ldots + f_n} \right) C = A + \left[\frac{\sum_{i=1}^{n} f_i d_i}{N} \right] C \quad \text{(indirect method)}$$

where d_i = deviation = $\frac{x_i - A}{c}$; f = frequency; C = class interval;

x = mid values of classes.

Arithmetic mean, when computed for the data of entire population, is represented by the symbol 'μ'. Where as when it is computed on the basis of sample data, it is represented as $\overline{X}$, which is the estimate of μ.

Properties of A.M.:

i) The algebraic sum of the deviations taken from arithmetic mean is zero
i.e. $\Sigma(x\text{-A.M.}) = 0$

ii) Let $\overline{x_1}$ be the mean of n_1 observations, $\overline{x_2}$ be the mean of the n_2 observations …… $\overline{x_k}$ be the mean of n_k observations then the mean $\bar{x}$ of n = $(n_1+n_2+\ldots..n_k)$ observations is given by

$$\bar{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + \ldots\ldots + n_k\overline{x_k}}{n_1 + n_2 + \ldots. + n_k} = \frac{\sum_{i=1}^{k} n_i \overline{x_i}}{\sum_{i=1}^{k} n_i}$$

Merits:

1. It is well defined formula defined
2. It is easy to understand and easy to calculate
3. It is based upon all the observations

4. It is amenable to further algebraic treatments, provided the sample is randomly obtained.
5. Of all averages, arithmetic mean is affected least by fluctuations of sampling

Demerits:

1. Cannot be determined by inspection nor it can be located graphically
2. Arithmetic mean cannot be obtained if a single observation is missing or lost
3. Arithmetic mean is affected very much by extreme values
4. Arithmetic mean may lead to wrong conclusions if the details of the data from which it is computed are not given
5. In extremely asymmetrical (skewed) distribution, usually arithmetic mean is not a suitable measure of location

Examples:

i) Ungrouped data:

If the weights of 7 ear-heads of sorghum are 89, 94, 102, 107, 108, 115 and 126 g. Find arithmetic mean by direct and deviation methods

Solution:

i) Direct Method: $\text{A.M.} = \frac{\sum_{i=1}^{n} x_i}{n}$

ii) Deviation Method: $\text{A.M.} = A + \frac{\sum_{i=1}^{n} d_i}{n}$

Where n = number of observations; x_i = given values;

A = arbitrary mean (assumed value); d = deviation $= x_i - A$

x_i	$d_i = x_i - A$
89	89-102= -13
94	94-102= -8
102	102-102= 0
107	107-102= 5
108	108-102= 6
115	115-102= 13
126	126-102= 24
$\sum x = 741$	$\sum d = 29$

here A = assumed value = 102

i) $\text{A.M.} = \frac{\sum_{i=1}^{n} x_i}{n}$ (ii) $\text{A.M.} = A + \frac{\sum_{i=1}^{n} d_i}{n}$

$$= \frac{741}{7} \qquad = 102 + \left(\frac{27}{7}\right)$$

$$= 105.86 \text{ g} \qquad = 105.86 \text{ g.}$$

ii) Grouped Data:

The following are the 405 soybean plant heights collected from a particular plot. Find the arithmetic mean of the plants by direct and indirect method:

Plant height (Cms)	8-12	13-17	18-22	23-27	28-32	33-37	38-42	43-47	48-52	53-57
No. of plants(f_i)	6	17	25	86	125	77	55	9	4	1

Solution:

a) Direct Method:

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

Where x_i = mid values of the corresponding classes

N = Total frequency = $\sum_{i=1}^{n} f_i$

f_i = frequency

b) Deviation Method:

$$\text{A.M.} = A + \left(\frac{\sum f_i d_i}{N}\right) C$$

Where d_i = deviation (i.e. $d_i = \frac{x_i - A}{C}$)

Length of class interval (C) = 5; Assumed value (A) = 30

C.I	f_i	x_i	$f_i x_i$	$d_i = \frac{x_i - A}{C}$	$f_i d_i$
8-12	6	10	60	-4	-24
13-17	17	15	255	-3	-51
18-22	25	20	500	-2	-50
23-27	86	25	2150	-1	-86
28-32	125	30	3750	0	0
33-37	77	35	2695	1	77
38-42	55	40	2200	2	110
43-47	9	45	405	3	27
48-52	4	50	200	4	16
53-57	1	55	55	5	5
Total	N = 405		$\Sigma f_i x_i = 12270$		$\Sigma f_i d_i = 24$

a) Direct Method:

$$\text{A.M.} = \frac{12270}{405} = 30.30 \text{ cms.}$$

b) Deviation Method:

$$\text{A.M.} = 30 + \left(\frac{24}{405}\right)5$$

$$= 30.30 \text{ cms.}$$

MEDIAN

The median is the middle most item that divides the distribution into two equal parts when the items are arranged in ascending order.

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

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

i) Arrange the data in ascending or descending order of magnitude

ii) Find out cumulative frequencies

iii) Apply formula: Median = Size of $\frac{N+1}{2}$, where N= $\sum f$

iv) Now look at the cumulative frequency column and find, that total which is either equal to $\frac{N+1}{2}$ or next higher to that and determine the value of the variable corresponding to it, which gives the value of median.

Continuous frequency distribution:

If the data are given with class intervals then the following procedure is adopted for the calculation of median.

i) find $\frac{N+1}{2}$, where N = Σf

ii) see the (less than) cumulative frequency just greater than $\frac{N+1}{2}$

iii) the corresponding value of x is median

In the case of continuous frequency distribution, the class corresponding to the cumulative frequency just greater than $\frac{N+1}{2}$ is called the median class and the value of median is obtained by the following formula:

$$\text{Median} = l + \left[\frac{\frac{N+1}{2} - m}{f}\right] C$$

Where l is the lower limit of median class

f is the frequency of the median class

m is the cumulative frequency of the class preceding the median class

C is the class length of the median class

N = total frequency

Examples:

Case-i) when the number of observations (n) is odd:

The number of runs scored by 11 players of a cricket team of a school are

5, 19, 42, 11, 50, 30, 21, 0, 52, 36, 27

To compute the median for the given data, we proceed as follows:

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude.

Let us arrange the values in ascending order:

0, 5, 11, 19, 21, 27, 30, 36, 42, 50, 52

$$\therefore \text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{value} = \left(\frac{11+1}{2}\right)^{th} \text{value}$$

$$= 6^{th} \text{ value}$$

Now the 6^{th} value in the data is 27.

$\therefore$ Median = 27 runs

Case-ii) when the number of observations (n) is even:

Find the median of the following heights of plants in Cms:

6, 10, 4, 3, 9, 11, 22, 18

In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.

Let us arrange the given items in ascending order

3, 4, 6, 9, 10, 11, 18, 22

In this data the number of items n = 8, which is even.

Median = average of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th terms.

= average of 9 and 10

$\therefore$ Median = 9.5 Cms.

Grouped Data:

Find out the median for the following frequency distribution of 180 sorghum ear-heads.

Weight of ear-heads (in g)	40-60	60-80	80-100	100-120	120-140	140-160	160-180	180-200
No.of ear-heads	6	28	35	45	30	15	12	9

Solution: Median = $l + \left(\frac{\frac{N+1}{2} - m}{f}\right) C$ ---------- (1)

Where l is the lower limit of the median class

f is the frequency of the median class

m is the cumulative frequency of the class preceding the median class

C is the class interval of the median class

and N = $\sum f$ = Total number of observations

Weight of ear-heads (in g)	No. of ear-heads	Cumulative Frequency (CF)
40-60	6	6
60-80	28	34
80-100	35	69 - m
100-120	45 – f	114 (Median class)
120-140	30	144
140-160	15	159
160-180	12	171
180-200	9	180
	N = $\sum f$ =180	

Here $\frac{N+1}{2} = \frac{181}{2} = 90.5$. Cumulative frequency just greater than 90.5 is 69 and the corresponding class is 100-120. The median class is 100-120.

N = 121; L =40; f = 24; m = 39 and C = 10

Substituting the above values in equation (1), we get

$$\text{Median} = 100 + \left(\frac{90.5 - 69}{45}\right)20$$

$$= 109.56 \text{ g}$$

Merits and Demerits of Median:

Merits:

1. It is rigidly defined.
2. It is easily understood and is easy to calculate. In some cases it can be located merely by inspection.
3. It is not at all affected by extreme values.
4. It can be calculated for distributions with open-end classes

Demerits:

1. In case of even number of observations median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms.

2. It is not amenable to algebraic treatment
3. As compared with mean, it is affected much by fluctuations of sampling.

MODE

Mode is the value which occurs most frequently in a set of observations or mode is the value of the variable which is predominant in the series.

In case of discrete frequency distribution mode is the value of x corresponding to maximum frequency

In case of continuous frequency distribution, mode is obtained from the formula:

$$\text{Mode} = l + \left[\frac{(f - f_1)}{2f - f_1 - f_2}\right] \text{C}$$

Where l is the lower limit of modal class

C is class interval of the modal class

f the frequency of the modal class

f_1 and f_2 are the frequencies of the classes preceding and succeeding the modal class respectively

If the distribution is moderately asymmetrical, the mean, median and mode obey the empirical relationship:

Mode = 3 Median – 2 Mean

Example: Find the mode value for the following data:

27, 28, 30, 33, 31, 35, 34, 33, 40, 41, 55, 46, 31, 33, 36, 33, 41, 33.

Solution: As seen from the above data, the item 33 occurred maximum number of times i.e. 5 times. Hence 33 is considered to be the modal value of the given data.

Grouped Data:

Example: The following table gives the marks obtained by 89 students in Statistics. Find the mode.

Marks	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
No. of students	4	6	10	16	21	18	9	5

Solution: $\text{Mode} = l + \frac{(f - f_1)}{2f - f_1 - f_2} \text{xC}$

Where l = the lower limit of the modal class ; C = length of the modal class

f = the frequency of the modal class

f_1 = the frequency of the class preceding modal class

f_2 = the frequency of the class succeeding modal class

Sometimes it so happened that the above formula fails to give the mode. In this case, the modal value lies in a class other than the one containing maximum frequency. In such cases we take the help of the following formula;

$$\text{Mode} = l + \frac{f_2}{f_1 + f_2} \text{xC}$$

Where f , c, f_1 and f_2 have usual meanings.

Marks	No. of students (f)
10-14	4
15-19	6
20-24	10
25-29	$16f_1$
30-34	21f
35-39	$18f_2$
40-44	9
45-49	5

From the above table it is clear that the maximum frequency is 21 and it lies in the class 30-34.

Thus the modal class is 29.5-34.5

Here L = 29.5 , c = 5, f = 21, f_1 =16, f_2 = 18

$$\text{Mode} = 30 + \left[\frac{21-16}{2*21-16-18}\right]*5$$

$$= 30 + 3.13$$

$$= 33.63 \text{ m}$$

Merits and Demerits of Mode:

Merits:

1. Mode is readily comprehensible and easy to calculate.
2. Mode is not at all affected by extreme values.
3. Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. Open-end classes also do not pose any problem in the location of mode

Demerits:

1. Mode is ill defined. It is not always possible to find a clearly defined mode. In some cases, we may come across distributions with two modes. Such distributions are called bi-modal. If a distribution has more than two modes, it is said to be multimodal.
2. It is not based upon all the observations.
3. It is not capable of further mathematical treatment.
4. As compared with mean, mode is affected to a greater extent by fluctuations of sampling.

Dispersion

Dispersion means scattering of the observations among themselves or from a central value (Mean/ Median/ Mode) of data. We study the dispersion to have an idea about the variation.

Suppose that we have the distribution of the yields (kg per plot) of two Ground nut varieties from 5 plots each. The distribution may be as follows:

Variety 1:	46	48	50	52	54
Variety 2:	30	40	50	60	70

It can be seen that the mean yield for both varieties is 50 k.g. But we can not say that the performances of the two varieties are same. There is greater uniformity of yields in the first variety where as there is more variability in the yields of the second variety. The first variety may be preferred since it is more consistent in yield performance.

Measures of Dispersion:

These measures give us an idea about the amount of dispersion in a set of observations. They give the answers in the same units as the units of the original observations. When the observations are in kilograms, the absolute measure is also in kilograms. If we have two sets of observations, we cannot always use the absolute measures to compare their dispersion. The absolute measures which are commonly used are:

1. The Range
2. The Quartile Deviation
3. The Mean Deviation
4. The Standard Deviation and Variance
5. Coefficient of Variation
6. Standard Error

Standard Deviation: It is defined as the positive square root of the arithmetic mean of the squares of the deviations of the given values from arithmetic mean. The square of the standard deviation is called variance.

Ungrouped data:

Let $x_1, x_2, \ldots\ldots, x_n$ be n observations then the standard deviation is given by the formula

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - A.M.)^2}{n}} \quad \text{where A.M.} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where n = no. of observations.

Simplifying the above formula, we have

or $$\text{S.D.} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)}$$

by linear transformation method, we have

$$\sigma = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} d_i^{\,2} - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right)}$$

where $d_i = x_i - A$; A= Assumed value; x_i = Given values

<u>Continuous frequency distribution</u>: (grouped data):

Let $f_1, f_2, \ldots\ldots, f_n$ be the 'n' frequencies corresponding to the mid values of the classes $x_1, x_2, \ldots\ldots, x_n$ respectively, then the standard deviation is given by

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{n} f(x_i - A.M.)^2}{N}} \qquad \text{where } \sum_{i=1}^{n} f_i = N$$

Simplifying the above formula, we have

$$\text{S.D.} = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{n} f_i x_i^{\,2} - \left(\frac{\sum_{i=1}^{n} f_i x_i}{N}\right)^2\right)}$$

by linear transformation method, we have

$$\text{S.D.} = C\sqrt{\frac{1}{N}\left(\sum f_i d_i^{\,2} - \frac{(\sum f_i d_i)^2}{N}\right)}$$

where $d_i = \frac{x_i - A}{C}$; A = assumed value; and C = class interval

S.D. for population data is represented by the symbol 'σ'

<u>Ungrouped data</u>:

Example:

Calculate S.D. for the kapas yields (in kg per plot) of a cotton variety recorded from seven plots 5, 6, 7, 7, 9, 4, 5

i) Direct method:

$$\text{S.D.} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)}$$

ii) Deviation Method:

$$\text{S.D.} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} d_i^{\,2} - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right)}$$

Where x_i = given values

Assumed value (A) = 7

d_i = deviation (i.e. $d_i = x_i - A$)

x_i	$x_i^{\,2}$	$d_i = x_i - A$	$d_i^{\,2}$
5	25	5-7=-2	4
6	36	6-7=-1	1
7	49	7-7=0	0
7	49	7-7=0	0
9	81	9-7=2	4
4	16	4-7=-3	9
5	25	5-7=-2	4
$\Sigma x_i = 43$	$\Sigma x_i^2 = 281$	$\Sigma d_i = -6$	$\Sigma d_i^{\,2} = 22$

i) Direct method:

$$\text{S.D.} = \sqrt{\frac{1}{7}\left[281 - \frac{(43)^2}{7}\right]}$$

$$= 1.55 \text{ kg.}$$

ii) Deviation Method:

$$\text{S.D.} = \sqrt{\frac{1}{7}\left[22 - \frac{(-6)^2}{7}\right]}$$

$$= 1.55 \text{ kg}$$

Grouped Data:

Example:

The following are the 381soybean plant heights in Cms collected from a particular plot. Find the Standard deviation of the plants by direct and deviation method:

Plant heights (Cms)	No. of Plants (f_i)
6.8 -7.2	9
7.3 -7.7	10
7.8-8.2	11
8.3-8.7	32
8.8-9.2	42
9.3-9.7	58
9.8-10.2	65
10.3-10.7	55
10.8-11.2	37
11.3-11.7	31
11.8-12.2	24
12.3-12.7	7

