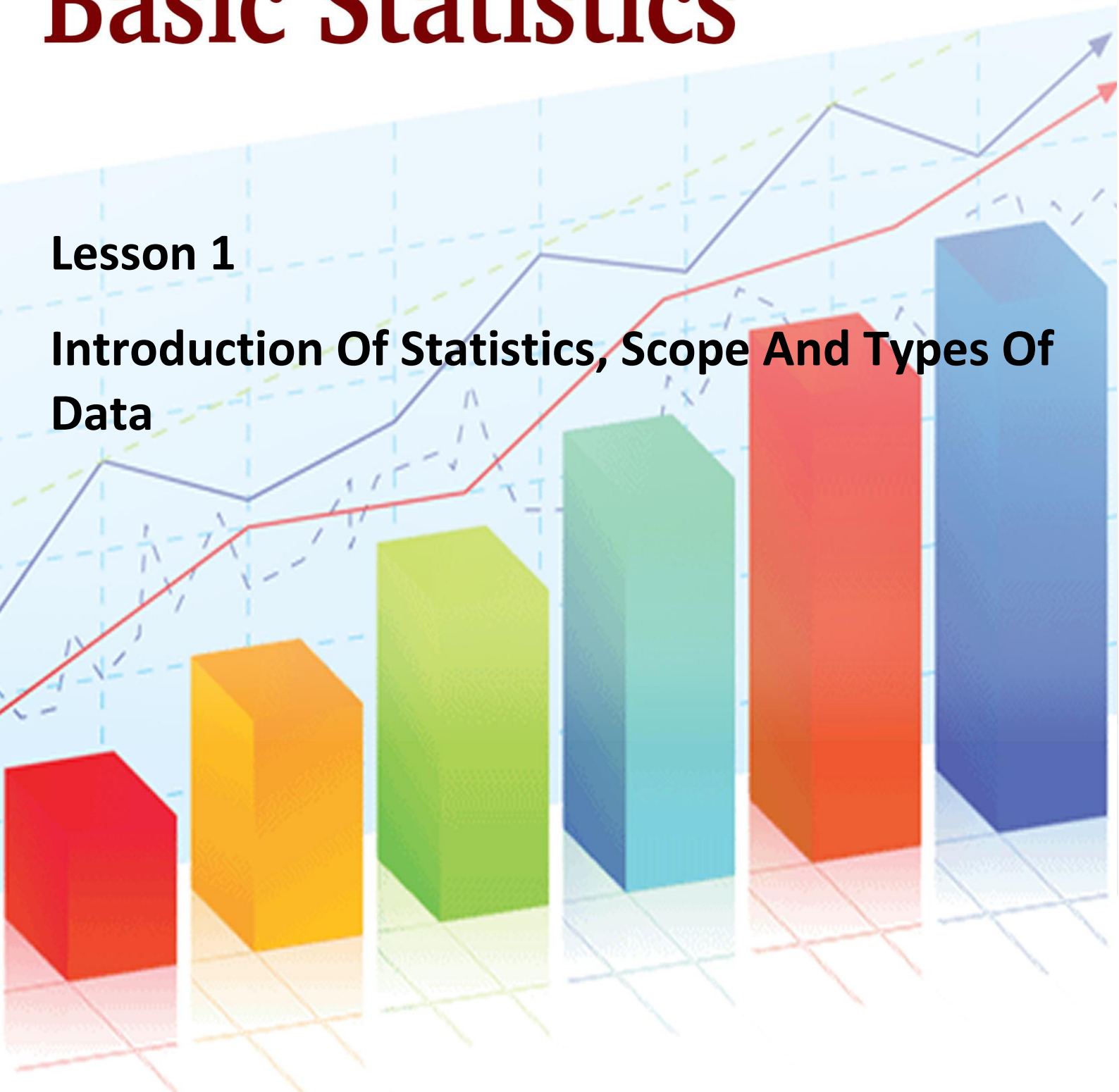
Basic Statistics



Lesson 1

Introduction Of Statistics, Scope And Types Of Data

Multiple Choice Questions

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 1	**Introduction Of Statistics, Scope And Types Of Data**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Objective of the Lesson:

1. Origin and growth of statistics
2. Importance and characteristics of statistics
3. Limitations and scope of statistics
4. Frequency distribution and its types

Glossary of Terms: Statistics, Scope, Frequency, Frequency Curve, etc

Introduction of Statistics, Scope and Types of Data

1.1 Introduction: Modern age is the age of science which requires that every aspect, whether it pertains to natural phenomena, politics, economics or any other field, should be expressed in an unambiguous and precise form. A phenomenon expressed in ambiguous and vague terms might be difficult to understand in proper perspective. Therefore, in order to provide an accurate and precise explanation of a phenomenon or a situation, figures are often used. The statement that prices in a country are increasing conveys only an incomplete information about the nature of the problem. However, if the figures of prices of various years are also provided, we are in a better position to understand the nature of the problem. In addition to this, these figures can also be used to compare the extent of price changes in a country vis-a-vis the changes in prices of some other country. Using these figures, it might be possible to estimate the possible level of prices at some future date so that some policy measures can be suggested to tackle the problem. The subject which deals with such type of figures, called data, is known as Statistics.

The word 'Statistics' is probably derived from the Latin word 'status' or the Italian word 'statista' or the German word

'statistik', each of which means a 'political state'. The word 'Statistics' is used in singular as well as in plural sense. As a plural, statistics may be defined as the numerical data relating to an aggregate of individuals and as a singular it is defined as the science of collection, organization, presentation, analysis and interpretation of numerical data.

1.2 Definition: Statistics has been defined differently by various authors from time to time. One can find more than hundred definitions in the literature but no definition can give a complete picture of the fast growing subject of Statistics. Important definitions are given below.

- It 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 by Lovitt
- The science which deals with the collection, analysis and interpretation of numerical data by Corxton and Cowden.
- The science of statistics is the method of judging collective, natural or social phenomenon form the results obtained from the analysis or enumeration or collection of estimates by Kings.
- 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) by Wallis and Roberts and this is the modern definition of statistics which covers the entire body of statistics.
- According to Sir R.A. Fisher "The science of Statistics is

essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data".

Fisher's definition is most exact in the sense that it covers all aspects and fields of Statistics viz. Collection, Organization, Presentation, Analysis and Interpretation of data.The credit for application of statistics to diverse fields of biological sciences goes to Sir R.A. Fisher (1890-1962) who is also known as "Father of Modern Statistics".

1.3 Scope of Statistics:

During last few decades statistics has penetrated into almost all sciences like agriculture, biology, business, social, engineering, medical, etc. Statistical methods are commonly used for analyzing and interpreting experimental data. Also, wide and varied applications have lead to the growth of many new branches of statistics such as Industrial Statistics, Biometrics, Biostatistics, Agricultural Statistics and the most recently developed Statistical Bioinformatics.

From definitions we may conclude that the Statistics is the science that transforms data into information and role of statisticians is to serve science and society through the development, understanding, and dissemination of state-of-the art techniques for collecting, presenting, analyzing, and drawing inferences from data. In brief we may summarize the scope of statistics as follows:

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

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

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

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

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

1.4. Limitations of Statistics:

i) Statistical methods are best applicable to quantitative data.

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

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

iv) Statistical conclusions are true on an average.

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

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

1.5 Concepts, Definitions, Frequency Distributions & Frequency Curves

1.5.1 Raw Data: The data collected by an investigator which have not been organized numerically and used by anybody else.

1.5.2 Array: An arrangement of raw numerical data in ascending or descending order of magnitude. The data can also be classified into Primary data and Secondary data.

1.5.3 Primary data: The data collected directly from the original source is called the primary data i.e. the data collected for the first time. The primary data may be collected by:

1.Direct interview method
2.Through mail
3.Through designed experiments

1.5.3.1 Direct interview method: In this method the investigator contacts the units/individuals and has personal interview. The information is recorded on the questionnaire or schedule. This information will be more reliable and correct but more expenditure may be involved and more time will be spent as the person himself will be going from place to place to collect the data.

1.5.3.2 Through mail: The data may be collected through correspondence. The questionnaire or schedules are sent by mail with the instructions for filling the same and return. It is less costly to get the data by mail. The main drawback of this method is the poor response. Usually the response by mail in surveys has been found to be about 40%.

1.5.3.3 Through designed experiments: Data are generated as outcome of the research conducted by the investigator himself.

1.5.4 Secondary Data: Sometimes we find that the data which we need had already been collected by some agencies for their study or the data are available in the published records. We may make use of such collected data, which is known as secondary data. The data, which have already been collected by some agency and have been processed or used at least once are called secondary data.

Secondary data may be collected from organizations or private agencies, government records, journals etc.

1.5.5 Variable: It is a common characteristics in biological science. A quantitative and qualitative characteristic that varies from observation to observations 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, yields of a crop, available nitrogen in soil, daily temperature, number of leaves per plant and number of eggs laid by insects are all variables. A quantity that varies from individual to individual is called a variable, e.g. height, weight etc. Variables are of two types i.e. discrete variable and continuous variable

1.5.6 Discrete Variable: A variable that takes only specific values in a given range, usually the integral values e.g. number of students in a college, number of petals in a flower, number of tillers in a plant etc.