Solution:

i) Direct method:

$$\text{A.M.} = = \frac{\sum_{i=1}^{n} f_i x_i}{N} \; ; \qquad \text{where N} = \sum_{i=1}^{n} f_i$$

$$\text{S.D.} = \sqrt{\frac{1}{N}\left[\sum_{i=1}^{n} f_i {x_i}^2 - \frac{\left(\sum_{i=1}^{n} f_i x_i\right)^2}{N}\right]}$$

ii) Deviation Method:

$$\text{A.M.} = A + \left(\frac{\sum_{i=1}^{n} f_i d_i}{N}\right) C$$

$$\text{S.D.} = C\sqrt{\frac{1}{N}\left[\sum_{i=1}^{n} f_i d_i^{\,2} - \frac{\left(\sum_{i=1}^{n} f_i d_i\right)^2}{N}\right]}$$

C.I.	f_i	x_i	$f_i x_i$	$f_i x_i^{\,2}$	$d_i = \frac{x_i - A}{C}$	$f_i d_i$	$f_i d_i^{\,2}$
6.8-7.2	9	7.0	63	441	-5	-45	225
7.3-7.7	10	7.5	75	562.5	-4	-40	160
7.8-8.2	11	8.0	88	704	-3	-33	99
8.3-8.7	32	8.5	272	2312	-2	-64	128
8.8-9.2	42	9.0	378	3402	-1	-42	42
9.3-9.7	58	9.5	551	5234.5	0	0	0
9.8-10.2	65	10	650	6500	1	65	65
10.3-10.7	55	10.5	577.5	6063.75	2	110	220
10.8-11.2	37	11.0	407	4477	3	111	333
11.3-11.7	31	11.5	356.5	4099.75	4	124	496
11.8-12.2	24	12.0	288	3456	5	120	600
12.3-12.7	7	12.5	87.5	1093.75	6	42	252
	N =381		$\Sigma f_i x_i$ =3793.5	$\Sigma f_i x_i^{\,2}$ =38346.25		$\Sigma f_i d_i$ = 348	$\Sigma f_i d_i^{\,2}$ = 2620

i) Direct method:

$$\text{A.M.} = \frac{3793.5}{381} = 9.96 \text{ Cms}$$

$$\text{S.D.} = \sqrt{\frac{1}{381}\left[38346.25 - \frac{(3793.5)^2}{381}\right]}$$

$$= \sqrt{\frac{1}{381}[38346.25 - 37770.71]}$$

$$= \sqrt{1.5106} = 1.23 \text{ Cms.}$$

ii) Deviation Method:

$$\text{A.M.} = 9.5 + \left(\frac{348}{381}\right)0.5 = 9.96 \text{ Cms}$$

$$\text{S.D.} = 0.5\sqrt{\frac{1}{381}\left[2620 - \frac{(348)^2}{381}\right]}$$

$$= 0.5\text{x}47.98 = 1.23 \text{ Cms.}$$

Measures of Relative Dispersion:

These measures are calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measure. If the original is in dollar or kilometers, we do not use these units with relative measure of dispersion. These are a sort of ratio and are called coefficients.

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal. Then their variability can be compared directly by using their standard deviations. However, if their means are widely different or if they are expressed in different units of measurement. We can not use the standard deviations as such for comparing their variability. We have to use the relative measures of dispersion in such situations.

Coefficient of Variation (C.V.)

Coefficient of variation is the percentage ratio of standard deviation and the arithmetic mean. It is usually expressed in percentage. The formula for C.V. is,

$$\text{C.V.} = \frac{S.D.}{Mean}\text{x}100$$

$$\text{Where S.D.} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)} \text{ and}$$

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n}$$

The coefficient of variation will be small if the variation is small of the two groups, the one with less C.V. said to be more consistent.

Note: 1. Standard deviation is absolute measure of dispersion

2. Coefficient of variation is relative measure of dispersion.

Example: Consider the distribution of the yields (per plot) of two ground nut varieties. For the first variety, the mean and standard deviation are 82 kg and 16 kg respectively. For the second variety, the mean and standard deviation are 55 kg and 8 kg respectively. Then we have, for the first variety

$$\text{C.V.} = \frac{16}{82} x100 = 19.5\%$$

For the second variety

$$\text{C.V.} = \frac{8}{55} x100 = 14.5\%$$

It is apparent that the variability in second variety is less as compared to that in the first variety. But in terms of standard deviation the interpretation could be reverse.

Example: Below are the scores of two cricketers in 10 innings. Find who is more 'consistent scorer' by Indirect method.

A	204	68	150	30	70	95	60	76	24	19
B	99	190	130	94	80	89	69	85	65	40

Solution: Let the player A = x

And the player B = y

Coefficient of variation of x= $(C.V.)_x = \frac{\sigma_x}{\bar{x}} X100$

Where $\bar{x} = A + \left(-\frac{\sum d_x}{n} \right)$ where $d_x = x_i - A$

Standard deviation of x $= \sigma_x = \sqrt{\frac{1}{n}\left[\sum d_x^{\,2} - \frac{\left(\sum dx\right)^2}{n} \right]}$

and coefficient of variation of y $= (C.V.)_y = \frac{\sigma_y}{\bar{y}} X100$

Standard deviation of y $= \sigma_y = \sqrt{\frac{1}{n}\left[\sum d_y^{\,2} - \frac{\left(\sum dy\right)^2}{n} \right]}$

Where $\bar{y} = B + \left(\frac{\sum d_y}{n} \right)$ where $d_y = y_i - B$

Here A = 150 and B = 190

x_i	y_i	$d_x=x_i-A$	$d_y=y_i-B$	dx_i^2	dy_i^2
204	99	54	-91	2916	8281
68	190	-82	0	6724	0
150	130	0	-60	0	3600
30	94	-120	-96	14400	9216
70	80	-80	-110	6400	12100
95	89	-55	-101	3025	10201
60	69	-90	-121	8100	14641
76	85	-74	-105	5476	11025
24	65	-126	-125	15876	15625
19	40	-131	-150	17161	22500
		$\Sigma d_x = -704$	$\Sigma d_y = -959$	$\Sigma d_x^2 = 80078$	$\Sigma d_y^2 = 107189$

$$\bar{x} = 150 + \left(\frac{-704}{10}\right)$$

$$= 79.6 \text{ runs}$$

$$\bar{y} = 190 + \left(\frac{-959}{10}\right)$$

$$= 94.1 \text{ runs}$$

$$\sigma_x = \sqrt{\frac{1}{10}\left[80078 - \frac{(-704)^2}{10}\right]}$$

$$= \sqrt{\frac{1}{10}[80078 - 49561.6]}$$

$$= 55.24 \text{ runs}$$

$$\sigma_y = \sqrt{\frac{1}{10}\left[107189 - \frac{(-959)^2}{10}\right]}$$

$$= \sqrt{\frac{1}{10}[107189 - 91968.1]}$$

$$= 39.01 \text{ runs}$$

$$(C.V.)_x = \frac{55.24}{79.6} X100$$

$$= 69.4\%$$

$$(C.V.)_y = \frac{39.01}{94.1} x100$$

$$= 41.46\%$$

Coefficient of variation of A is greater than coefficient of variation of B and hence we conclude that coefficient of player B is more consistent

NORMAL DISTRIBUTION

The Normal Distribution (N.D.) was first discovered by De-Moivre as the limiting form of the binomial model in 1733, later independently worked Laplace and Gauss.

The Normal distribution is 'probably' the most important distribution in statistics. It is a probability distribution of a continuous random variable and is often used to model the distribution of discrete random variable as well as the distribution of other continuous random variables. The basic from of normal distribution is that of a bell, it has single mode and is symmetric about its central values. The flexibility of using normal distribution is due to the fact that the curve may be centered over any number on the real line and it may be flat or peaked to correspond to the amount of dispersion in the values of random variable.

Definition: A random variable X is said to follow a Normal Distribution with parameter μ and and σ^2 if its density function is given by the probability law

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty;\ -\infty < \mu < \infty;\ \sigma > 0$$

where π = a mathematical constant equality = 22/7

e = Naperian base equaling 2.7183

μ = population mean

σ = population standard deviation

x = a given value of the random variable in the range $-\infty < x < \infty$

Characteristics of Normal distribution and normal curve:

The normal probability curve with mean μ and standard deviation σ is given by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad ; -\infty < x < \infty$$

and has the following properties

i. The curve is bell shaped and symmetrical, about the mean μ

ii. The height of normal curve is at its maximum at the mean. Hence the mean and mode of normal distribution coincides. Also the number of observations below the mean in a normal distribution is equal to the number of observations about the mean. Hence mean and median of N.D. coincides. Thus, N.D. has **Mean = median = mode**

iii. As 'x' increases numerically, f(x) decreases rapidly, the maximum probability occurring at the point x = μ, and given by

$$p[(x)]_{max} = \frac{1}{\sigma\sqrt{2\pi}}$$

iv. Skewness $\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}} = 0$

v. Kurtosis = $\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = 3$ ($\mu_i, \mu_2, \mu_3 and \mu_4$ are called central moments)

vi. All odd central moments are zero's

i.e. $\mu_1 = \mu_3 = \mu_5 = ... = 0$

vii. The first and third quartiles are equidistant from the median

viii. Linear combination of independent normal variates is also a normal variate

ix. The points of inflexion of the curve is given by

$$\left[x = \mu \pm \sigma, f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}} \right]$$

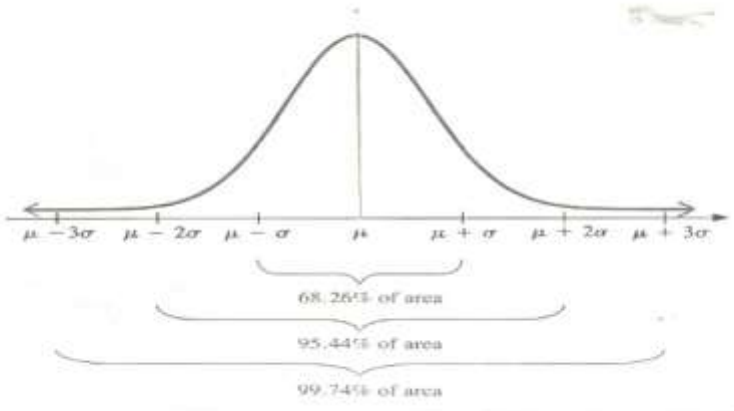
x. If $\int_{-\infty}^{+\infty} f(x)dx = 1$ then

the area under the normal curve is distributed as follows

i) μ-σ < x < μ+σ covers 68.26% of area

ii) μ-2σ < x < μ+2σ covers 95.44% of area

iii) μ-3σ < x < μ+3σ coves 99.73% of area



Area under Normal curve

The Normal Curve: The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

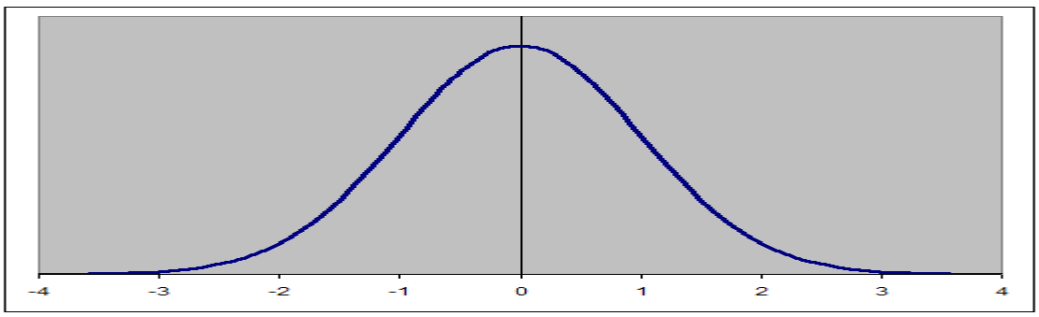
The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.

Standard Normal Distribution: If 'X' is a normal random variable with Mean μ and standard deviation σ, then Z = $\frac{X-\mu}{\sigma}$ is a standard normal variate with zero mean and standard deviation = 1.

The probability density function of standard normal variate 'z' is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad \text{and} \int_{-\omega}^{+\omega} f(z)dz = 1$$

A graph representing the density function of the Normal probability distribution is also known as a Normal Curve or a Bell Curve (see Figure below). To draw such a curve, one needs to specify two parameters, the mean and the standard deviation. The graph below has a mean of zero and a standard deviation of 1, i.e., (m =0, s =1). A Normal distribution with a mean of zero and a standard deviation of 1 is also known as the Standard Normal Distribution.



Standard Normal Distribution

Testing of Hypothesis

Introduction: The estimate based on sample values do not equal to the true value in the population due to inherent variation in the population. The samples drawn will have different estimates compared to the true value. It has to be verified that whether the difference between the sample estimate and the population value is due to sampling fluctuation or real difference. If the difference is due to sampling fluctuation only it can be safely said that the sample belongs to the population under question and if the difference is real we have every reason to believe that sample may not belong to the population under question. The following are a few technical terms in this context.

Hypothesis: The assumption made about any unknown characteristics is called hypothesis. It may or may be true.

Ex: 1. $\mu = 2.3$; μ be the population mean

2. $\sigma = 2.1$; σ be the population standard deviation

Population follows Normal Distribution. There are two types of hypothesis, namely null hypothesis and alternative hypothesis.

Null Hypothesis: Null hypothesis is the statement about the parameters. Such a hypothesis, which is usually a hypothesis of no difference is called null hypothesis and is usually denoted by H_0.

or

any statistical hypothesis under test is called null hypothesis. It is denoted by H_0.

1. H_0: $\mu = \mu_0$
2. H_0: $\mu_1 = \mu_2$

Alternative Hypothesis: Any hypothesis, which is complementary to the null hypothesis, is called an alternative hypothesis, usually denoted by H_1.

Ex: 1. H_1: $\mu \# \mu_0$

2. H_1: $\mu_1 \# \mu_2$

Parameter: A characteristics of population values is known as parameter. For example, population mean (μ) and population variance (σ^2).

In practice, if parameter values are not known and the estimates based on the sample values are generally used.

Statistic: A Characteristics of sample values is called a statistic. For example, sample mean ($\bar{x}$), sample variance (s^2) where $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$

$$\text{and } s^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i^{\,2}\right)}{n}\right]$$

Sampling distribution: The distribution of a statistic computed from all possible samples is known as sampling distribution of that statistic.

Standard error: The standard deviation of the sampling distribution of a statistic is known as its standard error, abbreviated as S.E.

S.E.($\bar{x}$) = $\frac{\sigma}{\sqrt{n}}$; where σ = population standard deviation and n = sample size

Sample: A finite subset of statistical objects in a population is called a sample and the number of objects in a sample is called the sample size.

Population: In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to objects belonging to a group. This group of objects under study is called population or universe.

Random sampling: If the sampling units in a population are drawn independently with equal chance, to be included in the sample then the sampling will be called random sampling. It is also referred as simple random sampling and denoted as SRS. Thus, if the population consists of 'N' units the chance of selecting any unit is 1/N. A theoretical definition of SRS is as follows: Suppose we draw a sample of size 'n' from a population size N; then there are $\binom{N}{n}$ possible samples of size 'n'. If all possible samples have an equal chance, $\frac{1}{\binom{N}{n}}$ of being drawn, then the sampling is said to be simple random sampling.

Simple Hypothesis: A hypothesis is said to be simple if it completely specifies the distribution of the population. For instance, in case of normal population with mean μ

and standard deviation σ, a simple null hypothesis is of the form H_0: $\mu = \mu_0$, σ is known, knowledge about μ would be enough to understand the entire distribution. For such a test, the probability of committing the type-1 error is expressed as exactly α.

Composite Hypothesis: If the hypothesis does not specify the distribution of the population completely, it is said to be a composite hypothesis. Following are some examples:

$H_0 : \mu \leq \mu_0$ and σ is known

$H_0 : \mu \geq \mu_0$ and σ is known

All these are composite because none of them specifies the distribution completely. Hence, for such a test the LOS is specified not as α but as 'at most α'.

Types of Errors: In testing of statistical hypothesis there are four possible types of decisions

1. Rejecting H_0 when H_0 is true
2. Rejecting H_0 when H_0 is false
3. Accepting H_0 when H_0 is true
4. Accepting H_0 when H_0 is false

1 and 4th possibilities leads to error decisions. Statistician gives specific names to these concepts namely Type-I error and Type-II error respectively.

the above decisions can be arranged in the following table

	H_0 is true	H_0 is false
Rejecting H_0	Type-I error (Wrong decision)	Correct
Accepting H_0	Correct	Type-II error

Type-I error: Rejecting H_0 when H_0 is true

Type-II error: Accepting H_0 when H_0 is false

The probabilities of type-I and type-II errors are denoted by α and β respectively.

Degrees of freedom: It is defined as the difference between the total number of items and the total number of constraints.

If 'n' is the total number of items and 'k' the total number of constraints then the degrees of freedom (d.f.) is given by d.f. = n-k

Level of significance(LOS): The maximum probability at which we would be willing to risk a type-I error is known as level of significance or the size of Type-I error is level of significance. The level of significance usually employed in testing of hypothesis are 5%

and 1%. The Level of significance is always fixed in advance before collecting the sample information. LOS 5% means the results obtained will be true is 95% out of 100 cases and the results may be wrong is 5 out of 100 cases.

Critical value: while testing for the difference between the means of two populations, our concern is whether the observed difference is too large to believe that it has occurred just by chance. But then the question is how much difference should be treated as too large? Based on sampling distribution of the means, it is possible to define a cut-off or threshold value such that if the difference exceeds this value, we say that it is not an occurrence by chance and hence there is sufficient evidence to claim that the means are different. Such a value is called the critical value and it is based on the level of significance.

Steps involved in test of hypothesis:

1. The null and alternative hypothesis will be formulated
2. Test statistic will be constructed
3. Level of significance will be fixed
4. The table (critical) values will be found out from the tables for a given level of significance
5. The null hypothesis will be rejected at the given level of significance if the value of test statistic is greater than or equal to the critical value. Otherwise null hypothesis will be accepted.
6. In the case of rejection the variation in the estimates will be called 'significant' variation. In the case of acceptance the variation in the estimates will be called 'not-significant'.

STANDARD NORMAL DEVIATE TESTS

OR

LARGE SAMPLE TESTS

If the sample size n >30 then it is considered as large sample and if the sample size n< 30 then it is considered as small sample.

SND Test or One Sample (Z-test)

Case-I: Population standard deviation (σ) is known

Assumptions:

1. Population is normally distributed
2. The sample is drawn at random

Conditions:

1. Population standard deviation σ is known
2. Size of the sample is large (say n > 30)

Procedure: Let $x_1, x_2, \ldots\ldots x_n$ be a random sample size of n from a normal population with mean μ and variance σ^2.

Let $\bar{x}$ be the sample mean of sample of size 'n'

Null hypothesis (H_0): population mean (μ) is equal to a specified value μ_0

i.e. $H_0 : \mu = \mu_0$

Under H_0, the test statistic is

$$Z = \frac{|\bar{x} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

i.e the above statistic follows Normal Distribution with mean '0' and varaince'1'.

If the calculated value of $|Z|$ < table value of Z at 5% level of significance, H_0 is accepted and hence we conclude that there is no significant difference between the population mean and the one specified in H_0 as μ_0.

Case-II: If σ is not known

Assumptions:

1. Population is normally distributed
2. Sample is drawn from the population should be random
3. We should know the population mean

Conditions:

1. Population standard deviation σ is not known
2. Size of the sample is large (say n > 30)

Null hypothesis (H_0) : $\mu = \mu_0$

under H_0, the test statistic

$$Z = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} \sim N(0,1) \qquad \text{where } s = \sqrt{\frac{1}{n}\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)}$$

$\bar{x}$ = Sample mean; n = sample size

If the calculated value of Z < table value of Z at 5% level of significance, H_0 is accepted and hence we conclude that there is no significant difference between the population mean and the one specified in H_0 otherwise we do not accept H_0.

The table value of Z at 5% level of significance = 1.96 and table value of Z at 1% level of significance = 2.58.

Two sample Z-Test or Test of significant for difference of means

Case-I: when σ is known

Assumptions:

1. Populations are distributed normally
2. Samples are drawn independently and at random

Conditions:

1. Populations standard deviation σ is known
2. Size of samples are large

Procedure: Let $\bar{x_1}$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2

Let $\bar{x_2}$ be the mean of a random sample of size n_2 from another population with mean μ_2 and variance σ_2^2

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

i.e. The null hypothesis states that the population means of the two samples are identical.

Under the null hypothesis the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\left(\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}\right)}} \sim N(0,1) \quad \rightarrow (1)$$

i.e the above statistic follows Normal Distribution with mean '0' and varaince'1'.

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say) i.e both samples have the same standard deviation then the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \quad \rightarrow (2)$$

If the calculated value of $|Z|$ < table value of Z at 5% level of significance, H_0 is accepted otherwise rejected.

If H_0 is accepted means, there is no significant difference between two population means of the two samples are identical.

Example: The Average panicle length of 60 paddy plants in field No. 1 is 18.5. cms and that of 70 paddy plants in field No. 2 is 20.3 cms. With common S.D. 1.15 cms. Test whether there is significant difference between tow paddy fields w.r.t panicle length.

Solution:

Null hypothesis: H_0: There is no significant difference between the means of two paddy fields w.r.t. panicle length.

H_0: $\mu_1 = \mu_2$

Under H_0, the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \text{ ---------- } (1)$$

Where $\overline{x_1}$ = first field sample mean = 18.5 inches

$\overline{x_2}$ = second field sample mean = 20.3 inches

n_1 = first sample size = 60

n_2 = second sample size = 70

σ = common S.D. = 1.15 inches

Substitute the given values in equation (1), we get

$$Z = \frac{|18.5 - 20.3|}{1.15\sqrt{\frac{1}{60} + \frac{1}{70}}} = \frac{1.8}{1.15x0.176} = 8.89$$

Calculated value of $|Z| = 5.1$

Cal. Value of $|Z|$ > table value of Z at 5% LOS(1.96), H_0 is rejected. This means, there is highly significant difference between two paddy fields w.r.t. panicle length.

Case-II: when σ is not known

Assumptions:

1. Populations are normally distributed
2. Samples are drawn independently and at random

Conditions:

1. Population standard deviation σ is not known
2. Size of samples are large

Null hypothesis $H_0 : \mu_1 = \mu_2$

Under the null hypothesis the test statistic becomes

$$Z = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)}} \sim N(0,1) \quad \rightarrow (2)$$

Where $\overline{x_1}$ = 1st sample mean , $\overline{x_2}$ = 2nd sample mean

${s_1}^2$ = 1st sample variance, ${s_2}^2$ = 2nd sample variance

n_1 = 1st sample size, n_2 = 2nd sample size

If the calculated value of $|Z|$ < table value of Z at 5% level of significance, H_0 is accepted otherwise rejected.

Example:

A breeder claims that the number of filled grains per panicle is more in a new variety of paddy ACM.5 compared to that of an old variety ADT.36. To verify his claim a random sample of 50 plants of ACM.5 and 60 plants of ADT.36 were selected from the experimental fields. The following results were obtained:

(ForACM.5)	(For ADT.36)
$\overline{x_1}$ = 139.4-grains/panicle	$\overline{x_2}$ = 112.9 grains/panicle
s_1 = 26.864	s_2 = 20.1096
n_1 = 50	n_2 = 60

Test whether the claim of the breeder is correct.

Sol: Null hypothesis $H_0 : \mu_1 = \mu_2$

(i.e. the average number of filled grains per panicle is the same for both ACM.5 and ADT.36)

Under H_0, the test statistic becomes

$$Z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1) \text{ ---------- (1)}$$

Where $\overline{x_1}$ = first variety sample mean = 139.4 grains/panicle

$\overline{x_2}$ = Second variety sample mean = 112.9 grains/panicle

s_1= first sample standard deviation = 26.864

s_2= second sample standard deviation = 20.1096

n_1 = first sample size = 50

and n_2 = second sample size = 60

Substitute the given values in equation (1), we get

$$Z = \frac{|139.4 - 112.9|}{\sqrt{\frac{(26.864)^2}{50} + \frac{(20.1096)^2}{60}}}$$

$$= \frac{|26.5|}{\sqrt{14.4335 + 6.7399}}$$

$$= 4.76$$

Calculated value of Z > Table value of Z at 5% LOS (1.96), H_0 is rejected. We conclude that the number of filled grains per panicle is significantly greater in ACM.5 than in ADT.36

SMALL SAMPLE TESTS

The entire large sample theory was based on the application of "normal test". However, if the sample size 'n' is small, the distribution of the various statistics, e.g., Z = $\frac{|\bar{x} - \mu_0|}{\frac{\sigma}{\sqrt{n}}}$ are far from normality and as such 'normal test' cannot be applied if 'n' is small.

In such cases exact sample tests, we use t-test pioneered by W.S. Gosset (1908) who wrote under the pen name of student, and later on developed and extended by Prof. R.A. Fisher.

Student's t-test: Let $x_1, x_2, \ldots\ldots\ldots, x_n$ be a random sample of size 'n' has drawn from a normal population with mean μ and variance σ^2 then student's t – is defined by the statistic

$$t = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}}$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - \frac{\left\{ \sum_{i=1}^{n} x_i \right\}^2}{n} \right)$

this test statistic follows a t – distribution with (n-1) degrees of freedom (d.f.). To get the critical value of t we have to refer the table for t-distribution against (n-1) d.f. and the specific level of significance. Comparing the calculated value of t with critical value, we can accept or reject the null hypothesis.

The Range of t – distribution is $-\infty$ to $+\infty$.

One Sample t – test

One sample t-test is a statistical procedure that is used to know the population mean and the known value of the population mean. In one sample t-test, we know the population mean. We draw a random sample from the population and then compare the sample mean with population mean and make a statistical decision as to whether or not the sample mean is different from the population mean. In one sample t-test, sample size should be less than 30.

Assumptions: 1. Population is normally distributed

2. Sample is drawn from the population and it should be random

3. We should know the population mean

Conditions: 1. Population S.D. σ is not known

2. Size of the sample is small (<30).

Procedure: Let : Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample of size 'n' has drawn from a normal population with mean μ and variance σ^2.

Null hypothesis (H_0): population mean (μ) is equal to a specified value μ_0

$$\text{i.e. } H_0: \mu = \mu_0$$

Under H_0, the test statistic becomes

$$t = \frac{\left|\bar{x} - \mu_0\right|}{\frac{s}{\sqrt{n}}}$$

and follows student's t – distribution with (n-1) degrees of freedom.

$$\text{where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

We now compare the calculated value of t with the tabulated value at certain level of significance

If calculated value of $|t|$ < table value of t at (n-1) d.f., the null hypothesis is accepted and hence we conclude that there is no significant difference between the population mean and the one specified in H_0 as μ_0.

Example: Based on field experiments, a new variety of greengram is expected to give an yield of 13 quintals per hectare. The variety was tested on 12 randomly selected farmer fields. The yields (quintal/hectare) were recorded as 14.3, 12.6, 13.7, 10.9,13.7, 12.0, 11.4, 12.0, 13.1, 12.6, 13.4 and 13.1. Do the results conform the expectation?

Solution:

Null Hypothesis: $H_0 : \mu = \mu_0 = 13$

i.e. the results conform the expectation

The test statistic becomes

$$t = \frac{\left|\bar{x} - \mu_0\right|}{\frac{s}{\sqrt{n}}} \sim t \text{ (n-1) d.f.}$$

Where $\bar{x} = \frac{\sum x}{n}$ and $s = \sqrt{\frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}$ is an unbiased estimate of σ

Let yield $= x_i$ (say)

x_i	x_i^2
14.3	204.49
12.6	158.76
13.7	187.69
10.9	118.81
13.7	187.69
12	144
11.4	129.96
12	144
13.1	171.61
12.6	158.76
13.4	179.56
13.1	171.61
$\Sigma x = 152.8$	$\Sigma x^2 = 1956.94$

$$\bar{x} = \frac{152.8}{12} = 12.73$$

$$s = \sqrt{\frac{1}{11}\left(1956.94 - \frac{(152.8)^2}{12}\right)}$$

$$= 1.01$$

$$t = \frac{|12.73 - 13|}{\frac{1.01}{\sqrt{12}}} = \frac{0.27}{0.29} = 0.93 \text{ qa/h.}$$

t-table value at (n-1) = 11d.f. at 5 percent level of significance is 2.20.

Calculated value of t < table value of t, H_0 is accepted and we may conclude that the results conform to the expectation.

t-test for Two Samples

Assumptions: 1. Populations are distributed normally

2. Samples are drawn independently and at random

Conditions: 1. Standard deviations in the populations are same and not known

2. Size of the sample is small

Procedure: If two independent samples x_i (i = 1,2,….,n_1) and y_j (j = 1,2, …..,n_2) of sizes $n_1$and n_2 have been drawn from two normal populations with means μ_1 and μ_2 respectively.

Null hypothesis $H_0 : \mu_1 = \mu_2$

The null hypothesis states that the population means of the two groups are identical, so their difference is zero.

Under H_0, the test statistics is $t = \dfrac{|\bar{x} - \bar{y}|}{S\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Where $S^2 = \quad = \dfrac{1}{n_1 + n_2 - 2}\left[\sum x^2 - \dfrac{(\sum x)^2}{n_1} + \sum y^2 - \dfrac{(\sum y)^2}{n_2}\right]$

or

$$S^2 = \text{pooled variance} = \frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}$$

where s_1^2 and s_2^2 are the variances of the first and second samples respectively.

and $\bar{x} = \dfrac{\sum_{i=1}^{n_1} x_i}{n_1}$ and $\bar{y} = \dfrac{\sum_{j=1}^{n_2} y_j}{n_2}$; where $\bar{x}$ and $\bar{y}$ are the two sample means.

Which follows Student's t – distribution with (n_1+n_2-2) d.f.

If calculated value of $|t|$ < table value of t with (n_1+n_2-2) d.f. at specified level of significance, then the null hypothesis is accepted otherwise rejected.

Example: Two verities of potato plants (A and B) yielded tubes are shown in the following table. Does the mean number of tubes of the variety 'A' significantly differ from that of variety B?

Tuber yield, kg/plant

Variety-A	2.2	2.5	1.9	2.6	2.6	2.3	1.8	2.0	2.1	2.4	2.3
Variety-B	2.8	2.5	2.7	3.0	3.1	2.3	2.4	3.2	2.5	2.9	

Solution:

Hypothesis $H_0 : \mu_1 = \mu_2$

i.e the mean number of tubes of the variety 'A' significantly differ from the variety 'B'

Statistic $t = \dfrac{|\bar{x} - \bar{y}|}{S\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2)d.f$

$n_1 = 1^{st}$ sample size; $n_2 = 2^{nd}$ sample size

$\bar{x}$ = Mean of the first sample; $\bar{y}$ = mean of the second sample

$\bar{x} = \frac{\sum x}{n_1}$ and $\bar{y} = \frac{\sum y}{n_2}$

Where $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\left\{\sum x^2 - \frac{(\sum x)^2}{n_1}\right\} + \left\{\sum y^2 - \frac{(\sum y)^2}{n_2}\right\}\right]$

x	y	x^2	y^2
2.20	2.80	4.84	7.84
2.50	2.50	6.25	6.25
1.90	2.70	3.61	7.29
2.60	3.00	6.76	9.00
2.60	3.10	6.76	9.61
2.30	2.30	5.29	5.29
1.80	2.40	3.24	5.76
2.00	3.20	4.00	10.24
2.10	2.50	4.41	6.25
2.40	2.90	5.76	8.41
2.30	$\Sigma y = 27.40$	5.29	$\Sigma y^2 = 75.94$
$\Sigma x = 24.70$		$\Sigma x^2 = 56.21$	

$\bar{x} = \frac{24.70}{11}$ $\bar{y} = \frac{27.40}{10}$

$= 2.25$ Kg $= 2.74$ Kg

Where $S^2 = \frac{1}{11+10-2}\left[\left\{56.21 - \frac{(24.70)^2}{11}\right\} + \left\{75.94 - \frac{(27.40)^2}{10}\right\}\right]$

$= \frac{1}{19}[\{56.21 - 55.46\} + \{75.94 - 75.07\}]$

$= 0.09$ Kg^2

$S = \sqrt{S^2} = 0.3$ Kg.