1.5.7 Continuous Variable: A variable which can theoretically assume any value between two given values is called a continuous variable. A continuous variable can take any value within a certain range, for example yield of a crop, height of plants and birth rates etc.

1.6 Classification of data on the basis of Scales

Four levels or scales of Data measurement are:

i) **Nominal Scale:** Lowest level where only names are meaningful

ii) **Ordinal Scale:** Ordinal adds an order to the names.

iii) **Interval Scale:** Interval adds meaningful differences

iv) **Ratio Scale:** Ratio adds a zero so that ratios are meaningful.

Frequency: The number of times an individual item is repeated in a series is called its frequency. In case of grouped data, the number of observations lying in any class is known as the frequency of that class.

Frequency Distribution: It is tabular arrangement of data values along with their along with their frequencies.

Cumulative Frequency (less than type): The cumulative frequency corresponding to any value or class is the number of observations less than or equal to that value or upper limit of that class. It may also be defined as the total of all frequencies up to the value or the class. On similar lines we can define more than type cumulative frequencies.

Relative Frequency: The relative frequency of a class is the frequency of the class divided by the total frequency of all the classes and is generally expressed as a percentage.

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

Rules for Constructing a Frequency Distribution:

The following points should be borne in mind while tabulating or classifying an observed frequency distribution.

1. The classes should be well defined and non-overlapping.
2. As far as possible the class interval should be of equal width.
3. The classes should be exhaustive i.e. the range of the classes should cover the entire range of the data.
4. As a general rule, the number of classes should be between 10 and 15 and never more than 20 and not less than 5. However the exact number depends upon the data in hand.
5. Open-ended classes should be avoided.

Note: It is not necessary to choose the smallest value as the lower limit of the lowest class or the largest value as the upper limit of the highest class.

Struge's formula: A numerical formula as suggested by H.A. Struge may be used for determining approximately the class size and the number of classes. According to this formula the number of classes (k) is given

k = 1+ 3.322 $\log_{10}$ N, where N is the number of observations. Then class size is determined as

$$\text{Class width (h)} = \frac{\text{Largest value} - \text{Smallest test value}}{\text{Number of Classes}}$$

$$= \frac{\text{Range}}{\text{k}}$$

Ungrouped or discrete frequency distribution:

When the number of observations in the data is small, then the listing of the frequency of occurrence against the value of variable is called the discrete frequency distribution. For example raw data showing the number of children of 20

families:

2, 0, 3, 1, 1, 3, 4, 2, 0, 3, 4, 2, 2, 1, 0, 4,1, 2, 2, 3

The number of children can be considered as the variables X and the frequency of occurrence can be listed as below:

Number of children	0		2	3	4
Frequency	3		6	4	3

Grouped (Continuous) frequency distribution:

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

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 class and lower limit of just class. It is continuous distribution.

For example:

Class	Frequency
0-10	2
10-20	4
20-30	5

30-40	3
40-50	1

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

For example:

Class	Frequency
0-9	2
10-19	4
20-29	5
30-39	3
40-49	1

One can convert discontinuous distribution to continuous distribution by subtracting half of the gap (0.5 in this case) from lower limit and by adding the same quantity to the upper limit.

Example 1: Construct a frequency distribution table for he 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, 6, 10, 21, 36, 45, 22, 18

Solution:

Number of observation (N) =

30

Number of classes(k) = 1 + 3.322 log 30 = 5.9 = 6(approx)

$$\textbf{Class size (h)} = \frac{\textbf{Maximum Value} - \textbf{Minimum Value}}{\textbf{Number of Classes}}$$

$$= \frac{46 - 6}{6} \cong 7$$

Inclusive Method:

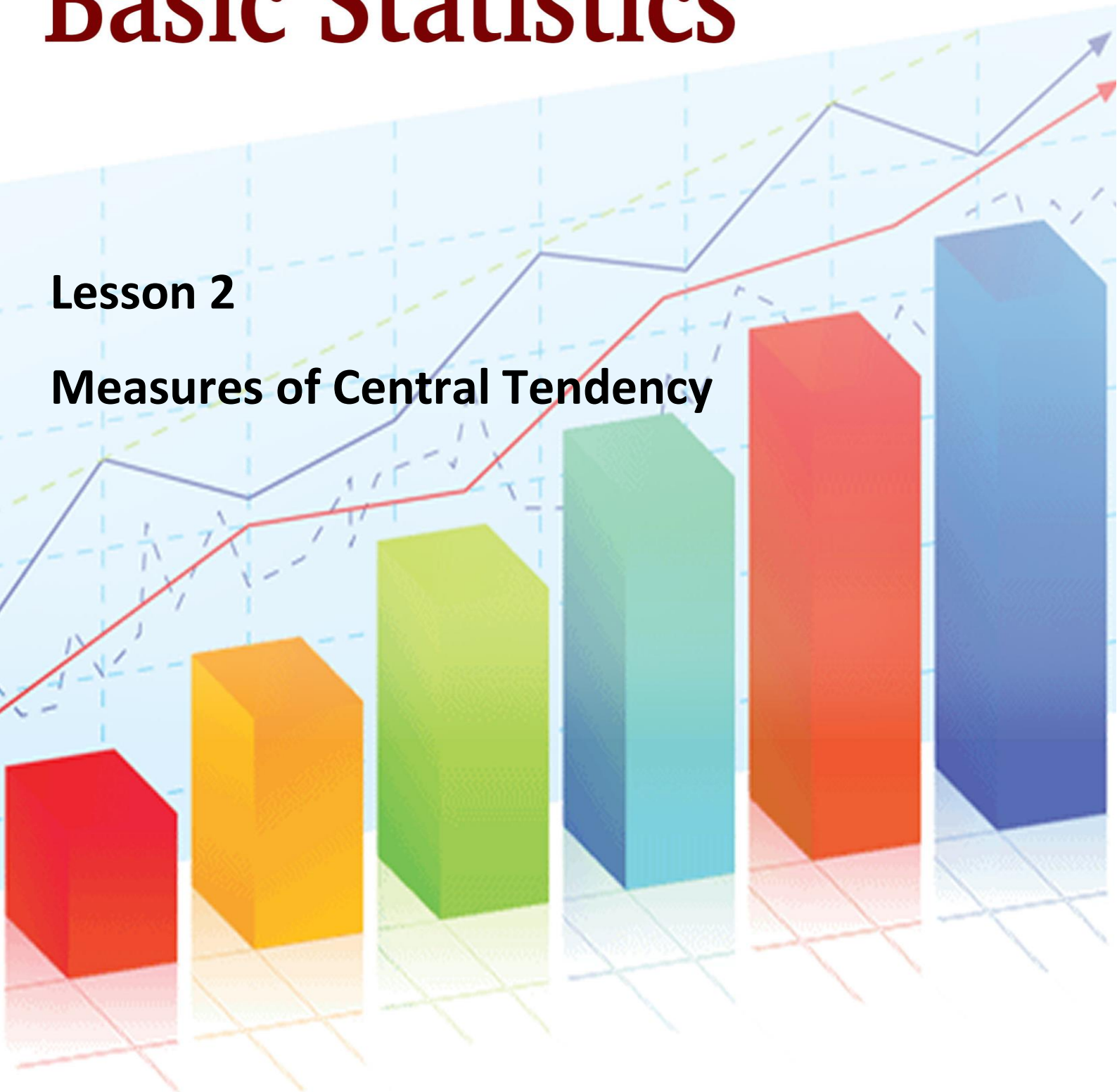
Class	Tally Marks	Frequency
6 - 11	IIII	4
12 - 17	I	1
18 - 23	IIII II	7
24 - 29	IIII	5
30 - 35	II	2
36 - 41	IIII	4
42 - 47	IIII II	7

Exclusive method:

Cla	Tally	Fre

ss	Marks	quency
5.5 – 11.5	IIII	4
11.5 – 17.5	I	1
17.5 – 23.5	IIIII II	7
23.5 – 29.5	IIIII	5
29.5 – 35.5	II	2
35.5 – 41.5	IIII	4
41.5 – 47.5	IIIII II	7

Basic Statistics



Lesson 2

Measures of Central Tendency

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications
ICAR-Indian Agricultural Statistics Research Institute

Course Name	**Basic Statistics**
Lesson 2	**Measures Of Central Tendency**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Objectives of the Lesson:

1. Characteristics for ideal averages
2. Various measures of Central tendency
3. Merit and Demerits of Various measures of central tendency
4. Deciles and Percentiles

Glossary of Terms: Mean, Median, Mode, Harmonic Mean, Geometric Mean, etc.

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations.

3.1 Types of Averages:

There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

3.2 Characteristics for a good or an ideal average:

The following properties should possess for an ideal average.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. Its definition shall be in the form of a mathematical formula.
5. It should be capable of further algebraic treatment.

6. It should have sampling stability.
7. It should be capable of being used in further statistical computations or processing.