Test statistic $t = \frac{|2.25 - 2.74|}{0.3\sqrt{\left(\frac{1}{11} + \frac{1}{10}\right)}}$

$= \frac{0.49}{0.13} = 3.77$

Calculated value of t = 3.77

Table value of t for 19 d.f. at 5 % level of significance is 2.09

Since the calculated value of t > table value of t, the null hypothesis is rejected and hence we conclude that the mean number of tubes of the variety ‘A’ significantly not differ from the variety ‘B’

Paired t – test

The paired t-test is generally used when measurements are taken from the same subject before and after some manipulation such as injection of a drug. For example, you can use a paired t test to determine the significance of a difference in blood pressure before and after administration of an experimental pressor substance.

Assumptions: 1. Populations are distributed normally

2. Samples are drawn independently and at random

Conditions: 1. Samples are related with each other

2. Sizes of the samples are small and equal

3. Standard deviations in the populations are equal and not known

Hypothesis H_0: $\mu_d = 0$

Under H_0, the test statistic becomes,

$$t = \frac{|\bar{d}|}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)} \text{ d.f.}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ and $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^{\,2} - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right]$

where S^2 is variance of the deviations

n = sample size; where $d_i = x_i - y_i$ (i = 1,2,……,n)

If calculated value of |t| < table value of t for (n-1)d.f. at α% level of significance, then the null hypothesis is accepted and hence we conclude that the two samples may belong to the same population. Otherwise, the null hypothesis rejected.

Example: The average number of seeds set per pod in Lucerne were determined for top flowers and bottom flowers in ten plants. The values observed were as follows:

Top flowers	4.2	5.0	5.4	4.3	4.8	3.9	4.2	3.1	4.4	5.8
Bottom flowers	4.6	3.5	4.8	3.0	4.1	4.4	3.6	3.8	3.2	2.2

Test whether there is any significant difference between the top and bottom flowers with respect to average numbers of seeds set per pod.

Solution:

Null Hypothesis H_0: $\mu_d = 0$

Under H_0 becomes, the test statistic is

$$t = \frac{|\bar{d}|}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)} d.f.$$

Where $\bar{d} = \frac{\sum d}{n}$ and $s^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$

x	y	d=x-y	d^2
4.2	4.6	-0.40	0.16
5.0	3.5	1.50	2.25
5.4	4.8	0.60	0.36
4.3	3.0	1.30	1.69
4.8	4.1	0.70	0.49
3.9	4.4	-0.50	0.25
4.2	3.6	0.60	0.36
3.1	3.8	-0.70	0.49
4.4	3.2	1.20	1.44
5.8	2.2	3.60	12.96
		Σd = 7.90	Σd^2 = 20.45

$$\bar{d} = \frac{\sum d}{n} = \frac{7.90}{10}$$

$$= 0.79$$

$$s^2 = \frac{1}{9}\left[20.45 - \frac{(7.90)^2}{10}\right]$$

$$= 2.27$$

$$s = \sqrt{s^2} = \sqrt{2.27}$$

$$= 1.51$$

$$t = \frac{0.79}{\frac{1.51}{\sqrt{10}}}$$

$$= 1.65$$

Calculated value of t = 1.65

Table value of t for 9 d.f. at 5% level of significance is 2.26

Calculated value of t < table value of t, the null hypothesis is accepted and we conclude that there is no significant difference between the top and bottom flowers with respect to average numbers of seeds set per pod.

F – Test

In agricultural experiments the performance of a treatment is assessed not only by its mean but also by its variability. Hence, it is of interest to us to compare the variability of two populations. In testing of hypothesis the equality of variances, the greater variance is always placed in the Numerator and smaller variance is placed in the denominator.

F- test is used to test the equality of two population variances, equality of several regression coefficients, ANOVA .

F- test was discovered by G.W. Snedecor. The range of F : 0 to ∞

Let $x_1, x_2, \ldots\ldots\ldots\ldots, xn_1$ and $y_1, y_2, \ldots\ldots y n_2$ be the two independent random samples of sizes n_1 and n_2 drawn from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. S_1^2 and S_2^2 are the sample variances of the two samples.

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$

Under H_0, the test statistic becomes

$$F = \frac{S_1^2}{S_2^2} \quad \text{where, } S_1^2 > S_2^2$$

Which follows F-distribution with (n_1-1, n_2-1)d.f.

Where $S_1^2 = \frac{1}{n_1 - 1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n_1}\right]$ and $S_2^2 = \frac{1}{n_2 - 1}\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n_2}\right]$

or $$F = \frac{S_2^2}{S_1^2} \quad \text{where } S_2^2 > S_1^2$$

Which follows F-distribution with (n_2-1, n_1-1)d.f.

If calculated value of F < table value of F with (n_2-1, n_1-1)d.f at specified level of significance, then the null hypothesis is accepted and hence we conclude that the variances of the populations are homogeneous otherwise heterogeneous.

Example: The heights in meters of red gram plants with two types of irrigation in two fields are as follows:

Tap water (x)	3.5	4.2	2.8	5.2	1.7	2.6	3.5	4.2	5.0	5.2
Saline water (y)	1.9	2.6	2.3	4.3	4.0	4.2	3.8	2.9	3.7	

Test whether the variances of the two system of irrigation are homogeneous.

Solution:

H_0: The variances of the two systems of irrigation are homogeneous.

i.e. $\sigma_1^2 = \sigma_2^2$

Under H_0, the test statistic becomes

$$F = \frac{S_1^2}{S_2^2} \quad ; \quad (S_1^2 > S_2^2)$$

Where S_1^2 = fist sample variance $= \frac{1}{n_1 - 1}\left(\sum x^2 - \frac{(\sum x)^2}{n_1}\right)$

and S_2^2 = second sample variance = $\frac{1}{n_2 - 1}\left(\sum y^2 - \frac{(\sum y)^2}{n_2}\right)$

x	y	x^2	y^2
3.5	1.9	12.25	3.61
4.2	2.6	17.64	6.76
2.8	2.3	7.84	5.29
5.2	4.3	27.04	18.49
1.7	4	2.89	16
2.6	4.2	6.76	17.64
3.5	3.8	12.25	14.44
4.2	2.9	17.64	8.41
5	3.7	25	13.69
5.2	$\sum y = 29.7$	27.04	$\sum y^2 = 104.33$
$\sum x$ =37.9		$\sum x^2 = 156.35$	

$$S_1^2 = \frac{1}{9}\left(156.35 - \frac{(37.9)^2}{10}\right) = 1.41 \text{ mt}^2$$

and

$$S_2^2 = \frac{1}{8}\left(104.33 - \frac{(29.7)^2}{9}\right) = 0.79 \text{ mt}^2$$

$$F = \frac{S_1^2}{S_2^2} = \frac{1.41}{0.79} = 1.78$$

F calculated value = 1.78

Table value of $F_{0.05}$ for (n_1-1, n_2-1) d.f. = 3.23

Calculated value of F < Table value of at 5% level of significance, H_0 is accepted and hence we conclude that the variances of the two systems of irrigation are homogeneous.

$$\frac{1}{F_i} = F_2 \quad \text{or} \quad F_1 = \frac{1}{F_2}$$

Chi-square (χ^2) test

The various tests of significance studied earlier such that as Z-test, t-test, F-test were based on the assumption that the samples were drawn from normal population. Under this assumption the various statistics were normally distributed. Since the procedure of testing the significance requires the knowledge about the type of population or parameters of population from which random samples have been drawn, these tests are known as parametric tests.

But there are many practical situations in which the assumption of any kind about the distribution of population or its parameter is not possible to make. The alternative technique where no assumption about the distribution or about parameters of population is made are known as non-parametric tests. Chi-square test is an example of the non parametric test. Chi-square distribution is a distribution free test.

If $X_i \rightarrow N(0,1)$ then $\sum x_i^2 \sim \chi^2_n$

Chi-square distribution was first discovered by Helmert in 1876 and later independently by Karl Pearson in 1900. The range of chi-square distribution is 0 to ∞.

Measuremental data: the data obtained by actual measurement is called measuremental data. For example, height, weight, age, income, area etc.,

Enumeration data: the data obtained by enumeration or counting is called enumeration data. For example, number of blue flowers, number of intelligent boys, number of curled leaves, etc.,

χ^2 – test is used for enumeration data which generally relate to discrete variable where as t-test and standard normal deviate tests are used for measure mental data which generally relate to continuous variable.

χ^2 –test can be used to know whether the given objects are segregating in a theoretical ratio or whether the two attributes are independent in a contingency table.

The expression for χ^2 –test for goodness of fit

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i = observed frequencies

E_i = expected frequencies

n = number of cells(or classes)

Which follows a chi-square distribution with (n-1) degrees of freedom

The null hypothesis H_0 = the observed frequencies are in agreement with the expected frequencies

If the calculated value of χ^2 < Table value of χ^2 with (n-1) d.f. at specified level of significance (α), we accept H_0 otherwise we do not accept H_0.

Conditions for the validity of χ^2 –test:

The validity of χ^2-test of goodness of fit between theoretical and observed, the following conditions must be satisfied.

i) The sample observations should be independent

ii) Constraints on the cell frequencies, if any, should be linear $\Sigma o_i = \Sigma e_i$

iii) N, the total frequency should be reasonably large, say greater than 50

iv) If any theoretical (expected) cell frequency is < 5, then for the application of chi-square test it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

Applications of Chi-square Test:

1. testing the independence of attributes
2. to test the goodness of fit
3. testing of linkage in genetic problems
4. comparison of sample variance with population variance
5. testing the homogeneity of variances
6. testing the homogeneity of correlation coefficient

Test for independence of two Attributes of (2x2) Contingency Table:

A characteristic which can not be measured but can only be classified to one of the different levels of the character under consideration is called an attribute.

2x2 contingency table: When the individuals (objects) are classified into two categories with respect to each of the two attributes then the table showing frequencies distributed over 2x2 classes is called 2x2 contingency table.

Suppose the individuals are classified according to two attributes say intelligence (A) and colour (B). The distribution of frequencies over cells is shown in the following table.

A / B	A_1	A_2	Row totals
B_1	a	B	R_1+(a+b)
B_2	c	D	R_2 = (c+d)
Column total	C_1 = (a+c)	C_2 = (b+d)	N = (R_1+R_2) or (C_1+C_2)

Where R_1 and R_2 are the marginal totals of 1st row and 2nd row

C_1 and C_2 are the marginal totals of 1st column and 2nd column

N = grand total

The null hypothesis H_0: the two attributes are independent (if the colour is not dependent on intelligent)

Based on above H_0, the expected frequencies are calculated as follows.

$$E(a) = \frac{R_1 x C_1}{N}; \quad E(b) = \frac{R_1 x C_2}{N}; \quad E(c) = \frac{R_2 x C_1}{N}; \quad E(d) = \frac{R_2 x C_2}{N}$$

Where N = a+b+c+d

To test this hypothesis we use the test statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

the degrees of freedom for mxn contingency table is (m-1)x(n-1)

the degrees of freedom for 2x2 contingency table is (2-1)(2-1) = 1

This method is applied for all rxc contingency tables to get the expected frequencies.

The degrees of freedom for rxc contingency table is (r-1)x(c-1)

If the calculated value of χ^2 < table value of χ^2 at certain level of significance, then H_0 is accepted otherwise we do not accept H_0

The alternative formula for calculating χ^2 in 2x2 contingency table is

$$\chi^2 = \frac{(ad - bc)^2 N}{R_1 x R_2 x C_1 x C_2}$$

Example: Examine the following table showing the number of plants having certain characters, test the hypothesis that the flower colour is independent of the shape of leaf.

Flower colour	Shape of leaf		Totals
	Flat leaves	Curled leaves	
White flowers	99 (a)	36 (b)	R_1 = 135
Red flowers	20(c)	5 (d)	R_2 = 25
Totals	C_1 = 119	C_2 = 41	N = 160

Solution:

Null hypothesis H_0: attributes 'flower colour' and 'shape of leaf' are independent of each other.

Under H_0 the statistic is

$$\chi^2 = \frac{\sum_{i=1}^{n}(o_i - e_i)^2}{e_i}$$

where o_i = observed frequency

and e_i = expected frequency

Expected frequencies are calculated as follows.

E(a) = $\frac{R_1 * C_1}{N} = \frac{135*119}{160} = 100.40$ where R_1 and R_2 = Row totals

E(b) = $\frac{R_1 * C_2}{N} = \frac{135*41}{160} = 34.59$ C_1 and C_2 = column totals

E(c) = $\frac{R_2 * C_1}{N} = \frac{25*119}{160} = 18.59$ N= Grand totals

E(d) = $\frac{R_2 * C_2}{N} = \frac{25*41}{160} = 6.406$

o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2$	$\frac{(o_i - e_i)2}{e_i}$
99	100.40	-1.4	1.96	0.02
36	34.59	1.41	1.99	0.06
20	18.59	1.41	1.99	0.11
5	6.41	-1.41	1.99	0.31
				$\frac{\sum_{i=1}^{n}(o_i - e_i)^2}{e_i} = 0.49$

Calculated value of $\chi^2 = \frac{\sum_{i=1}^{n}(o_i - e_i)^2}{e_i} = 0.49$

Direct Method:

Statistic: $\chi^2 = \frac{N(ad-bc)^2}{R_1 R_2 C_1 C_2}$

here a = 99, b = 36, c = 20 and d = 5 and N = 160

$$\chi^2 = \frac{160(99*5 - 36*20)^2}{135*25*119*41}$$

$$= \frac{160 * 50625}{164666.25}$$

$$= \frac{18100000}{16466625} = 0.49$$

Calculated value of $\chi^2 = 0.40$

Table value of χ^2 for (2-1) (2-1) = 1 d.f. is 3.84

Calculated value of χ^2 < Table value of χ^2 at 5% LOS for 1 d.f. , Null hypothesis is accepted and hence we conclude that two characters, flower colour and shape of leaf are independent of each other.

Yates correction for continuity in a 2x2 contingency table:

In a 2x2 contingency table, the number of d.f. is (2-1)(2-1) = 1. If any one of Expected cell frequency is less than 5, then we use of pooling method for χ^2 –test results with `0' d.f. (since 1 d.f. is lost in pooling) which is meaningless. In this case we apply a correction due to Yates, which is usually known a Yates correction for continuity.

Yates correction consists of the following steps; (1) add 0.5 to the cell frequency which is the least, (2) adjust the remaining cell frequencies in such a way that the row and column totals are not changed. It can be shown that this correction will result in the formula.

$$\chi^2 \text{ (corrected)} = \frac{N\left[|ad - bc| - \frac{N}{2}\right]^2}{R_1 R_2 C_1 C_2}$$

Example: The following data are observed for hybrids of Datura.

Flowers violet, fruits prickly =47

Flowers violet, fruits smooth = 12

Flowers white, fruits prickly = 21

Flowers white, fruits smooth = 3. Using chi-square test, find the association between colour of flowers and character of fruits.