3.3 Arithmetic mean

The arithmetic mean (or, simply average or mean) of a set of numbers is obtained by dividing the sum of numbers of the set by the number of numbers. If the variable x assumes n values $x_1, x_2 \ldots x_n$ then the mean, is given by

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

Example 1: Calculate the mean for 2, 4, 6, 8, and 10.

Solution:

$$Mean = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

Direct method : If the observations $x_1, x_2 \ldots x_n$ have frequencies $f_1, f_2, f_3, \ldots, f_n$ respectively, then the mean is given by :

$$Mean(\bar{X}) =) \frac{(f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum f_i x_i}{\sum f_i}$$

This method of finding the mean is called the direct method.

Example 2: Given the following frequency distribution, calculate the arithmetic mean

Marks (x)	50	55	60	65	70	75
No of Students (f)	2	5	4	4	5	5

Solution:

Marks (x)	50	55	60	65	70	75	Total

No of Students (f)	2	5	4	4	5	5	25
fx	100	275	240	260	350	375	1600

$$Mean(\bar{X}) =) \frac{(f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{1600}{25} = 64$$

(ii) Short cut method: In some problems, where the number of variables is large or the values of xi or fi are larger, then the calculations become tedious. To overcome this difficulty, we use short cut or deviation method in which an approximate mean, called assumed mean is taken. This assumed mean is taken preferably near the middle, say A, and the deviation di = xi − A is calculated for each variable x_i. Then the mean is given by the formula:

$$Mean(\bar{X}) = A + \frac{\sum f_i x_i}{\sum f_i}$$

Mean for a grouped frequency distribution

Example 3: Given the following frequency distribution, calculate the arithmetic mean

Marks (x)	50	55	60	65	70	75
No of Students (f)	2	5	4	4	5	5

Solution:

x	f	fx	d=x-A	fd
50	2	100	-10	-20

55	5	275	-5	-25
60	4	240	0	00
65	4	260	+5	20
70	5	350	+10	50
75	5	375	+15	75
	25	1600		100

By Direct method:

$$Mean(\bar{X}) =) \frac{(f_1 d_1 + f_2 d_2 + \ldots + f_n d_n)}{f_1 + f_2 + \cdots + f_n} = \frac{\sum f_i d_i}{\sum f_i}$$

$$= \frac{1600}{25} = 64$$

By Short-cut method:

$$Mean(\bar{X}) = A + \frac{\sum f_i d_i}{N}$$

$$= 60 + \frac{100}{25} = 60 + 4 = 64$$

Mean for a grouped frequency distribution

Find the class mark or mid-value x, of each class, as

$$x_i = class\ marks = \left(\frac{lower\ limit + upper\ limit}{2}\right)$$

Then

$$\bar{X} = \frac{\sum f_i x_i}{\sum f_i}\ or\ \bar{X} = A + \frac{\sum f_i d_i}{\sum f_i}, d_i = x_i - A$$

Example 4: Following is the distribution of persons according to different income groups. Calculate arithmetic mean.

Inco	0	1	2	3	4	5	6
me	-	0	0	0	0	0	0

SR (100)	10	-20	-30	-40	-50	-60	-70
Number of persons	6	8	10	12	7	4	3

Solution:

Income C.I	Number of Persons (f)	Mid X	$d_i = \frac{(x_i - A)}{h}$	fd
0-10	6	5	-3	-18
10-20	8	15	-2	-16
20-30	10	25	-1	-10
30-40	12	A =35	0	0
40-50	7	45	1	7
50-60	4	55	2	8
60-70	3	65	3	9
Total	50			-20

$$\bar{X} = A + \frac{\sum f_i d_i}{\sum f_i} \times h, d_i = \frac{(x_i - A)}{h}$$

$$A + \frac{-20}{50} \times 10 = 35 - 4 = 31$$

3.3.1 Merits and demerits of Arithmetic mean:

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits:

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

3.4 Harmonic mean (H.M.):

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If $x_1, x_2, .., x_n$ are n observations,

$$HM = \frac{n}{\sum_{i=1}^{n}(1/x_i)}$$

For a frequency distribution

$$HM = \frac{n}{\sum_{i=1}^{n} f\left(\frac{1}{x_i}\right)}$$

Example 5: From the given data calculate H. M. 5, 10, 17, 24, and 30.

Solution:

x	$\frac{1}{x}$
5	0.2000
10	0.1000
17	0.0588
24	0.0417
30	0.0333
Total	0.4338

Hence,

$$HM = \frac{n}{\sum_{i=1}^{n}(1/x_i)}$$

$$= \frac{5}{0.4338} = 11.52$$

Example 6: The marks secured by some students of a class are given below. Calculate the harmonic mean.

Marks	20	21	22	23	24	25
Number of Students	4	2	7	1	3	1

Solution:

Marks x	No of students	$\frac{1}{x_i}$	$\frac{f_i}{x_i}$
20	4	0.0500	0.2000
21	2	0.0476	0.0952

22	7	0.0454	0.3178
23	1	0.0435	0.0435
24	3	0.0417	0.1251
25	1	0.0400	0.0400
	18		0.8216

Hence,

$$HM = \frac{n}{\sum_{i=1}^{n} f(1/x_i)}$$

$$\frac{6}{0.8216} = 21.91$$

3.5 Geometric Mean (G.M.):

The geometric mean of a series containing n observations is the nth root of the product of the values. If $x_1, x_2, \ldots, x_n$ are observations then

G.M. = $\sqrt[n]{x_1 . x_2 \ldots . x_n}$

$(x_1 . x_2 \ldots . x_n)^{\frac{1}{n}}$

$$Log\ G.M. = \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log x_n)$$

$$Log\ G.M. = \frac{\sum \log x_i}{n}$$

$$G.M. = Antilog \frac{\sum \log x_i}{n}$$

Example 7: Calculate the geometric mean (G.M.) of the following series of monthly income of a batch of families 180, 250, 490, 1400, 1050.

Solution:

x	Log x
180	2.2553

250	2.3979
490	2.6902
1400	3.1461
1050	3.0212
	13.5107

$$G.M. = Antilog \frac{\sum \log x_i}{n} = Antilog \frac{13.5107}{5} = Antilog\ 2.70 = 503.6$$

Example 8: Calculate the average income per head from the data given below .Use geometric mean.

Class of people	Number of families	Monthly income per head (SR)
Landlords	2	5000
Cultivators	100	400
Landless – labours	50	200
Money – lenders	4	3750
Office Assistants	6	3000
Shop keepers	8	750
Carpenters	6	600
Weavers	10	300

Solution:

Class of people	**Annual income (SR) X**	**Number of families (f)**	**Log x**	**f logx**
Landlords	5000	2	3.6990	7.398

Cultivators	400	100	2.6021	260.210
Landless – labours	200	50	2.3010	115.050
Money – lenders	3750	4	3.5740	14.296
Office Assistants	3000	6	3.4771	20.863
Shop keepers	750	8	2.8751	23.2008
Carpenters	600	6	2.7782	16.669
Weavers	300	10	2.4771	24.771
		186		482.257

$$\text{G. M.} = \text{Antilog}\frac{\sum f_i \log x_i}{n}$$
$$= Antilog\frac{482.257}{186} = Antilog(2.5928)$$
$$= 391.50$$

Combined Mean:

If the arithmetic averages and the number of items in two or more related groups are known, the combined or the composite mean of the entire group can be obtained by

$$Combined\ Mean, \bar{X} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ \ldots + n_n\bar{x}_n}{n_1 + n_2 + \cdots + n_n}$$

Example 9: Find the combined mean for the data given below:

$$n_1 = 20; \bar{x}_1 = 4; n_2 = 30 \; and \; \bar{x}_2 = 3$$

Solution:

$$Combined\ Mean, \bar{X} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{4 \times 20 + 3 \times 30}{20 + 30} = \frac{80 + 90}{50}$$
$$= \frac{170}{50} = 3.4$$

3.6 Positional Averages (Median and Mode):

These averages are based on the position of the given observation in a series, arranged in an ascending or descending order. The magnitude or the size of the values does matter as was in the case of arithmetic mean. It is because of the basic difference that the median and mode are called the positional measures of an average.

3.6.1 Median:

The median is the middle value of a distribution i.e., median of a distribution is the value of the variable which divides it into two equal parts. It is the value of the variable such that the number of observations above it is equal to the number of observations below it.

Ungrouped or Raw data:

Arrange the given values in the increasing or decreasing order. If the numbers of values are odd, median is the middle value. If the numbers of values are even, median is the mean of middle two values.

By formula,

Median, Md = $\left(\frac{N+1}{2}\right)^{\text{th}}$ item

When odd numbers of values are given:-

Example 10: Find median for the following data

25, 18, 27, 10, 8, 30, 42, 20, 53

Solution:

Arranging the data in the increasing order 8, 10, 18, 20, 25, 27, 30, 42, 53

Here, numbers of observations are odd (N= 9)

Hence, Median, Md = $\left(\frac{N+1}{2}\right)^{th}$ item = $\left(\frac{9+1}{2}\right)^{\text{th}}$ item= $(5)^{\text{th}}$ item

The middle value is the 5th item i.e., 25 is the median value.

When even numbers of values are given:-

Example 11: Find median for the following data

5, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the data in the increasing order 2, 5, 8, 10, 12, 18, 22, 30

Here median is the mean of the middle two items (i.e) mean of (10, 12) i. e., $\left(\frac{10+12}{2}\right)$ = 11

Example 12: The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Accountancy.

Serial No	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Accountancy)	57	45	24	31	25	84	43	80	32	72

Solution: For such question, median is the most suitable measure of central tendency. The marks in the two subjects are first arranged in increasing order as follows:

Serial No	1	2	3	4	5	6	7	8	9	10
Marks in Statistics	28	30	32	35	46	47	52	53	55	60
Marks in Accountancy	24	25	31	32	43	45	57	72	80	84

Median value for Statistics $=$ (Mean of 5th and 6th items) $= = \left(\frac{46+47}{2}\right) = 46.5$

Median value for Accountancy $=$ (Mean of 5th and 6th items)

$$= \left(\frac{43+45}{2}\right) = 44$$

Therefore, the level of knowledge in Statistics is higher than that in Accountancy.

Grouped Data:

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

Discrete Series:

Step1: Find cumulative frequencies.

Step 2: Find $\left(\frac{N+1}{2}\right)$

Step 3: See in the cumulative frequencies the value just greater than $\left(\frac{N+1}{2}\right)$

Step4: Then the corresponding value of x will be median.

Example 13: The following data are pertaining to the number of members in a family. Find median size of the family.

Number of members x	1	2	3	4	5	6	7	8	9	10	11	12
Frequency F	1	3	5	6	10	13	9	5	3	2	2	1

Solution:

X	f	cf
1	1	1
2	3	4
3	5	9

4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
N=	60	

$$\text{Median} = \text{Size of } \left(\frac{N+1}{2}\right)\text{th item} = \text{Size of } \left(\frac{60+1}{2}\right)\text{th item}$$
$$= 30.5\text{th item.}$$

The cumulative frequency just greater than 30.5 is 38 and the value of x corresponding to 38 is 6. Hence the median size is 6 members per family.

Note:

It is an appropriate method because a fractional value given by mean does not indicate the average number of members in a family.

Continuous Series:

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step 2: Find $\left(\frac{N}{2}\right)$

Step3: See in the cumulative frequency the value first greater than $\left(\frac{N}{2}\right)$ Then the c

class interval is called the Median Class. Then apply the formula for Median

$$Md = l + \frac{\frac{N}{2} - cf}{F} \times h$$

Where,

$l = lower\ limit\ of\ the\ median\ class$

$\Sigma fi = n = number\ of\ Observations$

$f = frequency\ of\ the\ median\ class$

$h = size\ of\ the\ median\ class\ (assuming\ class\ size\ to\ be\ equal)$

$cf = cumulative\ frequency\ of\ the\ class\ preceding\ the\ median\ class.$

$N = Total\ frequency.$

Note:

If the class intervals are given in inclusive type convert them into exclusive type and call it as true class interval and consider lower limit in this.

3.7 Quartiles:

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower) quartile (Q1) marks off the first one-fourth, the third (upper) quartile (Q3) marks off the three-fourth.

Raw or ungrouped data:

First arrange the given data in the increasing order and use the formula for Q1 and Q3 then quartile deviation, Q.D. is given by

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

where, $Q_1 = \left(\frac{N+1}{4}\right)$ th item and $Q_3 = 3\left(\frac{N+1}{4}\right)$ th item

Example 14: Compute quartiles for the data given below 25,18, 30, 8, 15, 5, 10, 35, 40, 45

Solution:

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

$$Q_1 = \left(\frac{N+1}{4}\right) th\ item$$

$$= \left(\frac{10+1}{4}\right) th\ item$$

$$= (2.75)^{th}\ item.$$

$$= 2nd\ item + \left(\frac{3}{4}\right)(3rd\ item - 2nd\ item)$$

$$= 8 + \left(\frac{3}{4}\right)(10 - 8)$$

$$= 8 + \left(\frac{3}{4}\right) \times 2$$

$$= 9.5$$

$$Q_3 = 3\left(\frac{N+1}{4}\right)^{th}\ item$$

$$= 3(2.75)^{th}\ item.$$

$$= 8.25^{th}\ item$$

$$= 8th\ item + \left(\frac{3}{4}\right)(9th\ item - 8th\ item)$$

$$= 35 + \left(\frac{3}{4}\right)(40 - 35)$$

$$= 35 + 1.25$$

$= 36.25$

Discrete Series:

Step 1: Find cumulative frequencies

Step 2: Find $\left(\frac{N+1}{4}\right)$

Step 3: See in the cumulative frequencies, the value just greater than $\left(\frac{N+1}{4}\right)$ the corresponding value of x is Q_1

Step 4: Find $3\left(\frac{N+1}{4}\right)$

B See in the cumulative frequencies, the value just greater than $3\left(\frac{N+1}{4}\right)$ then the corresponding value of x is Q_3.

Example 15: Compute quartiles for the data given below.

X	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

Solution:

x	f	c.f
5	4	4
8	3	7
12	2	9
15	4	13

19	5	18
24	2	20
30	4	24
Total	24	

$$Q1 = \left(\frac{N+1}{4}\right)^{th} item$$

$$= \left(\frac{24+1}{4}\right)^{th} item$$

$$= \left(\frac{25}{4}\right)^{th} item$$

$$= 6.25^{th} = 8$$

$$Q_3 = 3\left(\frac{N+1}{4}\right)^{th} item$$

$$= (3 \times 6.25)^{th} item$$

$$= 18.75^{th} item = 24$$

Continuous Series:

Step 1: Find cumulative frequencies;

Step 2: Find $\left(\frac{N}{4}\right)$

Step 3: See in the cumulative frequencies, the value just greater $\left(\frac{N}{4}\right)$, then the corresponding class interval is called first quartile class.

Step 4: Find $3\left(\frac{N}{4}\right)$, See in the cumulative frequencies the value just greater than $3\left(\frac{N}{4}\right)$ then the corresponding class interval is called 3rd quartile class. Then apply the respective formulae

$$Q_1 = l_1 + \frac{\frac{N}{4} - cf_1}{f_1} \times h_1$$

$$Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - cf_3}{f_3} \times h_3$$

$where\ l_1 = lower\ limit\ of\ the\ first\ quartile\ class$

$f_1 = frequency\ of\ the\ first\ quartile\ class$

$h_1 = width\ of\ the\ first\ quartile\ class$

$cf_1 = c.f.preceding\ the\ first\ quartile\ class$

$l_3 = 1ower\ limit\ of\ the\ 3rd\ quartile\ class$

$f_3 = frequency\ of\ the\ 3rd\ quartile\ class$

$h_3 = width\ of\ the\ 3rd\ quartile\ class$

$cf_3 = c.f.preceding\ the\ 3rd\ quartile\ class$

3.8 Deciles:

These are the values, which divide the total number of observation into 10 equal parts. These are 9 deciles D1, D2…D9. These are all called first decile, second decile…etc.

Deciles for Raw data or ungrouped data

Example 16: Compute D5 for the data given below 5, 24, 36, 12, 20, 8.

Solution: Arranging the given values in the increasing order 5, 8, 12, 20, 24, 36

$$D5 = 5\left(\frac{N+1}{10}\right)^{th} observation$$

$$= 5\left(\frac{6+1}{10}\right)^{th} observation$$

$= (3.5)^{th}\ observation$

$= 3^{rd}\ item + \left(\frac{1}{2}\right)[4^{th}\ item - 3^{rd}\ item]$

$= 12 + \left(\frac{1}{2}\right)[20 - 12]$

$= 16.$

Deciles for Grouped data:

Same as quartile.

3.9 Percentiles:

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The percentile (Pk) is that value of the variable up to which lie exactly k% of the total number of observations.

Relationship:

$P_{25} = Q_1;\ P_{50} = D_5 = Q_2 = Median\ and\ P_{75} = Q_3$

Percentile for Raw Data or Ungrouped Data:

Example 17: Calculate P15 for the data given below: 5, 24, 36 , 12 , 20 , 8.

Solution: Arranging the given values in the increasing order. 5, 8, 12, 20, 24, 36

$P_{15} = 15\left(\frac{N+1}{100}\right)^{th}\ item$

$= 15\left(\frac{6+1}{100}\right)^{th}\ item$

$= (1.05)^{th}\ item$

$= 1^{st}\ item + 0.5(2^{nd}\ item - 1^{st}\ item)$

$= 5 + 0.5(8 - 5)$

$= 5.15$

Percentile for Grouped Data:

Example 18: Find P_{53} for the following frequency distribution.

Class Interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	5	8	12	16	20	10	4	3

Solution:

Class Interval	Frequency	cf
0-5	5	5
5-10	8	13
10-15	12	25
15-20	16	41
20-25	20	61
25-30	10	71
30-35	4	75
35-40	3	78
Total	78	

$$P_{53} = l + \frac{\frac{53N}{100} - cf}{f} \times h = 20 + \frac{41.34 - 41}{20} \times 5 = 20.085$$

3.10 Mode:

The mode or modal value of a distribution is that value of the variable for which the frequency is the maximum. It refers to that value in a distribution which occurs most frequently. It shows the center of concentration of the frequency in around a given value. Therefore, where the purpose is to

know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order for shoes or ready-made garments the modal size helps because these sizes and other sizes around in common demand.