Sol: H_0: The two attributes colour of flowers and fruits are independent.

We cannot use Yate's correction for continuity based on observed values. If only expected frequency less than 5, we use Yates's correction for continuity.

The test statistic is

$$\chi^2_{\text{(corrected)}} = \frac{N\left[|ad-bc|-\frac{N}{2}\right]^2}{R_1R_2C_1C_2}$$

	Flowers Violet	Flowers white	Total
Fruits Prickly	47(48.34)	21(19.66)	68
Fruits smooth	12(10.66)	3(4.34)	15
Total	59	24	83

The figures in the brackets are the expected frequencies

$$\chi^2_{(corrected)} = \frac{83\left[|(47*3)-(21*12)|-\frac{83}{2}\right]^2}{68*15*59*24}$$

$$= \frac{83\left[|141-252|-41.5\right]^2}{68*15*59*24}$$

$$= \frac{400910.75}{1444320} = 0.28$$

Calculated value of $\chi^2 = 0.28$

Table value of χ^2 for (2-1) (2-1) = 1 d.f. is 3.84

Calculated value of χ^2 < table value of χ^2, H_0 is accepted and hence we conclude that colour of flowers and character of fruits are not associated

CORRELATION

When there are two continuous variables which are concomitant their joint distribution is known as bivariate normal distribution. If there are more than two such variables their joint distribution is known as multivariate normal distributions. In case of bivariate or multivariate normal distributions, we may be interested in discovering and measuring the magnitude and direction of the relationship between two or more variables. For this purpose we use the statistical tool known as correlation.

Definition: If the change in one variable affects a change in the other variable, the two variables are said to be correlated and the degree of association ship (or extent of the relationship) is known as correlation.

Types of correlation:

a). Positive correlation: If the two variables deviate in the same direction, i.e., if the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other variable, correlation is said to be direct or positive.

Ex: (i) Heights and weights

(ii) Household income and expenditure

(iii) Amount of rainfall and yield of crops

(iv) Prices and supply of commodities

(v) Feed and milk yield of an animal

(vi) Soluble nitrogen and total chlorophyll in the leaves of paddy.

b). Negative correlation: If the two variables constantly deviate in the opposite direction i.e., if increase (or decrease) in one variable results in corresponding decrease (or increase) in the other variable, correlation is said to be inverse or negative.

Ex: (i) Price and demand of a goods

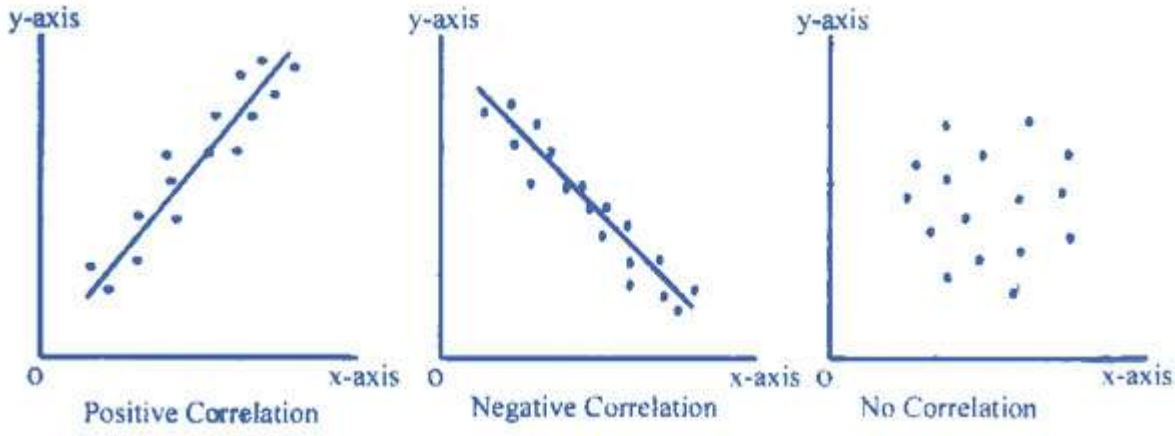
(ii) Volume and pressure of perfect gas

(iii) Sales of woolen garments and the day temperature

(iv) Yield of crop and plant infestation

c) No or Zero Correlation: If there is no relationship between the two variables such that the value of one variable change and the other variable remain constant is called no or zero correlation.

Figures:

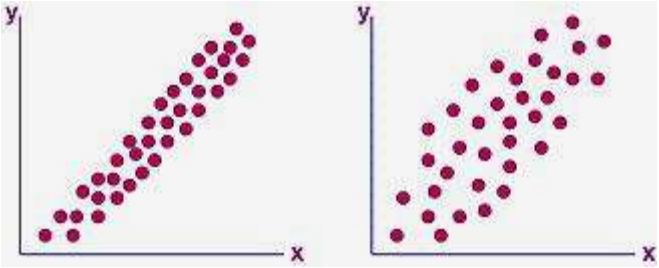


Methods of studying correlation: 1. Scatter Diagram 2. Karl Pearson's Coefficient of Correlation 3. Spearman's Rank Correlation 4. Regression Lines

1. Scatter diagram: It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution (x_i,y_i); $i = j = 1,2,\ldots n$, If the values of the variables X and Y be plotted along the X-axis and Y-axis respectively in the xy-plane, the diagram of dots so obtained is known as scatter diagram. From the scatter diagram, if the points are very close to each other, we should expect a fairly good amount of correlation

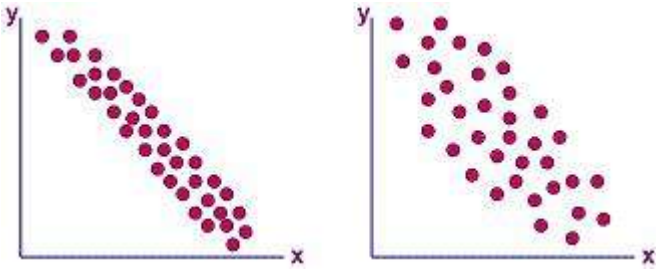
between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is fairly large.

If the plotted points shows an upward trend of a straight line then we say that both the variables are positively correlated



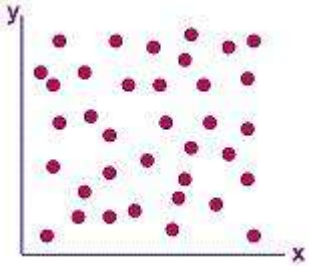
Positive Correlation

When the plotted points shows a downward trend of a straight line then we say that both the variables are negatively correlated



Negative Correlation

If the plotted points spread on whole of the graph sheet, then we say that both the variables are not correlated.



No Correlation

Karl Pearson's Coefficient of Correlation: Prof. Karl Pearson, a British Biometrician suggested a measure of correlation between two variables. It is known as Karl Pearson's coefficient of correlation. It is useful for measuring the degree of linear relationship between the two variables X and Y. It is usually denoted by r_{xy} or 'r'.

i) Direct Method: $$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{V(x)V(y)}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\text{After simplification} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

or

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}}\sqrt{\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

ii) Deviation method:

$$r = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\{\sum dx^2 - \frac{(\sum dx)^2}{n}\}}\sqrt{\{\sum dy^2 - \frac{(\sum dy)^2}{n}\}}}$$

Where σ_x = S.D. of x and σ_y = S.D. of Y

n = number of items; $d_x = x-A$, $d_y = y-B$

A = assumed value of and B = assumed value of y

The correlation coefficient never exceed unity. It always lies between –1 and +1 (i.e. $-1 \le r \le 1$)

If r = +1 then we say that there is a perfect positive correlation between x and y

If r = -1 then we say that there is a perfect negative correlation between x and y

If r = 0 then the two variables x and y are called uncorrelated variables

No unit of measurement.

Test for significance of correlation coefficient

If 'r' is the observed correlation coefficient in a sample of 'n' pairs of observations from a bivariate normal population, then Prof. Fisher proved that under the null hypothesis

$$H_0 : \rho = 0$$

the variables x, y follow a bivariate normal distribution. If the population correlation coefficient of x and y is denoted by ρ, then it is often of interest to test whether ρ is zero or different from zero, on the basis of observed correlation coefficient 'r'. Thus if 'r' is the sample correlation coefficient based on a sample of 'n' observations, then the appropriate test statistic for testing the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis H_1: $\rho \neq 0$ is

$$t = \frac{|r|}{\sqrt{1-r^2}}\sqrt{(n-2)}$$

t follows Student's t – distribution with (n-2) d.f.

If calculated value of $|t|$ > table value of t with (n-2) d.f. at specified level of significance, then the null hypothesis is rejected. That is, there may be significant correlation between the two variables. Otherwise, the null hypothesis is accepted.

Example: From a paddy field, 12 plants were selected at random. The length of panicles in cm (x) and the number of grains per panicle (y) of the selected plants were recorded. The results are given in the following table. Calculate correlation coefficient and its testing.

y	112	131	147	90	110	106	127	145	85	94	142	111
x	22.9	23.9	24.8	21.2	22.2	22.7	23.0	24.0	20.6	21.0	24.0	23.1

Solution:

a) Direct Method:

$$\text{Correlation coefficient } r_{xy} = r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}}\sqrt{\{\sum y^2 - \frac{(\sum y)^2}{n}\}}}$$

Where n = number of observations

Testing the correlation coefficient:

Null hypothesis H_0: Population correlation coefficient 'ρ' = 0

Under H_0, the test statistic becomes

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)} \text{ d.f.}$$

x	y	x^2	y^2	xy
112	22.9	12544	524.41	2564.8
131	23.9	17161	571.21	3130.9
147	24.8	21609	615.04	3645.6
90	21.2	8100	449.44	1908.0
110	22.2	12100	492.84	2442.0
106	22.7	11236	515.29	2406.2
127	23.0	16129	529.00	2921.0
145	24.0	21025	576.00	3480.0
85	20.6	7225	424.36	1751.0
94	21.0	8836	441.00	1974.0
142	24.0	20164	576.00	3408.0
111	23.1	12321	533.61	2564.1
$\sum x = 1400$	$\sum y = 273.4$	$\sum x^2 = 168450$	$\sum y^2 = 6248.20$	$\sum xy = 32195.6$

$$r = \frac{32195.6 - \frac{(1400)x273.4)}{12}}{\sqrt{\left(168450 - \frac{(1400)^2}{12}\right)}\sqrt{\left(6248.20 - \frac{(273.4)^2}{12}\right)}}$$

$$= \frac{298.9333}{71.5308x19.2367} = 0.95$$

The value of coefficient of determination (r^2) = 0.90

$$t = \frac{0.95\sqrt{12-2}}{\sqrt{1-(0.95)^2}}$$

$$= \frac{3.0042}{0.3122}$$

$$= 9.6$$

t critical (table) value for 10 d.f. at 5% LOS is 2.23

Since calculated value i.e. 9.6 is > t table value i.e. 2.23, it can be inferred that there exists significant positive correlation between (x, y).

b) Indirect Method:

Here A = 127 and B = 24

x	y	$d_x = x - A$	$d_y = y - B$	d_x^2	d_y^2	$d_x d_y$
112	22.9	-15	-1.1	225	1.21	16.5
131	23.9	4	-0.1	16	0.01	-0.4
147	24.8	20	0.8	400	0.64	16
90	21.2	-37	-2.8	1369	7.84	103.6
110	22.2	-17	-1.8	289	3.24	30.6
106	22.7	-21	-1.3	441	1.69	27.3
127	23	0	-1	0	1	0
145	24	18	0	324	0	0
85	20.6	-42	-3.4	1764	11.56	142.8
94	21	-33	-3	1089	9	99
142	24	15	0	225	0	0
111	23.1	-16	-0.9	256	0.81	14.4
		$\sum d_x =$ -124	$\sum d_y =$ -14.6	$\sum d_x^2 =$ 6398	$\sum d_d^2 =$ 37	$\sum d_x d_y =$ 449.8

$$r = \frac{\sum d_x d_y - \frac{\left(\sum d_x\right)\left(\sum d_y\right)}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$= \frac{449.8 - \frac{(-124)(-14.6)}{12}}{\sqrt{6398 - \frac{(-124)^2}{12}}\sqrt{37 - \frac{(14.6)^2}{12}}}$$

$$= \frac{298.9333}{\sqrt{5116.6667}x\sqrt{19.2367}}$$

$$= \frac{298.9333}{71.5309x4.3860} = 0.95$$

REGRESSION

The term `regression' literally means "stepping back towards the average". It was first used by a British Biometrician Sir Francis Galton.

The relationship between the independent and dependent variables may be expressed as a function. Such functional relationship between two variables is termed as regression.

In regression analysis independent variable is also known as regressor or predictor or explanatory variable while dependent variable is also known as regressed or explained variable.

When only two variables are involved the functional relationship is known as simple regression. If the relationship between two variables is a straight line, it is known as simple linear regression, otherwise it is called as simple non-linear regression.

Direct Method:

The regression equation of Y on X is given as

$$Y = a + bX$$

Where Y= dependent variable; X = independent variable and a = intercept

b = the regression coefficient (or slope) of the line. a and b are also called as Constants

the constants a and b can be estimated with by applying the `least squares method'. This involves minimizing $\sum_{i}^{n} e^2 = \sum_{i}^{i} (y - a - bx)^2$. This gives

$$b_{yx} = b = \frac{Cov(XY)}{v(X)} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

and $a = \overline{Y} - b\overline{X}$

where b is called the estimate of regression coefficient of Y on X and it measures the change in Y for a unit change in X.

Similarly, the regression equation of X on Y is given as

$$X = a^1 + b^1 Y$$

where X = dependent variable and Y = independent variable

$$b_{xy} = b^1 = = \frac{Cov(XY)}{v(Y)} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum Y^2 - \frac{(\sum Y)^2}{n}}$$

$$\text{and } a^1 = \overline{X} - b^1 \overline{Y}$$

where b^1 is known as the estimate of regression coefficient of X on Y and `a' is intercept

Deviation Method:

The regression equation of Y on X is

$$(Y - \overline{Y}) = b\,(X - \overline{X})$$

$$\Rightarrow Y = \overline{Y} + b(X - \overline{X})$$

The regression equation of X on Y

$$(X - \overline{X}) = b^1 (Y - \overline{Y})$$

$$\Rightarrow X = \overline{X} + b^1 (Y - \overline{Y})$$

Properties of regression coefficient:

1. Correlation coefficient is the geometric mean of the two regression coefficients.

   i.e $r = \pm\sqrt{b.b^1}$

2. If one of the regression coefficient is greater than unity, the other must be less than unity.
3. Arithmetic mean of the regression coefficients is greater than the correlation coefficient `r'.
4. Regression coefficients are independent of the change of origin but not of scale.
5. Units of 'b' are same as that of the dependent variable.
6. Regression is only a one-way relationship between y (dependent variable) and x (independent variable).
7. The range of b is from $-\infty$ to ∞. $-\infty$ for negative b and $+\infty$ for positive b.

Note:

1. Both the lines regression pass through the point $(\overline{X}, \overline{Y})$. In other words, the mean values $(\overline{X}, \overline{Y})$ can be obtained as the point of intersection of the two regression lines

2. If r = 0, the two variables are uncorrelated, the lines of regression become perpendicular to each other

3. If $r = \pm 1$, in this case the two lines of regression either coincide or they are parallel to each other

4.If the regression coefficients are positive, r is positive and if the regression coefficients are negative, r is negative

Distinguish between correlation and regression:

Correlation	Regression
1. Correlation is the relationship between two or more variables. Where the change in one variable affects a change in other variable	1. Regression is mathematical measure of the average relationship between two or more variables. Where one variable is dependent and other variable is independent
2. correlation coefficient measures extent of relationship between two variables	2. regression coefficient estimates the change in one variable for a unit change in other related variable
3. correlation is a two way relationship	3. regression is a one way relationship
4. correlation coefficient is independent of units of the variables.	4. regression coefficient is expressed in the units of dependent variable
5. correlation coefficient always lies between –1 and +1	5. regression coefficient lies between $-\infty$ and $+\infty$
6. In correlation both the variables are random	6. In regression one variable will be in dependent and other will be dependent
7. correlation coefficient calculated by $r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{V(x)V(y)}}$	7. the regression coefficient of y on x is $b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$ the regression coefficient of x on y is $b^1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}}$
8. $r = \sqrt{bb^1}$ we can not predict	$b = r\sqrt{\frac{\sum (y-\bar{y})^2}{\sum (x-\bar{x})^2}}$;we can predict

Example: The following data give the yield per plant (gm) (Y) and days to flowering (X) of 11 pigeonpea plants. Fit regression equations of Y on X and X on Y. Also estimate Y when the flowering period is 149.

Solution:

The regression equation of Y on X is given by

$$Y = a + bX$$

Where Y= dependent variable; X = independent variable and a = intercept

b = the regression coefficient of Y on X

The constants a and b can be estimated with the help of `least squares method'

$$\text{Where } b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}; \qquad a = \overline{Y} - b\overline{X}$$

Similarly, the regression equation of X on Y is given by

$$X = a^1 + b^1 Y$$

Where X = dependent variable and Y = independent variable

$$\text{Where } b^1 = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum Y^2 - \frac{(\sum Y)^2}{n}}; \text{ and } a^1 = \overline{X} - b^1 \overline{Y}$$

Y	X	Y^2	X^2	XY
24.72	91	611.08	8281	2249.52
20.25	76	410.06	5776	1539.00
38.56	98	1486.87	9604	3778.88
74.72	136	5583.08	18496	10161.92
72.75	128	5292.56	16384	9312.00
78.45	119	6154.40	14161	9335.55
69.8	142	4872.04	20164	9911.60
80.4	135	6464.16	18225	10854.00
160.2	147	25664.04	21609	23549.40
165.75	145	27473.06	21025	24033.75
77.56	122	6015.55	14884	9462.32
$\sum Y$ =863.16	$\sum X$ =1339	$\sum Y^2$ = 90026.91	$\sum X^2$ = 168609	$\sum XY$ = 114187.94

$$b = \frac{114187.94 - \frac{(1339)(863.16)}{11}}{168609 - \frac{(1339)^2}{11}}$$

$$= \frac{9117.83}{5616.18}$$

$$= 1.62$$

$$\overline{X} = \frac{\sum X}{n} = \frac{1339}{11} = 121.73 \text{ and}$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{863.16}{11} = 78.47$$

and $a = 78.47 - 1.62*121.73 = -118.74$

The regression equation of Y on X is $Y = a+bX = Y = -118.74 + 1.62X$

To estimate Y when X= 149

$$\hat{Y} = -118.74 + 1.62 * 149 = 122.64$$

$$b^1 = \frac{114187.94 - \frac{(1339)(863.16)}{11}}{90026.91 - \frac{(863.16)^2}{11}} = \frac{9117.83}{22295.53} = 0.41$$

$$a^1 = \overline{X} - b^1 \overline{Y} = 121.73 - 0.41*78.47 = 89.56$$

The regression equation of X on Y is $X = 89.56 + 0.41Y$

Result:

1.The regression equation of Y on X is $Y = a + bX = Y = -118.74 + 1.62X$

2. The regression equation of X on Y is $X = 89.56 + 0.41Y$

3. Estimated Yield is 122.64

Analysis of Variance (ANOVA)

The ANOVA is a powerful statistical tool for tests of significance. The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation when we have two or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis that all the samples have been drawn from the same population. For example, if three fertilizers are to be compared to find their efficacy, this could be done by a field experiment, in which each fertilizer is applied to 10 plots and then the 30 plots are later harvested with the crop yield being calculated for each plot. Now we have 3 groups of ten figures and we wish to know if there are any differences between these groups. The answer to this problem is provided by the technique of ANOVA.

The term ANOVA was introduced by Prof. R.A. Fisher in 1920's to deal with problem in the analysis of agronomical data. Variation is inherent in nature

The total variation in any set of numerical data is due to a number of causes which may be classified as i) Assignable causes and ii) chance causes

The variation due to assignable causes can be detected and measured where as the variation due to chance causes is beyond the control of humans and cannot be traced separately

ANOVA: The ANOVA is a simple arithmetical process of sorting out the components of variation in a given data

Types of ANOVA: There are two types i) One way classification and ii) Two way classification

Assumptions of ANOVA:

1. The observations are independent
2. Parent population from which observations are taken is normal
3. Various treatment and environmental effects are additive in nature
4. The experimental errors are distributed normally with mean zero and variance σ^2

Experimental Designs

In order to verify a hypothesis pertaining to some scientific phenomena we have to collect data. Such data are obtained by either observation or by experimentation. The main topics connected with data collection are Theory of Sample Surveys and Experimental Designs. In sample survey, a researcher makes observations on existing population and records data without interfering with the process that is being observed. In

experimentation, on the other hand, the researcher controls or manipulates the environment of the subjects that constitute the population. The experiments allow a researcher to study the factors of his interest and show that these factors actually cause certain effects. Hence, whenever the objective is to study the effects of variables rather than simply to describe a population, we prefer the data collection through experimentation.

Modern concepts of experimental design are due primarily to R.A. Fisher. He developed them in the planning of agricultural field experiments. They are now used in many fields of science.

Basic concepts:

Blocks: In agricultural experiments, most of the times we divide the whole experimental unit (field) into relatively homogeneous sub-groups or strata. These strata, which are more uniform amongst themselves than the field as a whole are known as blocks.

Treatments: the objects of comparison in an experiment are defined as treatments

For example: i) suppose an Agronomist wishes to know the effect of different spacings on the yield of a crop, different spacings will be treatments. Each spacing will be called a treatment.

ii) If different of fertilizer are tried in an experiment to test the responses of a crop to the fertilizer doses, the different doses will be treatments and each dose will be a treatment.

ii) A teacher practices different teaching methods on different groups in his class to see which yields the best results.

iii) A doctor treats a patient with a skin condition with different creams to see which is most effective.

Experimental unit: Experimental unit is the object to which treatment is applied to record the observations.

For example i) In laboratory insects may be kept in groups of five or six. To each group, different insecticides will be applied to know the efficacy of the insecticides. In this study different groups of insects will be the experimental unit.

ii) If treatments are different varieties, then the objects to which treatments are applied to make observations will be different plot of land. The plots will be called experimental units.

Basic principles of experimental designs:

The purpose of designing an experiment is to increase the precision of the experiment. In order to increase the precision, we try to reduce the experimental error. For reducing the experimental error, we adopt some techniques. These techniques form the basic principles of experimental designs. The basic principles of the experimental designs are replication, randomization and local control.

1. Replication: Repetition of treatment to different experimental units is known as Replication. In other words, the repetition of treatments under investigation is known as replication. We have no means of knowing about the variations in the results of a treatment. Only when we repeat the treatment several times we can estimate the experimental error.

A replication is used (i) to secure more accurate estimate of the experimental error, a term which represents the differences that would be observed if the same treatments were applied several times to the same experimental units;

(ii) to reduce the experimental error and thereby to increase precision, which is a measure of the variability of the experimental error.

The standard error of treatment mean is $\frac{\sigma}{\sqrt{r}}$. Where σ is S.D. of treatment in the population and `r' is the number of replications. As `r' increases, the standard error of mean decreases. Also in the analysis of variance the replication of treatments provides estimate of experimental error which is essential for the application of F-tes.

2. Randomization: when all the treatments have equal chances of being allocated to different experimental units it is known as randomization

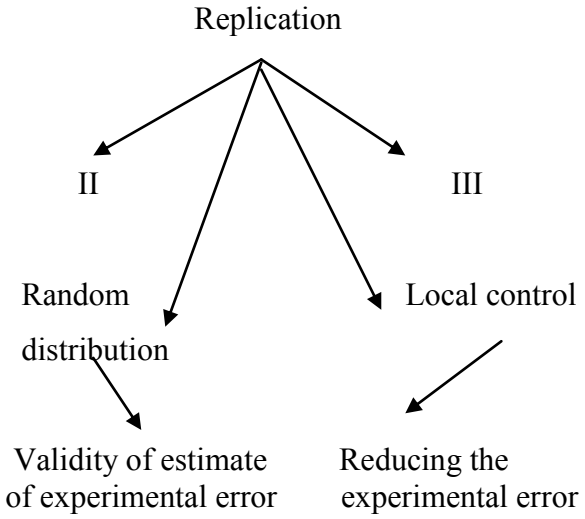
or

Random allocation of treatments to different experimental units known as randomization.

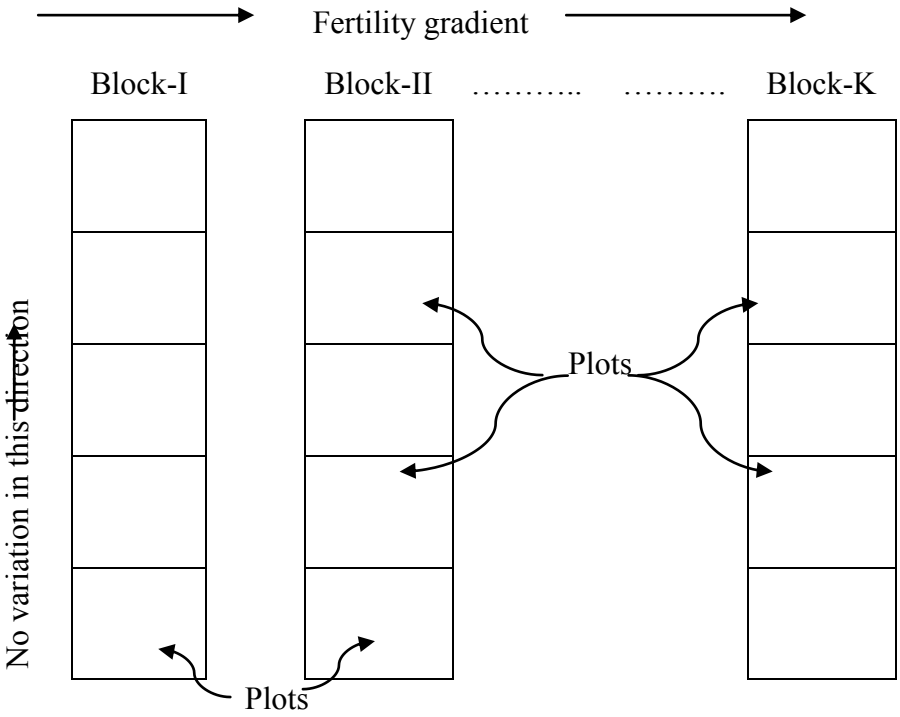
The purpose of randomization is to remove bias and other sources of extraneous variation which are not controllable. Another advantage of randomization (accompanied by replication) is that it forms the basis of any valid statistical test. Hence the treatments must be assigned at random to the experimental units. Randomization is usually done by using tables of random numbers.

3.Local control: It has been observed that all extraneous sources of variation are not removed by randomization and replication. This necessitates a refinement in the experimental technique. For this purpose, we make use of local control, a term referring to the grouping of homogeneous experimental units.

The main purpose of the principle of local control is to increase the efficiency of an experimental design by decreasing the experimental error.



Shape of blocks and plots: the shape and size of the blocks will usually depend up on the shape and size of the plots. In order to control the experimental error it is desirable to divide the whole experimental area into different subgroups (blocks) such that within each block there is as much homogeneity as possible but between blocks there is maximum variation. Further each block is to be divided into as many plots as the number of treatments. For maximum precision the plots should be rectangular in shape with their long sides parallel to the direction of the fertility gradient and the blocks should be arranged one after the other along the fertility gradient as shown in the figure.



COMPLETELY RANDOMIZED DESIGN (CRD)

The CRD is the simplest of all the designs. In this design, treatments are allocated at random to the experimental units over the entire experimental material. In case of field experiments, the whole field is divided into a required number of plots equal size and then the treatments are randomized in these plots. Thus the randomization gives every experimental unit an equal probability of receiving the treatment.

In field experiments there is generally large variation among experimental plots due to soil heterogeneity. Hence, CRD is not preferred in field experiments. In laboratory experiments and green house studies, it is easy to achieve homogeneity of experimental materials. Therefore, CRD is most useful in such experiments.

Layout of CRD: The placement of the treatments on the experimental units along with the arrangement of experimental units is known as the layout of an experiment.

For example, suppose that there are 5 treatments A,B,C,D and E. each with 4 replications, we need 20 experimental units. Here, since the number of units is 20, a two digit random number of table will be consulted and a series of 20 random numbers will be taken excluding those which are greater than 20. suppose, the random numbers are 4,18,2,14,3,7,13,1,6,10,17,20,8,15,11,5,9,12,16,19. After this the plots will be serially numbered and the treatment A will be allotted to the plots bearing the serial numbers 4, 18, 2, 14 and so on.

1 B	2 A	3 B	4 A	5 D
6 C	7 B	8 D	9 E	10 C
11 D	12 E	13 B	14 A	15 D
16 E	17 C	18 A	19 E	20 C

Statistical analysis:

Let us suppose that there are `k' treatments applied to `r' plots. These can be represented by the symbols as follows:

Treatments	1….2…….. j……..n	Totals	means
t_1	y_{11} y_{12}, … y_{1j}… y_{1r}	T_1	$\overline{T}_1$
t_2	y_{21} y_{22}, … y_{2j}… y_{2r}	T_2	$\overline{T}_2$
.	. . . .	.	.
.	. . . .	.	.
.	. . . .	.	.
ti	y_{i1} y_{i2}, … y_{ij}…….. y_{ir}	Ti	$\overline{T}$ i
.	. . . .	.	.
.	. . . .	.	.
t_k	y_{k1} y_{k2}, … y_{kj}… y_{kr}	T_k	$\overline{T}_k$
		G.T.	

Mathematical model:

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ (i = 1,2,….k; j = 1,2,……r)

Where y_{ij} is the j^{th} replication of the i^{th} treatment

μ = general mean effect

α_i = the effect due to i^{th} treatment = $\left[\overline{T_i} - \mu\right]$

ε_{ij} = error effect ($\varepsilon_{ij} \sim N(0,\sigma^2)$)

Null hypothesis H_0 : There is no significant difference between the treatment effects

$\Rightarrow \alpha_1 = \alpha_2 = ... = \alpha_k = 0$

Where $\alpha_i = \mu_i - \mu$, (i = 1, 2,……..,k)

The null hypothesis can be verified by applying the ANOVA procedure. The steps involved in this procedure are as follows:

1) Correction factor $= \frac{(G.T.)^2}{N}$

2) Treatment Sum of Square (Tr.S.S.) $= \frac{\left(T_1^2 + T_2^2 + ... + T_K^2\right)}{n} - CF$

$$= \frac{\sum_{i=1}^{k} T_i^2}{n} - CF$$

3) Total Sum of Square (TSS) $= \{ y_{11}^2 + y_{12}^2 + y_{13}^2 + \ldots\ldots\ldots + y_{kn}^2 \} - CF$

$= \Sigma\Sigma y_{ij}^2 - CF$

4) Error Sum of Square (ESS) = TSS-Tr.S.S.