Computation of the mode:

Ungrouped or Raw Data:

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example 19: 2, 7, 10, 15, 10, 17, 8, 10, 2

$$Mode = M_0 = 10$$

In some cases the mode may be absent while in some cases there may be more than one mode

Grouped Data:

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

Continuous distribution:

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the following formula:

$$Mode, Mo = l + \frac{f - f_1}{2f - f_1 - f_2} \times h$$

$Where, l = lower\ limit\ of\ the\ modal\ class$

$f = frequency\ of\ the\ modal\ class$

$f_1 = frequency\ of\ the\ class\ preceding\ the\ modal\ class$

$f_2 = frequency\ of\ the\ class\ following\ the\ modal\ class.$

$h = size\ of\ the\ modal\ class$

Remarks:

If $(2f_1 - f_0 - f_2)$ comes out to be zero, then mode is obtained by the following formula taking absolute differences within vertical lines;

$$Mode, M_0 = \frac{f - f_1}{|f - f_1| + |f - f_2|} \times \mathrm{h}$$

If mode lies in the first class interval, then f is taken as zero.

The computation of mode poses no problem in distributions with open-end classes, unless the modal value lies in the open-end class.

Example 20: Calculate mode for the following:

CI	f
0-50	5
50-100	14
100-150	40
150-200	91
200-250	150
250-300	87
300-350	60
350-400	38
400 and above	15

Solution:

The highest frequency is 150 and corresponding class interval is 200 – 250, which is the modal class.

$Here, l = 200;\ f = 150;\ f_1 = 91;\ f_2 = 87\ and\ h = 50$

$$Mode, Mo = l + \frac{f - f_1}{2f - f_1 - f_2} \times h$$

$$= 200 + \frac{150-91}{2\times150-91-87} \times 50$$

$$= 200 + 24.18$$

$$= 224.18$$

Determination of Modal class:

For a frequency distribution modal class corresponds to the maximum frequency. But in any one (or more) of the following cases-

If the maximum frequency is repeated

If the maximum frequency occurs in the beginning or at the end of the distribution

If there are irregularities in the distribution, the modal class is determined by the method of grouping.

Steps for Calculation:

We prepare a grouping table with 6 columns

1) In column I, we write down the given frequencies;
2) Column II is obtained by combining the frequencies two by two;
3) Leave the 1st frequency and combine the remaining frequencies two by two and write in column III;
4) Column IV is obtained by combining the frequencies three by three;
5) Leave the 1st frequency and combine the remaining frequencies three by three and write in column V;
6) Leave the 1st and 2nd frequencies and combine the remaining frequencies three by three and write in column VI.

Mark the highest frequency in each column. Then form an analysis table to find the modal class. After finding the modal class, use the formula to calculate the modal value.

3.11 Empirical Relationship between Averages

In a symmetrical distribution the three simple averages mean = median = mode. For a moderately asymmetrical distribution, the relationship between them are brought by Prof. Karl Pearson as

$$Mode = 3\,Median - 2\,Mean$$

Example 21: If the mean and median of a moderately asymmetrical series are 26.8 and 27.9 respectively, what would be its most probable mode?

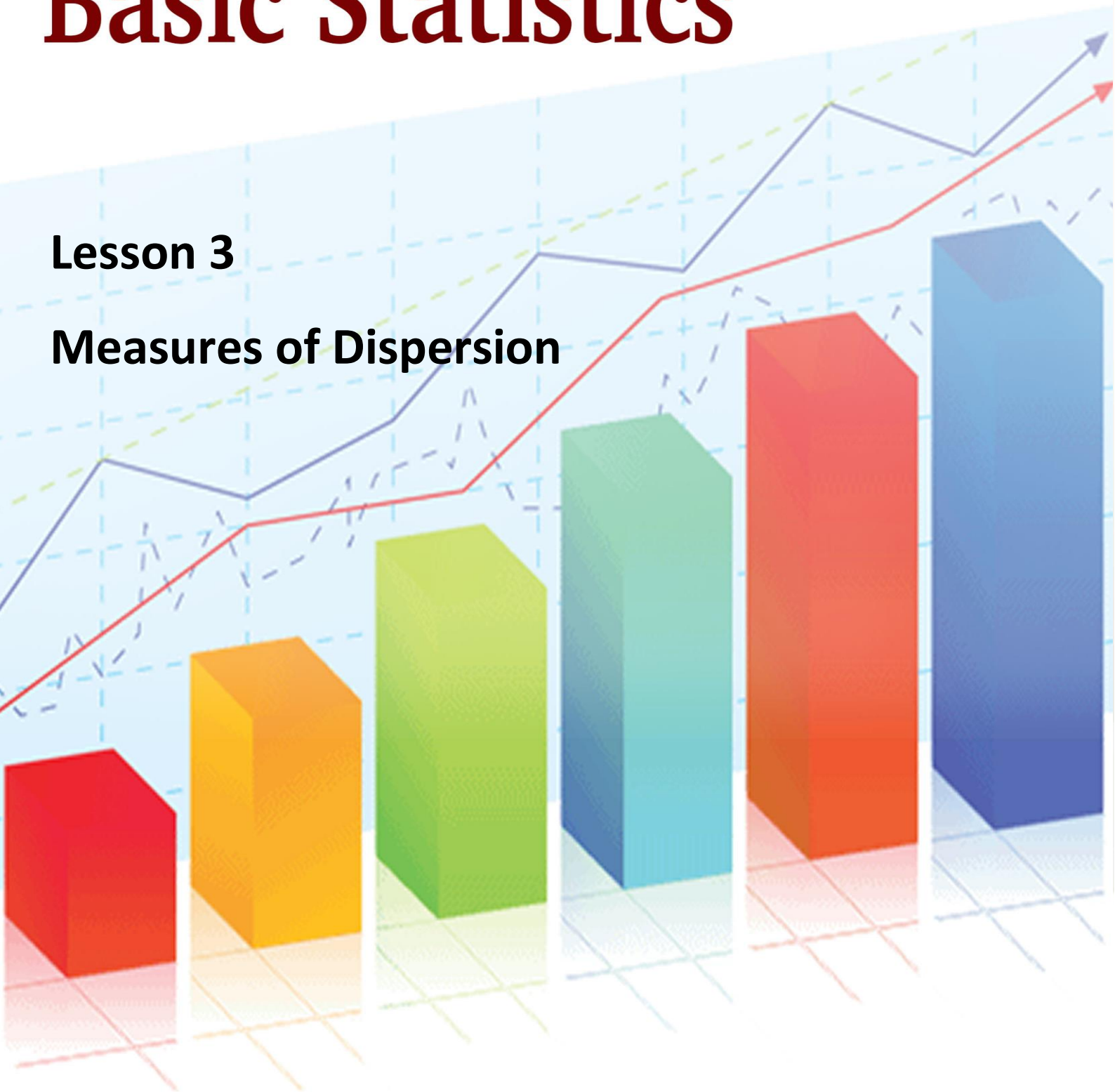
Solution:

Using the empirical formula

$$Mode = 3\,median - 2\,mean$$

$$= 3 \times 27.9 - 2 \times 26.8 = 30.1$$

Basic Statistics



Lesson 3

Measures of Dispersion

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 3	**Measures Of Dispersion**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Lesson-3

Objectives of the lesson:

1. Characteristics of good measure of Dispersion
2. Various absolute and relative measures of Dispersion
3. Mean deviation, Standard deviation and Coefficient of variation

Glossary of Terms: Dispersion, Range, Quartile Deviation, Mean Deviation, Standard Deviation, Coefficient of Variation etc.

4.1 Introduction:

The measures of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. For example consider the following marks of two students.

Student I	Student II
68	85
75	90
65	80
67	25
70	65

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

4.2 Characteristics of a good measure of dispersion:

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate

4.3 Absolute and Relative Measures:

There are two kinds of measures of dispersion, namely

1) Absolute measure of dispersion and
2) Relative measure of dispersion.

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations. They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations.

The various absolute and relative measures of dispersion are listed below.

Absolute measure	Relative measure
Range	Co-efficient of Range
Quartile deviation	Co-efficient of Quartile deviation
Mean deviation	Co-efficient of Mean deviation
Standard deviation	Co-efficient of variation

4.3.1 Range and coefficient of Range:

Range:

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, $Range = L - S$.

Where L = Largest value. S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed,

Method 1:

L = Upper boundary of the highest class

S = Lower boundary of the lowest class

Method 2:

L = Mid value of the highest class.

S = Mid value of the lowest class.

Co-efficient of Range:

$$Co - efficient\ of\ Range = \frac{L - S}{L + S}$$

Example 1: Find the value of range and its co-efficient for the following data.

7, 9, 6, 8, 11, 10, 4

Solution:

$$L = 11, S = 4.$$

$$Range = L - S$$

$$= 11 - 4 = 7$$

$$\text{Co} - \text{efficient of Range} = \frac{\text{L} - \text{S}}{\text{L} + \text{S}} = \frac{11 - 4}{11 + 4} = \frac{7}{15} = 0.4667$$

Example 2: Calculate range and its co efficient from the following distribution.

Size:	60- 63	63- 66	66- 69	69- 72	72- 75
Number:	5	18	42	27	8

Solution:

$$L = Upper\ boundary\ of\ the\ highest\ class = 75$$

$$S = Lower\ boundary\ of\ the\ lowest\ class = 60$$

$$Range = L - S = 75 - 60 = 15$$

$$Co - efficient\ of\ Range = \frac{L - S}{L + S} = \frac{75 - 60}{75 + 60} = \frac{15}{135} = 0.1111$$

4.3.2 Quartile Deviation and Co efficient of Quartile Deviation:

Quartile Deviation (Q.D.):

Definition: Quartile Deviation is half of the difference between the first and third quartiles. Hence, it is called Semi Inter Quartile Range.

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

Among the quartiles Q_1, Q_2 and Q_3, the range Q_3- Q_1 is called inter quartile range and $\frac{Q_3 - Q_1}{2}$, semi inter quartile range.

Co-efficient of Quartile Deviation:

$$Co - efficient\ of\ Quartile\ Deviation\ = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 3: Find the Quartile Deviation for the following data:

391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

Solution: Arrange the given values in ascending order.

384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488

$$Position\ of\ Q_1\ is\ \frac{N+1}{4} = \frac{10+1}{4} = 2.75^{th}\ item$$

$$Q_1 = 2^{nd}\ value + 0.75\ (3^{rd}\ value - 2^{nd}\ value)$$

$$= 391 + 0.75\ (407 - 391)$$

$$= 391 + 0.75 \times 16$$

$$= 391 + 12$$

$$= 403$$

$$Position\ of\ Q_3\ is\ 3\left(\frac{N+1}{4}\right) = 3 \times 2.75 = 8.25^{th}\ item$$

$$Q_3 = 8^{th}\ value + 0.25\ (9^{th}\ value - 8^{th}\ value)$$

$$= 777 + 0.25\ (1490 - 777)$$

$$= 777 + 0.25\ (713)$$

$$= 777 + 178.25 = 955.25$$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{955.25 - 403}{2} = \frac{552.25}{2} = 276.125$$

Example 4: Weekly wages of labors are given below. Calculate Q.D. and Coefficient of Q.D.

Weekly Wage (Rs.)	100	200	400	500	600
No. of Weeks	5	8	21	12	6

Solution:

Weekly Wage (Rs.)	No. of Weeks	Cum. No. of Weeks
100	5	5
200	8	13
400	21	34
500	12	46
600	6	52
Total	N=52	

$$Position\ of\ Q_1\ is\ \frac{N+1}{4} = \frac{52+1}{4} = 13.25^{th}\ item$$

$$Q1 = 13^{th}\ value + 0.25\ (14^{th}\ Value - 13^{th}\ value)$$
$$= 13^{th}\ value + 0.25\ (400 - 200)$$
$$= 200 + 0.25\ (400 - 200)$$
$$= 200 + 0.25\ (200)$$
$$= 200 + 50 = 250$$

$$Position\ of\ Q_3\ is\ 3\left(\frac{N+1}{4}\right) = 3 \times 13.25 = 39.25^{th}\ item$$

$$Q3 = 39^{th}\ value + 0.75\ (40^{th}\ value - 39^{th}\ value)$$

$$= 500 + 0.75\ (500 - 500)$$

$$= 500 + 0.75 \times 0$$

$$= 500$$

$$Q.D. = \frac{Q3 - Q1}{2} = \frac{500 - 250}{2} = \frac{250}{2} = 125$$

$$Co - efficient\ of\ Quartile\ Deviation = \frac{Q3 - Q1}{Q3 + Q1} = \frac{500 - 250}{500 + 250}$$

$$= \frac{250}{750} = 0.33$$

Example 5: For the date given below, give the quartile deviation and coefficient of quartile deviation.

X :	351 – 500	501 – 650	651 – 800	801– 950	951– 1100
f :	48	189	88	4	28

Solution:

x	F	True class Intervals	Cumulative frequency
351-500	48	350.5-500.5	48
501-650	189	500.5-650.5	237

651-800	88	650.5-800.5	325
801-950	47	800.5-950.5	372
951-1100	28	950.5-1100.5	400
Total	N = 400		

$Since, \frac{N}{4} = 100$

$Therefore, Q_1\ Class\ is\ 500.5 - 650.5$

$Hence, l_1 = 500.5; \frac{n}{4} = 100; cf1 = 48; f_1 = 189; h_1 = 150$

$$Q_1 = l_1 + \frac{\frac{N}{4} - cf_1}{f_1} \times h_1$$

$$= 500.5 + \frac{100 - 48}{189} \times 150 = 541.77$$

$Now, for\ Q3$

$3\left(\frac{N}{4}\right) = 3 \times 100 = 300$

$Hence, Q_3\ Class\ is\ 650.5 - 800.5$

$l_3 = 650.5; 3\left(\frac{N}{4}\right) = 300; cf_3 = 237; f_3 = 88; h_3 = 150$

$$Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - cf_3}{f_3} \times h_3$$

$$= 650.5 + \frac{300 - 237}{88} \times 150 = 757.89$$

$$Q.D. = \frac{Q3 - Q1}{2} = \frac{757.89 - 541.77}{2} = \frac{216.12}{2} = 108.06$$

$$\text{Co} - \text{efficient of Quartile Deviation} = \frac{\text{Q3} - \text{Q1}}{\text{Q3} + \text{Q1}}$$

$$= \frac{757.89 - 541.77}{757.89 + 541.77} = \frac{216.12}{1299.66}$$

$$= 0.1663$$

4.3.3 Mean Deviation and Coefficient of Mean Deviation:

Mean Deviation: The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation is measure of dispersion based on all items in a distribution.

Definition:

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored.

We usually compute mean deviation about any one of the three averages mean, median or mode. Sometimes mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

Coefficient of mean deviation:

Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

$$\text{Coefficient of Mean Deviation} = \frac{\text{Mean Deviation (M. D.)}}{\text{Mean or Median or Mode}}$$

If the result is desired in percentage,

$$\text{The coefficient of mean deviation} = \frac{\text{Mean Deviation (M.D.)}}{\text{Mean or Median or Mode}} \times 100$$

Computation of mean deviation – Individual Series:

1) Calculate the average mean, median or mode of the series.
2) Take the deviations of items from average ignoring signs and denote these deviations by |D|.
3) Compute the total of these deviations, i.e., Σ |D|
4) Divide this total obtained by the number of items.

Symbolically,

$$M.D. = \frac{\sum |D|}{n}$$

Example 6: Calculate mean deviation from mean and median for the following data:

100, 150, 200, 250, 360, 490, 500, 600, and 671.

Also calculate co- efficient of M.D.

Solution:

$$Mean = \frac{\sum X}{n}$$

$$= \frac{100 + 150 + 200 + 250 + 360 + 490 + 500 + 600 + 671}{9}$$

$$= \frac{3321}{9} = 369$$

Now arrange the data in ascending order

100, 150, 200, 250, 360, 490, 500, 600, 671

$$\text{Median, Md} = \text{Value of } \left(\frac{N+1}{2}\right)^{th} \text{ item} = \left(\frac{9+1}{2}\right)^{th} \text{ item} = 5^{th} \text{ item}$$
$$= 360$$

X	$\boldsymbol{D} = \lvert \boldsymbol{x} - \boldsymbol{Mean} \rvert$	$\lvert \boldsymbol{D} \rvert = \lvert \boldsymbol{x} - \boldsymbol{M_d} \rvert$
100	269	260
150	219	210
200	169	160
250	119	110
360	9	0
490	121	130
500	131	140
600	231	240
671	302	311
3321	1570	1561

$$M.D. from\ mean = \frac{\sum |D|}{n} = \frac{1570}{9} = 174.44$$

$$Co-efficient\ of\ M.D. = \frac{Mean\ Deviation\ (M.D.)}{Mean} = \frac{174.44}{369} = 0.47$$

$$M.D. from\ median = \frac{\sum |D|}{n} = \frac{1561}{9} = 173.44$$

$$Co-efficient\ of\ M.D. = \frac{Mean\ Deviation\ (M.D.)}{Median} = \frac{173.44}{360} = 0.48$$

Mean Deviation- Discrete Series:

$$M.D. = \frac{\sum f\ |D|}{n}$$

Example:7

Compute Mean deviation from mean and median from the following data:

Height in cms	158	159	160	161	162	163	164	165	166
No. of persons	15	20	32	35	33	22	20	10	8

Also compute coefficient of mean deviation.