ANOVA table

Sources of variation	*D.F.*	*S.S.*	*M.S.*	*F-cal.value*	*F-table value*
Treatments	*k-1*	*Tr.S.S.*	$TMS = \frac{Tr.S.S.}{k-1}$	$F_t = \frac{TMS}{EMS}$	*F[k-1, N-k] at* α *% LOS*
Error	*N-k*	*ESS*	$EMS = \frac{ESS}{N-k}$	-	
Total	*N-1*	*TSS*			

If the calculated value of F < table value of F at certain level of significance, we accept H_0 and hence we may conclude that there is no significant difference between the treatment means

If the calculated value of F > table vale of F, H_0 is rejected. Then the problem is to know which of the treatment means are significantly different. For this, we calculate critical difference (CD)

CD = SED x t – table value for error d.f. at 5% LOS.

where SED = Standard Error of Difference between the Treatments.

SED = $\sqrt{\frac{2EMS}{r}}$ (equal No. of replications), where r is number of replications

SED = $\sqrt{EMS\left(\frac{1}{r_i}+\frac{1}{r_j}\right)}$, where i = 1,2,…..,k and j = 1,2,…..,k (unequal No. replications)

The treatment means are arranged first in descending order of magnitude. If the difference between the two treatment means is less than CD value, it will he declared as non significant otherwise significant

Advantages and disadvantages of CRD:

1. This design is most commonly used in laboratory experiments such as in Ag. Chemistry, plant pathology, and animal experiments where the experimental material is expected to be homogeneous.
2. This design is useful in pot cultural experiments where the same type of soil is usually used. However, in greenhouse experiments care has to be taken with regard to sunshade, accessibility of air along and across the bench before conducting the experiment.

3. Any number of replications and treatments can be used. The number of replications may vary from treatment to treatment.
4. The analysis remains simple even if information on some units are missing
5. This design provides maximum number of degrees of freedom for the estimation of error than the other designs
6. The only draw back with this design is that when the experimental material is heterogeneous, the experimental error would be inflated and consequently the treatments are less precisely compared. The only way to keep the experimental error under control is to increase the number of replications thereby increasing the degrees of freedom for error.

Applications: 1. CRD is most useful in laboratory technique and methodological studies.

Ex: in physics, chemistry, in chemical and biological experiments, in some greenhouse studies etc.

2. CRD is also recommended in situations where an appreciable fraction of units is likely to be destroyed or fail to respond.

Case i) CRD with equal repetitions:

Example: In order to find out the yielding abilities of five varieties of sesamum an experiment was conducted in the greenhouse using a completely randomized design with four pots per variety. Analyze the data and state your conclusions

Seed yield of sesamum, g/pot

Varieties				
1	2	3	4	5
8	10	18	12	8
8	12	17	10	11
6	13	13	15	9
10	9	16	11	8

Sol: H_0: There is no significant difference between effect varities.

i.e. $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$

Correction factor (CF) = $\frac{(GT)^2}{N}$

Total sum of squares (TSS) = $\{y_{11}^2 + y_{12}^2 + y_{13}^2 + \ldots\ldots\ldots\ldots + y_{45}^2\} - CF$

Variety Sum of squares (VSS) =

$$\frac{(\sum v_1)^2 + (\sum v_2)^2 + (\sum v_3)^2 + (\sum v_4)^2 + (\sum v_5)^2}{r} - \text{CF}$$

Error Sum of Square (ESS) = TSS- VSS

	Varieties					
	1	2	3	4	5	
	8	10	18	12	8	
	8	12	17	10	11	
	6	13	13	15	9	
	10	9	16	11	8	
Variety totals	32	44	64	48	36	GT = 224
Means	8	11	16	12	9	

$$\text{CF} = \frac{(224)^2}{20} = 2508.8$$

$$\text{TSS} = \{(8)^2 + (10)^2 + (18)^2 + \ldots\ldots\ldots\ldots + (16)^2 + (11)^2 + (8)^2 - 2508.8 = 207.2$$

$$\text{VSS} = \frac{(32)^2 + (44)^2 + (64)^2 + (48)^2 + (36)^2}{4} - 2508.8 = 155.2$$

ESS= TSS-VSS = 207.2-155.2 = 52.0

ANOVA TABLE

Sources of variation	d.f.	S.S.	M.S.	F-cal value	F- table value
Varieties	5-1 =4	155.20	38.80	11.19	$F_{0.05}(4,15) = 3.06$
Error	19-4=15	52.0	3.47		
Total	20-1=19	207.20			

Calculated value of F > Table value of F at 5%LOS, H_0 is rejected and hence we conclude that there is significant difference between variety means.

Critical difference (CD): SED x $t_{0.05}$ for error d.f.

$$\text{Where SED} = \sqrt{\frac{2EMS}{r}}$$

t-table value for 15 d.f at 5%LOS is 2.13

$$\text{CD} = \sqrt{\frac{2*3.47}{4}} \text{ x } 2.13 = 2.80$$

The varieties means are arranging descending order of magnitude. If the difference between the verities mean is less than CD value, it will be declared as non significant otherwise significant

V_3	V_4	V_2	V5	V_1
16	12	11	9	8

The varieties which do not differ significantly have been underlined by a bar.

Coefficient of variation (CV) = Coefficient of variation = $\frac{\sqrt{EMS}}{\overline{X}}$ x100

Where $\overline{X}$ = Grand mean

Coefficient of variation = $\frac{\sqrt{3.47}}{11.2}x100 = = 16.6\%$

Conclusion: Lesser CV% indicates more consistency in the data

Case –ii) CRD with unequal repetitions:

Example: A Completely Randomized Design was conducted with the three treatments A B and C where treatment A is replicated 6 times and B is replicated 4 times and C is replicated 5 times. Analyze the data and state your conclusions.

A	B	C
16.5	15.0	18.2
17.0	13.8	24.3
16.0	14.0	25.0
12.0	17.9	18.9
18.0	-	21.0
14.0	-	-

Sol: Null hypothesis H_0: There is no significant difference between the effect treatments

i.e. $\alpha_1 = \alpha_2 = \alpha_3$

First treatment A is replicated r_1 = 6 times

Second treatment B is replicated r_2 = 4 times

Third treatment C is replicated r_3 = 5 times

Correction factor (CF) = $\frac{(GT)^2}{N}$

Total sum of squares (TSS) = $\{y_{11}^2 + y_{12}^2 + y_{13}^2 + \ldots\ldots\ldots\ldots + y_{35}^2\} - CF$

Treatment of Sum of square (Tr.S.S.) = $\frac{(\sum A)^2}{r_1} + \frac{(\sum B)^2}{r_2} + \frac{(\sum c)^2}{r_3}$ - CF

Error Sum of square(ESS) = TSS-Tr.S.S.

Standard error of difference = SED = $\sqrt{EMS\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$

Where i = 1, 2, ….., k and j = 1,2,….., r

Coefficient of variation = $\frac{\sqrt{2EMS}}{\overline{X}}$ X100; where $\overline{X}$ = Grand mean

Treatments							Totals	Means
A	16.5	17.0	16.0	12.0	18.0	14.0	ΣA=93.5	15.58
B	15.0	13.8	14.0	17.9	-	-	ΣB=60.7	15.18
C	18.2	24.3	25.0	18.9	21.0	-	ΣC=107.4	21.48
							GT=261.6	

$$CF = \frac{(261.6)^2}{15} = 4562.30$$

$$TSS = (16.5)^2 + (17.0)^2 + (16.0)^2 + (12.0)^2 + (18.0)^2 + (14.0)^2 + (15.0)^2 + (13.8)^2 + (14.0)^2 + (17.9)^2 + (18.2)^2 + (24.3)^2 + (25.0)^2 + (18.9)^2 + (21.0)^2 - 4562.30$$

$$= 4758.04 - 4562.30$$

$$= 195.74$$

$$Tr.S.S. = \frac{(93.5)^2}{6} + \frac{(60.7)^2}{4} + \frac{(107.4)^2}{5} - 4562.30$$

$$= 4685.11 - 4562.30$$

$$= 122.81$$

$$ESS = TSS\text{-}Tr.S.S.$$

$$= 195.74 - 122.81$$

$$= 72.93$$

AN0VA TABLE

Sources	d.f	S.S.	M.S.	F-cal. Value	F- table Value
Treatments	3-1=2	122.81	61.405	10.10	$F_{0.05}(2,12) = 3.89$
Error	12	72.93	6.0775		
Total	15-1=14	195.74			

Calculated value F(Tr) > Table value of F, H_0 is rejected and hence we conclude that there is significant difference between treatment means.

In CRD with unequal number of replications, we have to calculate CD for each pair of treatment means. As there are 3 treatments, we have to calculate $3C_2 = 3$ CD values.

Treatment pair	SED = $\sqrt{EMS\left(\frac{1}{r_i}+\frac{1}{r_j}\right)}$	CD = SED x $t_{0.05}$ d.f.
AB	$\sqrt{6.08\left(\frac{1}{6}+\frac{1}{4}\right)} = 1.59$	=1.59*2.18=3.47
AC	$\sqrt{6.08\left(\frac{1}{6}+\frac{1}{5}\right)} = 1.49$	=1.49*2.18=3.25
BC	$\sqrt{6.08\left(\frac{1}{4}+\frac{1}{5}\right)} = 1.65$	=1.65*2.18=3.61

Bar Notation:

The treatment means are arranged according to their ranks

T_3	T_1	T_2
21.48	15.58	15.18

i) Those pairs not scored are significant

ii) Those pairs under scored are non-significant

Coefficient of variation = $\frac{\sqrt{EMS}}{\bar{X}}$ x100

where $\bar{x}$ = Grand mean

$$=\frac{\sqrt{6.08}}{17.44}x100 \quad =14.14\%$$

Among three treatments, the third treatment i.e. T_3 is found to be superior one

RANDOMIZED BLOCK DESIGN (RBD)

We have seen that in a completely randomized design no local control measure was adopted excepting that the experimental units should be homogeneous. Usually, when experiments require a large number of experimental units, completely randomized designs cannot ensure precision of the estimates of treatment effects.

In agricultural field experiments, usually the experimental materials are not homogeneous. In such situations the principle of local control is adopted and the experimental material is grouped into homogeneous sub groups. The subgroup is

commonly termed as block. Since each block will consist the entire set of treatments a block is equivalent to a replication.

The blocks are formed with units having common characteristics which may influence the response under study. In agricultural field experiments the soil fertility is an important character that influences the crop responses. The uniformity trial is used to identify the soil fertility of a field. If the fertility gradient is found to run in one direction (say from north to south) then the blocks are formed in the opposite direction (from east to west).

If the number of experimental units within each group is same as the number of treatments and if every treatment appears precisely once in each group, then such an arrangement is called a randomized block design.

Layout: Let us consider 5 treatments A, B, C, D and E each replicated 4 times. We divide the whole experimental area into 4 relatively homogeneous blocks and each block into 5 plots. Treatments are then allocated at random to the plots of a block, fresh randomization being done for each block. A particular layout as follows.

Block-1	A	E	B	D	C

Block-2	E	D	C	B	A

Block-3	C	B	A	E	D

Block-4	A	D	E	C	B

Let us select one digit random numbers in the order of their occurrence in the table leaving 0 and greater than 5. suppose we get random numbers from 1 to 5 as: 1,3,5,4,2. So in the First block we allocate treatment A to the 1^{st} plot and B to 3^{rd} plot and so on.

Statistical Analysis: The results from RBD can be arranged in two way table according to the replications (blocks) and treatments; there will be `rk' observations in total. The data can be arranged in the following table.

Treatments	Blocks b_1…..b_2 ……b_j……. ..b_r	Treatment Totals	Means
t_1	y_{11} y_{12}, … y_{1j}……y_{1r}	T_1	$\overline{T}_1$
t_2	y_{21} y_{22}, … y_{2j}……y_{2r}	T_2	$\overline{T}_2$
.	. . . .	.	.
.	. . . .	.	.
.	. . . .	.	.
t_i	y_{i1} y_{i2}, … y_{ij}…….. y_{ir}	Ti	$\overline{T}$ i
.	. . . .	.	.
.	. . . .	.	.
t_k	y_{k1} y_{k2}, … y_{kj}……. y_{kr}	T_k	$\overline{T}_k$
Block totals	B_1 B_2 .....B_j …. …B_r	G.T.	
Means	$\overline{B}_1$ $\overline{B}_2$ ..... $\overline{B}_j$ …. . $\overline{B}_r$		

Mathematical model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (i = 1,2,....k;\ j = 1,2,.....r)$$

Where y_{ij} is the response of the j^{th} block and i^{th} treatment

μ = general mean effect

α_i = the effect due to i^{th} treatment

β_j = the effect due to j^{th} block

ε_{ij} is the error effect ($\varepsilon_{ij} \sim N(0,\sigma^2)$)

Null hypothesis: i) H_{01} : There is no significant difference between the treatment effects.

i.e. $\alpha_1 = \alpha_2 = ... = \alpha_k$

ii) H_{02}: There is no significant difference between the block effects

i.e. $\beta_1 = \beta_2 = ... = \beta_r$

The null hypothesis can be verified by applying the ANOVA procedure. The different steps are in the analysis of data are:

1) Correction factor $= \dfrac{(G.T)^2}{rk}$

2) Treatment Sum of Squares (Tr.S.S.) $= \dfrac{(T_1)^2 + (T_2{}^2) + ... + (T_K{}^2)}{r} - CF$

$$= \frac{\sum_{i=1}^{k} T_k{}^2}{r} - CF$$

3) Block Sum of Squares (BSS) $= \dfrac{(B_1^{\,2}) + (B_2^{\,2}) + \ldots\ldots\ldots + (B_r^{\,2})}{k} - CF$

$$= \frac{\sum_{j=1}^{r} B_j^{\,2}}{k} - CF$$

4) Total Sum of Squares (TSS) $= \{y_{11}^{\,2} + y_{12}^{\,2} + y_{13}^{\,2} + \ldots\ldots\ldots + y_{kr}^{\,2}\} - CF$

$$= \sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}^{\,2} - CF$$

5) Error Sum of Square (ESS) $= TSS - Tr.S.S. - SSB$

ANOVA TABLE

Sources of variation	D.F	S.S	M.S.	F-Cal.Value	F-table value At 5% LOS
Treatments	*k-1*	*Tr.S.S.*	$TMS = \dfrac{Tr.S.S.}{k-1}$	$F_t = \dfrac{TMS}{EMS}$	*F[k-1,{(r-1)(k-1)}]*
Blocks (Replications)	*r-1*	*BSS*	$BMS = \dfrac{BSS}{r-1}$	$F_b = \dfrac{BMS}{EMS}$	*F[r-1, {(r-1)(k-1)}]*
Error	*(r-1)(k-1)*	*ESS*	$EMS = \dfrac{ESS}{(r-1)(k-1)}$		
Total	rk-1	TSS			

If the calculated value of F (Treatments) < table value of F, we accept H_0, and hence we may conclude that there is no significant difference between the treatment means.

If calculated value of F (Treatments) > table value of F, we reject H_0 and hence we may conclude that there is significant difference between the treatment means.

If the treatments are significantly different, the comparison of the treatments is carried out on the basis of Critical Difference (C.D.).

C.D. = SED(Tr) x $t_{(r-1)(k-1)}$ at α level of significance

$$\text{Where SED} = \sqrt{\frac{2xEMS}{r}}$$

i) F (Blocks) should be not significant, if the planning of experiment is well manner.

ii) Desirable C.V. (%) in field experiment and lab experiment.

Advantages and disadvantages of RBD:

1. The principle advantage of RBD is that it increases the precision of the experiment. This is due to the reduction of experimental error by adoption of local control.
2. The amount of information obtained in RBD is more as compared to CRD. Hence, RBD is more efficient than CRD.
3. Flexibility is another advantage of RBD. Any number of replications can be included in RBD. If large number of homogeneous units are available, large number of treatments can be included in this design.
4. Since the layout of RBD involves equal replication of treatments, statistical analysis is simple. Even when some observations are missing of certain treatments, the data can be analysed by the use of missing plot technique.
5. When the number of treatments is increased, the block size will increase. If the block size is large it may be difficult to maintain homogeneity within blocks. Consequently, the experimental error will be increased. Hence, RBD may not be suitable for large number of treatments. But for this disadvantage, the RBD is a versatile design. It is the most frequently used design in agricultural experiments.
6. The optimum blocks size in field experiments is 21 plots. i.e. we can not compare treatments which are > 21 in RBD to preserve homogeneity of plots, within a block.