Solution:

Height X	**No. of persons f**	$d = x - A$, $A = 162$	fd	$\|D\| = \|X - mean\|$	$f\|D\|$
158	15	- 4	-60	3.51	52.65
159	20	- 3	-60	2.51	50.20
160	32	- 2	-64	1.51	48.32
161	35	- 1	-35	0.51	17.85
162	33	0	0	0.49	16.17
163	22	1	22	1.49	32.78
164	20	2	40	2.49	49.80
165	10	3	30	3.49	34.90
166	8	4	32	4.49	35.92
	195		-95		338.59

$$Mean = A + \frac{\sum f\,d}{N} = 162 + \frac{-95}{195} = 162 - 0.49 = 161.51$$

$$M.D. = \frac{\sum f\,|D|}{n} = \frac{338.59}{195} = 1.74$$

$$Coefficient\ of\ M.D. = \frac{MD}{Mean} = \frac{1.74}{161.51} = 0.0108$$

Height x	No. of persons (f)	c.f.	$\|X - Median\|$	$f\|D\|$
158	15	15	3	45
159	20	35	2	40
160	32	67	1	32
161	35	102	0	0
162	33	135	1	33
163	22	157	2	44
164	20	177	3	60
165	10	187	4	40
166	8	195	5	40
	195			334

$$Median = Size\ of \left(\frac{N+1}{2}\right)^{th} item = \left(\frac{195+1}{2}\right)^{th} item = (96)^{th}\ item = 161$$

$$M.D. = \frac{\sum f|D|}{n} = \frac{334}{195} = 1.71$$

$$Coefficient\ of\ M.D. = \frac{MD}{Median} = \frac{1.71}{161} = 0.0106$$

Mean Deviation-Continuous Series:

The method of calculating mean deviation in a continuous series same as the discrete series. In continuous series we have to find out the mid points of the various classes and take deviation of these points from the average selected. Thus

$$M.D. = \frac{\sum f|D|}{n}$$

Example 8: Find out the mean deviation from mean and median from the following series.

Age in years	No of persons
0-10	20
10-20	25

20-30	32
30-40	40
40-50	42
50-60	35
60-70	10
70-80	8

Also compute co-efficient of mean deviation.

Solution:

X	M	F	$d = \frac{m - A}{c}$, $A = 35, c = 10$	fd	$D = \lvert m - \bar{x} \rvert$	$f\lvert D \rvert$
0-10	5	20	-3	-60	31.5	630.0
10-20	15	25	-2	-50	21.5	537.5
20-30	25	32	-1	-32	11.5	368.0
30-40	**35**	40	0	0	1.5	60.0
40-50	45	42	1	42	8.5	357.0
50-60	55	35	2	70	18.5	647.5
60-70	65	10	3	30	28.5	285.0
70-	75	8	4	32	38.5	308.0

80						
		212		32		3193.0

$$Mean = A + \frac{\sum f\,d}{N} \times C$$

$$= 35 + \frac{32}{212} \times 10 = 36.5$$

$$M.D. = \frac{\sum f\,|D|}{n} = \frac{3193}{212} = 15.06$$

Calculation of median and M.D. from median:

X	M	f	Cf	$\|D\| = \|m - M_d\|$	f|D|
0-10	5	20	20	32.25	645.00
10-20	15	25	45	22.25	556.25
20-30	25	32	77	12.25	392.00
30-40	35	40	117	2.25	90.00
40-50	45	42	159	7.75	325.50
50-60	55	35	194	17.75	621.25
60-70	65	10	204	27.75	277.50
70-80	75	8	212	37.75	302.00
Total		N=212			3209.50

$$\frac{N}{2} = \frac{212}{2} = 106$$

$$l = 30;\ \frac{N}{2} = 106;\ cf = 77;\ f = 40;\ h = 10$$

$$Median, Md = l + \frac{\frac{N}{2} - cf}{f} \times h = 30 + \frac{106 - 77}{40} \times 10 = 37.25$$

$$M.D. = \frac{\sum f\,|D|}{n} = \frac{3209.50}{212} = 15.14$$

$$Coefficient\ of\ M.D. = \frac{MD}{Median} = \frac{15.14}{37.25} = 0.41$$

4.3.3.1 Merits and Demerits of M.D:

Merits:

1. It is simple to understand and easy to compute.
2. It is rigidly defined.
3. It is based on all items of the series.
4. It is not much affected by the fluctuations of sampling.
5. It is less affected by the extreme items.
6. It is flexible, because it can be calculated from any average.
7. It is better measure of comparison.

4.3.3.2 Demerits:

1. It is not a very accurate measure of dispersion.
2. It is not suitable for further mathematical calculation.
3. It is rarely used. It is not as popular as standard deviation.
4. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

4.3.4 Standard Deviation and Coefficient of variation:

Standard Deviation:

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by the

Greek letter σ (sigma).

Calculation of Standard deviation-Individual Series:

There are two methods of calculating Standard deviation in an individual series.

a) Deviations taken from Actual mean; and

b) Deviation taken from Assumed mean

(a) Deviation taken from Actual mean:

This method is adopted when the mean is a whole number.

Steps:

1. Find out the actual mean of the series ($\bar{x}$)
2. Find out the deviation of each value from the mean $X = (x - \bar{x})$
3. Square the deviations and take the total of squared deviations x2
4. Divide the total (ΣX^2) by the number of observation, $\frac{\Sigma X^2}{N}$
5. The square root of , $\frac{\Sigma X^2}{N}$ is standard deviation.

Thus, Standard Deviation (SD or σ^2) $= \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$

(b) Deviations taken from assumed mean:

This method is adopted when the arithmetic mean is fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, we apply short– cut method; deviations are taken from an assumed mean. The formula is:

$$SD\ or\ \sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

Where d-stands for the deviation from assumed mean $= (x - A)$

Steps:

Assume any one of the item in the series as an average (A)

1. Find out the deviations from the assumed mean; i.e., X-A denoted by d and also the total of the deviations Σd

2. Square the deviations; i.e., d^2 and add up the squares of deviations, i.e, $\sum d^2$
3. Then substitute the values in the following formula:

$$SD\ or\ \sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$$

Example 9: Calculate the standard deviation from the following data. 14, 22, 9, 15, 20, 17, 12, 11

Solution:

Deviations from actual mean.

Value (x)	$d = x - \bar{x}$	$(x - \bar{x})^2$
14	-1	1
22	7	49
9	-6	36
15	0	0
20	5	25
17	2	4
12	-3	9
11	-4	16
120		140

$$Mean, \bar{x} = \frac{\sum x}{N} = \frac{120}{8} = 15$$

$$Thus, Standard\ Deviation\ (SD\ or\ \sigma) = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{(x - \bar{x})^2}{N}} = \sqrt{\frac{140}{8}} = \sqrt{17.5} = 4.18$$

Example 10: The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

Student Nos	1	2	3	4	5	6	7	8	9	10
Marks	43	48	65	57	31	60	37	48	78	59

Solution: (Deviations from assumed mean)

Nos.	Marks (x)	$d = X - A, (A = 57)$	d^2
1	43	-14	196
2	48	-9	81
3	65	8	64
4	57	0	0
5	31	-26	676
6	60	3	9
7	37	-20	400
8	48	-9	81
9	78	21	441
10	59	2	4
n = 10		∑d=-44	$\sum d^2$ =1952

$$SD\ or\ \sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{195.2}{10} - \left(\frac{44}{10}\right)^2} = \sqrt{195.2 - 19.36} = \sqrt{175.84} = 13.26$$

Calculation of standard deviation:

Discrete Series:

There are three methods for calculating standard deviation in discrete series:

a) Actual mean methods: If the actual mean in fractions, the calculation takes lot of time and labour; and as such this method is rarely used in practice.

b) Assumed mean method: Here deviations are taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.

c) Step-deviation method: If the variable values are in equal intervals, then we adopt this method.

Example 11: Calculate Standard deviation from the following data.

X :	20	22	25	31	35	40	42	45

f	5	12	15	20	25	14	10	6

Solution:

Deviations from assumed mean

X	F	$d = x - A, (A = 31)$	d^2	fd	fd^2
20	5	-11	121	-55	605
22	12	-9	81	-108	972
25	15	-6	36	-90	540
31	20	0	0	0	0
35	25	4	16	100	400
40	14	9	81	126	1134
42	10	11	121	110	1210
45	6	14	196	84	1176
	N=107			∑fd=167	$\sum fd^2$= 6037

$$SD\ or\ \sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum d}{\sum f}\right)^2} = \sqrt{\frac{6037}{107} - \left(\frac{167}{107}\right)^2} = \sqrt{56.42 - 2.44} = \sqrt{53.98} = 7.35$$

Calculation of Standard Deviation –Continuous Series:

In the continuous series the method of calculating standard deviation is almost the same as in a discrete series. But in a continuous series, mid-values of the class intervals are to be found out. The step- deviation method is widely used.

Coefficient of Variation:

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into

a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

The coefficient of variation is obtained by dividing the standard deviation by the mean and multiplies it by 100. Symbolically,

$$Coefficient\ of\ Variation\ (CV)\ =\ \frac{SD}{Mean}\times 100\ =\ \frac{\sigma}{X}\times 100$$

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

Example 12: In two factories A and B located in the same industrial area, the average weekly wages (in SR) and the standard deviations are as follows:

Factory	Average (x)	Standard Deviation (σ)	No. of workers
A	34.5	5	476
B	28.25	4.5	524

Which factory A or B pays out a larger amount as weekly wages?