Example: The yields of 6 varieties of a crop in lbs., along with the plan of the experiment, are given below. The number of blocks is 5, plot of size is 1/20 acre and the varieties have been represented by A, B, C, D and E and analyze the data and state your conclusions

B-I	B 12	E 26	D 10	C 15	A 26	F 62
B-II	E 23	C 16	F 56	A 30	D 20	B 10
B-III	A 28	B 9	E 35	F 64	D 23	C 14
B-IV	F 75	D 20	E 30	C 14	B 7	A 23
B-V	D 17	F 70	A 20	C 12	B 9	E 28

Solution:

Null hypothesis H_{01}: There is no significant difference between variety means

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6$$

H_{02}: There is no significant difference between block means

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$$

Correction factor $= \dfrac{(G.T)^2}{rk}$

Variety Sum of Squares due to varieties (VSS) $= \dfrac{\sum v_i^2}{r} - CF$

Block Sum of square(BSS) $= \dfrac{\sum b_j^2}{k} - CF$

Total sum of squares (TSS) $= \Sigma\Sigma y_{ij}^2 - CF$

Error Sum of Square (ESS) = TSS- VSS- BSS

First rearrange the given data

Blocks	Varieties						Block totals	Means
	A	B	C	D	E	F		
B_1	26	12	15	10	26	62	$\Sigma B_1 = 151$	25.17
B_2	30	10	16	20	23	56	$\Sigma B_2 = 155$	25.83
B_3	28	9	14	23	35	64	$\Sigma B_3 = 173$	28.83
B_4	23	7	14	20	30	75	$\Sigma B_4 = 169$	28.17
B_5	20	9	12	17	28	70	$\Sigma B_5 = 156$	26.00
Variety totals	$\Sigma A = 127$	$\Sigma B = 47$	$\Sigma C = 71$	$\Sigma D = 90$	$\Sigma E = 142$	$\Sigma F = 327$	GT = 804	-
Means	25.4	9.4	14.2	18	28.4	65.4	-	-

$$CF = \frac{(804)^2}{30} = 21547.2$$

$$VSS = \frac{(127)^2 + (47)^2 + (71)^2 + (90)^2 + (142)^2 + (327)^2}{5} - 21547.2$$

$$= 31714.4 - 21547.2 = 10167.2$$

$$BSS = \frac{(151)^2 + (155)^2 + (173)^2 + (169)^2 + (156)^2}{6} - 21547.2$$

$$= 21608.67 - 21547.2 = 61.47$$

$$TSS = (12)^2 + (26)^2 + (10)^2 + (15)^2 + + \ldots\ldots\ldots + (12)^2 + (9)^2 + (28)^2 - 21547.2$$

= 32194 – 21547.2

= 10646.8

ESS = TSS – BSS – Tr.S.S.

= 10646.8 – 61.47 – 10167.2

= 418.13

ANOVA TABLE

Sources of variation	d.f	S.S.	M.S.	F-cal. Value	F- table Value
Blocks	5-1=4	61.47	15.37	0.74	$F_{0.05}$ (4, 20) =2.87
Varieties	6-1=5	10167.2	2033.44	97.25	$F_{0.05}$ (5, 20) = 2.71
Error	29-4-5= 20	418.13	20.91		
Total	30-1-29	10646.8			

Calculated value of F (Treatments) > Table value of F, H_0 is rejected and hence we conclude that there is highly significant difference between variety means.

Where SEm = $\sqrt{\frac{EMS}{r}} = \sqrt{\frac{20.91}{5}} = 2.04$

SED = $\sqrt{2}$ * SEm = 1.414 * 2.04 = 2.88

Critical difference = SED x t-table value for error d.f. at 5% LOS

∴ CD = 2.88 * 2.09

= 6.04

Coefficient of variation = $\frac{\sqrt{EMS}}{\overline{X}}$ X100 = $\frac{\sqrt{20.91}}{26.8}$ X100 = 17%

Bar Notation:

$\overline{F}$	$\overline{E}$	$\overline{A}$	$\overline{D}$	$\overline{C}$	$\overline{B}$
65.4	28.4	25.4	18.0	14.2	9.40

i) Those pairs not scored are significant

ii) Those pairs underscored are non-significant

Variety F gives significantly higher yield than all the other varieties; varieties D,C and B are on par and gives significantly higher yield than variety A.

LATIN SQUARE DESIGN (LSD)

When the experimental material is divided into rows and columns and the treatments are allocated such that each treatment occurs only once in a row and once in a column, the design is known as latin square design. In this design eliminating fertility variations consists in an experimental layout which will control variation in two perpendicular directions

[Latin square designs are normally used in experiments where it is required to remove the heterogeneity of experimental material in two directions. This design requires that the number of replications (rows) equal the number of treatments]. In LSD the number of rows and number of columns are equal. Hence the arrangement will form a square.

Layout of LSD: In this design the number of rows is equal to the number of columns and it is equal to the number of treatments. Thus in case of `m' treatments, there have to be mxm = $m^2$ experimental units (plots) arranged in a square so that each row as well as each column contain `m' plots. The `m' treatments are then allocated at random to these rows and columns in such a way that every treatment occurs once and only once in each row and each column such a layout is known as mxm L.S.D and is extensively used in agricultural experiments. The minimum and maximum number of treatments required for layout of LSD is 5 to 12.

In LSD the treatments are usually denoted by alphabets like A,B,C…etc. For a latin square with five treatments the arrangement may be as follows

Square –I

A	B	C	D	E
B	A	E	C	D
C	D	A	E	B
D	E	B	A	C
E	C	D	B	A

Square – II…….. etc.,

A	B	C	D	E
B	A	D	E	C
C	E	A	B	D
D	C	E	A	B
E	D	B	C	A

Statistical analysis: the mathematical model for LSD is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk} \qquad (i = j = k = 1,2,\ldots\ldots,m)$$

where Y_{ijk} denote the response from the unit (plot) in the i^{th} row, j^{th} column and receiving the k^{th} treatment

μ = general mean effect

α_i = i[th] row effect

β_j = j[th] column effect

γ_k = k[th] treatment effect;

$\in_{ijk}$ = error component

we know that total variation = variation due to rows + variation due to columns +

Variation due to treatments + variation due to error

Null hypothesis (H_0) = There is no significant difference between Rows, Columns and Treatment effects.

i.e. i) $H_{01}: \alpha_1 = \alpha_2 = ... = \alpha_m$

ii) $H_{02}: \beta_1 = \beta_2 = ... = \beta_m$ and

iii) $H_{03}: \gamma_1 = \gamma_2 = ... = \gamma_m$

The steps in the analysis of the data for verifying the null hypothesis are:

Different component variations can be calculated as follows:

1) $C.F = \frac{(G.T)^2}{m^2}$

2).Row Sum of Squares (RSS) $= \frac{\sum_{i=1}^{m} r_i^2}{m} - C.F$

3) Column Sum of Squares (CSS) $= \frac{\sum_{j=1}^{m} c_j^2}{m} - C.F$

4) Treatment Sum of Squares (Tr..S.S). $= \frac{\sum_{k=1}^{m} t_k^2}{m} - C.F$

5) Total Sum of Squares (TSS) $= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} y_{ijk}^2 - C.F$

6) Error Sum of Squares (ESS) = TSS –RSS-CSS-Tr.S.S.

ANOVA TABLE

Sources	D.F	S.S	M.S.	F-cal. value	F-table value at 5%LOS
Rows	*m-1*	*RSS*	$RMS = \frac{RSS}{m-1}$	$F_R = \frac{RMS}{EMS}$	$F_R[m-1, \{(m-1)(m-2)\}]$
Columns	*m-1*	*CSS*	$CMS = \frac{CSS}{m-1}$	$F_C = \frac{CMS}{EMS}$	〃
Treatments	*m-1*	*Tr.S.S.*	$TMS = \frac{Tr.S.S.}{m-1}$	$F_T = \frac{TMS}{EMS}$	〃
Error	*(m-1)(m-2)*	*ESS*	$EMS = \frac{ESS}{m-1}$		
Total	m^2*-1*	*TSS*	-		

If calculate value of F(Tr) < table value of Fat 5%LOS, H_0 is accepted and hence we may conclude that there is no significance difference between treatment effects.

If calculate value of F(Tr) > table value of F at 5%LOS, H_0 is rejected and hence we may conclude that there is significance difference between treatments effects.

If the treatments are significantly different, the comparison of the treatments is carried out on the basis of Critical Difference (C.D.).

C.D. = SED (Tr) x $t_{(r-1)(k-1)}$ at α level of significance

Where SED = $\sqrt{\frac{2xEMS}{m}}$, where m = number of rows

If F is significant, the significance of any treatment contrast can be tested by using the CD value.

Advantages of Latin Square Design: 1) With two way grouping or stratification LSD controls more of the variation than C.R.D. or R.B.D.

2) L.S.D. is an incomplete 3-way layout. Its advantage over complete 3-way layout is that instead of m^3 experimental units only m^2 units are needed. Thus a 4x4 L.S.D. results in saving of 64-16 = 48 observations over a complete 3-way layout.

3) The statistical analysis is simple though slightly complicated than for R.B.D. Even with missing data the analysis remains relatively simple.

4) More than one factor can be investigated simultaneously.

5) The missing observations can be analysed by using missing plot technique.

Example: An experiment on cotton was conducted to study the effect of foliar application of urea in combinations with insecticidal sprays in the cotton yield. Five treatments were tried in a 6x6 Latin Square Design. The layout plan and yield is given below:

T_2 4.9	T_4 6.4	T_5 3.3	T_1 9.5	T_3 11.8
T_3 9.3	T_1 4.0	T_2 6.2	T_5 5.1	T_4 5.4
T_4 7.0	T_3 15.4	T_1 6.5	T_2 6.0	T_5 4.6
T_5 5.3	T_2 7.6	T_3 13.2	T_4 8.6	T_1 4.9
T_1 9.3	T_5 6.3	T_4 11.8	T_3 15.9	T_2 7.6

Analyze the data and state your conclusions

Sol:

Null hypothesis: Rows, Columns and Treatments effects are equal

or

$H_{01}: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5;$

$H_{02}: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

$H_{03}: \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5$

$$CF = \frac{(GT)^2}{m^2}$$

$$RSS = \frac{\sum_{i=1}^{m} R_i^2}{m} - CF$$

$$CSS = \frac{\sum_{j=1}^{m} C_j^2}{m} - CF$$

$$Tr.S.S = \frac{\sum_{k=1}^{m} T_k^2}{m} - CF$$

$$TSS = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} y_{ij}^2 - CF$$

ESS = TSS-RSS-CSS-Tr.S.S.

ANOVA Table

Rows	Columns					Row totals
	1	2	3	4	5	
1	T_2 4.9	T_4 6.4	T_5 3.3	T_1 9.5	T_3 11.8	$R_1 = 35.9$
2	T_3 9.3	T_1 4.0	T_2 6.2	T_5 5.1	T_4 5.4	$R_2 = 30.0$
3	T_4 7.0	T_3 15.4	T_1 6.5	T_2 6.0	T_5 4.6	$R_3 = 9.5$
4	T_5 5.3	T_2 7.6	T_3 13.2	T_4 8.6	T_1 4.9	$R_4 = 39.6$
5	T_1 9.3	T_5 6.3	T_4 11.8	T_3 15.9	T_2 7.6	$R_5 = 50.9$
Column totals	$C_1 = 35.8$	$C_2 = 39.7$	$C_3 = 41.0$	$C_4 = 45.1$	$C_5 = 34.3$	GT=195.9

Treatment Totals: $\Sigma T_1 = 34.2$; $\Sigma T_2 = 32.3$; $\Sigma T_3 = 65.6$; $\Sigma T_4 = 39.2$; $\Sigma T_5 = 24.6$

$$CF = \frac{(195.9)^2}{25} = 1535.07$$

$$RSS = \frac{R_1^2 + R_2^2 + ... + R_5^2}{5} - CF$$

$$= \frac{(35.9)^2 + (30.0)^2 + (39.5)^2 + (39.6)^2 + (50.9)^2}{5} - 1535.07$$

$$= 46.54$$

$$CSS = \frac{C_1^2 + C_2^2 + ... + C_5^2}{5} - CF$$

$$= \frac{(35.8)^2 + (39.7)^2 + (41.0)^2 + (45.1)^2 + (34.3)}{5} - 1535.07$$

$$= 14.77$$

$$Tr.S.S = \frac{(\sum T_1)^2 + (\sum T_2)^2 + ... + (\sum T_5)^2}{5} - CF$$

$$= \frac{(34.2)^2 + (32.3)^2 + ((65.6)^2 + (39.2)^2 + (24.6)^2}{5} - 1535.07$$

$$= 196.55$$

$$TSS = Y_{11}^2 + Y_{12}^2 + Y_{13}^2 + \ldots\ldots\ldots\ldots + Y_{55}^2 - CF$$

$$= (4.9)^2 + (6.4)^2 + (3.3)^2 + \ldots\ldots..+ (15.9)^2 + (7.6)^2 - 1535.07$$

$$= 1821.07 - 1535.07 = 286.0$$

ESS = TSS – RSS – CSS – Tr.S.S

=286.0 - 46.54 - 14.77 - 196.55 = 28.14

ANVOA table

Sources	d.f.	S.S.	M.S.	F cal . value	F table value
Rows	5-1 = 4	46.54	11.64	4.95	$F_{0.05}(4, 12) = 3.26$
Columns	5-1 = 4	14.77	3.69	1.57	$F_{0.05}(4, 12) = 3.26$
Treatments	5-1 = 4	196.55	49.14	20.91	$F_{0.05}(4, 12) = 3.26$
Error	24 – 12 = 12	28.14	2.35		
Total	25-1 = 24	286.0			

Calculated value of F (treatments) > Table value of t at 5% LOS, H_0 is rejected and hence effect of foliar application of urea, have significant effect in the yield. To determine which of the treatment pairs differ significantly we have to calculate the critical difference (C.D.)

$$SEm = \sqrt{\frac{EMS}{m}} = \sqrt{\frac{2.35}{5}} = 0.69$$

$$SED = \sqrt{2} * SEm = 1.414 * 0.69 = 0.97$$

CD = SED x t – table value for 12 d.f. at 5 % LOS

= 0.98 * 2.18

= 2.11

$$\text{Coefficient of variation (CV)} = \frac{\sqrt{EMS}}{\overline{X}} * 100 = \frac{\sqrt{2.35}}{7.84} \text{ X100 } = 19.55\%$$

Bar Notation:

$\overline{T_3}$	$\overline{T_4}$	$\overline{T_1}$	$\overline{T_2}$	$\overline{T_5}$
13.12	7.84	6.84	6.46	4.92

i) The pairs not scored are significant

ii) The pairs under scored are non significant

From the bar chart it can be concluded that third treatment i.e. T_3 significantly higher than all the other treatments.

References Books:

1. Statistics for Agricultural Sciences, G. Nageswara Rao, Second Edition, BS Publications, Hyderabad
2. A Text book of Agricultural Statistics, R. Rangaswamy, New Age International (P) Limited, publishers
3. Statistical Methods, K.P. Dhamu and K. Ramamoorthy, AGROBIOS (INDIA)
4. Fundamentals of Mathematical Statistics, S.C. Gupta and V.K. Kapoor, Sultan Chand & Sons Educational Publications
5. Fundamentals Applied Statistics, S.C. Gupta and V.K. Kapoor, Sultan Chand & Sons Educational Publications
6. Design Resources Server: www.iasri.res.in