Which factory A or B has greater variability in individual wages?

Solution:

$$Given\ N1\ =\ 476;\ \bar{X}_1\ =\ 34.5\ and\ \sigma_1\ =\ 5$$

$$N2\ =\ 524, \bar{X}2\ =\ 28.5, \sigma 2\ =\ 4.5$$

1. Total wages paid by factory A

$$=\ 34.5\times 476$$

$$=\ \text{SR}16.422$$

Total wages paid by factory B

$$= 28.5 \times 524$$
$$= \text{SR. } 14{,}934$$

Therefore factory A pays out larger amount as weekly wages.

2. C. V. of distribution of weekly wages of factory A and B are

$$\text{CV (A)} = \frac{\sigma_1}{\bar{X}_1} \times 100 = \frac{5}{34.5} \times 100 = 14.49$$

$$\text{CV (B)} = \frac{\sigma_2}{\bar{X}_2} \times 100 = \frac{4.5}{28.5} \times 100 = 15.79$$

Factory B has greater variability in individual wages, since C.V. of factory B is greater than C.V of factory A.

Example 13: Prices of a particular commodity in five years in two cities are given below:

Price in city A	Price in city B
20	10
22	20
19	18
23	12
16	15

Which city has more stable prices?

Solution: Actual mean method

City A			City B		
Prices (X)	Deviations from $\bar{X}$ = 20 dx	dx^2	Prices (Y)	Deviations from $\bar{Y}$	Dy^2
20	0	0	10	-5	25

22	2	4	20	5	25
19	-1	1	18	3	9
23	3	9	12	-3	9
16	-4	16	15	0	0
Σx=100	Σdx=0	Σdx^2=30	Σy=75	Σdy=0	Σdy^2 =68

City A:

$$\text{Mean} = \frac{\sum X}{N} = \frac{100}{5} = 20;$$

$$\text{SD}\ (\sigma) = \sqrt{\frac{\sum dx^2}{N}} = \sqrt{\frac{30}{5}} = \sqrt{6} = 2.45$$

$$\text{CV (City A)} = = \frac{\text{SD}}{\text{Mean}} \times 100 = \frac{2.45}{20} \times 100 = 12.25\%$$

City B:

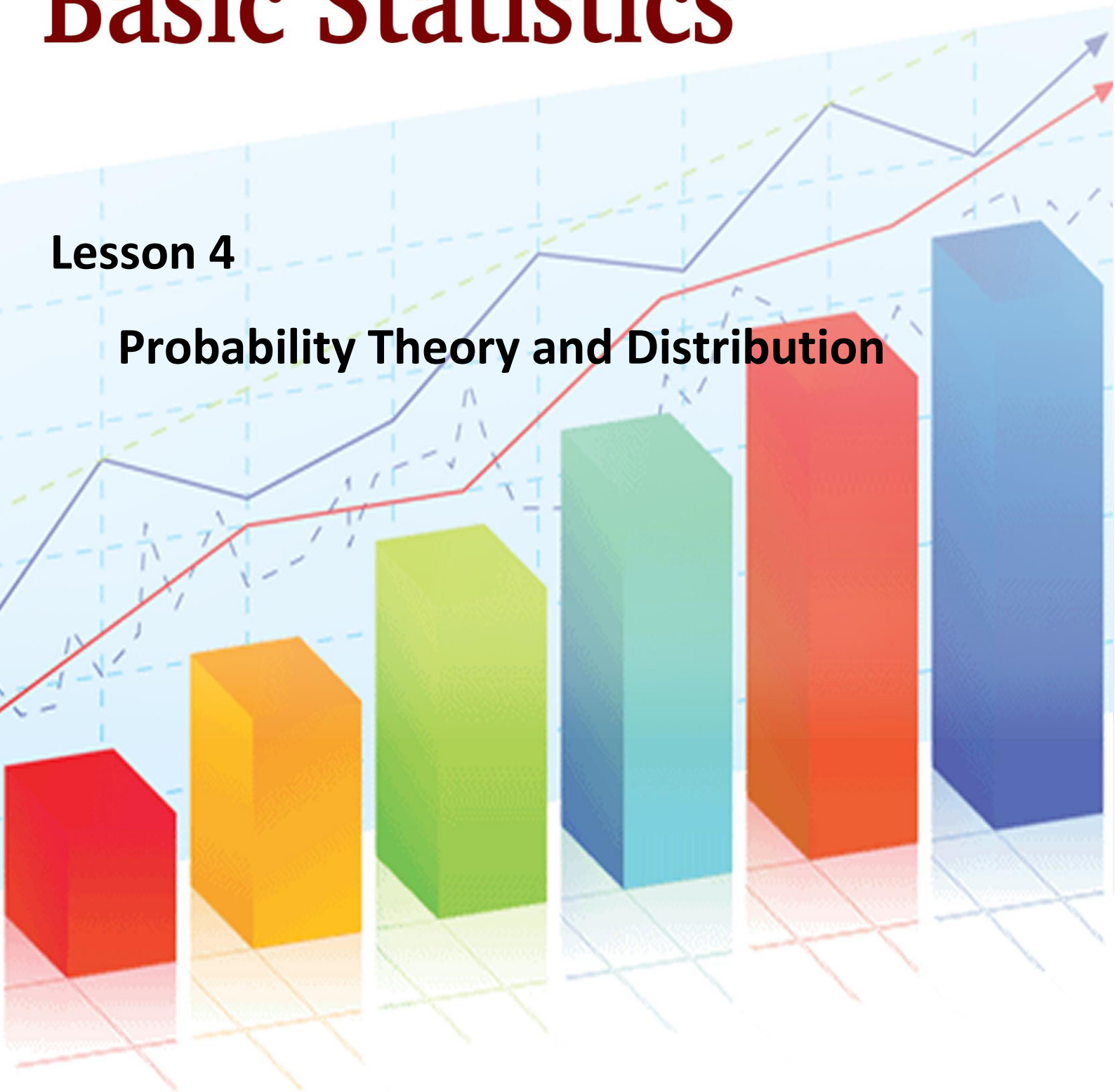
$$Mean = \frac{\sum X}{N} = \frac{75}{5} = 15;$$

$$SD\ (\sigma) = \sqrt{\frac{\sum dx^2}{N}} = \sqrt{\frac{68}{5}} = \sqrt{13.6} = 3.69$$

$$CV\ (City\ B) = = \frac{SD}{Mean} \times 100 = \frac{3.69}{15} \times 100 = 24.6\%$$

Therefore, City A had more stable prices than City B, because the coefficient of variation is less in City A.

Basic Statistics



Lesson 4

Probability Theory and Distribution

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 4	**Probability Theory and Distribution**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Lesson-4

Objectives of the Lesson:

1. Probability – Basic concepts
2. Equally likely, mutually exclusive, independent event
3. Additive and Multiplicative laws
4. Normal Distribution and its properties

Glossary of Terms: Sample Space, Event, Addison Law, Conditional Probability, Normal Distribution etc.

4.1 Introduction:

The concept of probability is difficult to define in precise terms. In ordinary language, the word probable means likely (or) chance. Generally the word, probability, is used to denote the happening of a certain event, and the likelihood of the occurrence of that event, based on past experiences. By looking at the clear sky, one will say that there will not be any rain today. On the other hand, by looking at the cloudy sky or overcast sky, one will say that there will be rain today. In the earlier sentence, we aim that there will not be rain and in the latter we expect rain. On the other hand a mathematician says that the probability of rain is '0' in the first case and that the probability of rain is '1' in the second case. In between 0 and 1, there are fractions denoting the chance of the event occurring. In ordinary language, the word probability means uncertainty about happenings. In Mathematics and Statistics, a numerical measure of uncertainty is provided by the important branch of statistics – called theory of probability. Thus we can say, that the theory of probability describes certainty by 1 (one), impossibility by 0 (zero) and uncertainties by the co-efficient which lies between 0 and 1.

Trial and Event An experiment which, though repeated under essentially identical (or) same conditions does not give unique results but may result

in any one of the several possible outcomes. Performing an experiment is known as a trial and the outcomes of the experiment are known as events.

Example 1: Seed germination – either germinates or does not germinates are events. In a lot of 5 seeds none may germinate (0), 1 or 2 or 3 or 4 or all 5 may germinate.

Sample space (S)

A set of all possible outcomes from an experiment is called sample space. For example, a set of five seeds are sown in a plot, none may germinate, 1, 2, 3, 4 or all five may germinate. i.e the possible outcomes are {0, 1, 2, 3, 4, 5. The set of numbers is called a sample space. Each possible outcome (or) element in a sample space is called sample point.

Exhaustive Events

The total number of possible outcomes in any trial is known as exhaustive events (or) exhaustive cases.

Example 1:

When pesticide is applied a pest may survive or die. There are two exhaustive cases namely (survival, death)

In throwing of a die, there are six exhaustive cases, since anyone of the 6 faces. 1, 2, 3, 4, 5, 6 may come uppermost.

In drawing 2 cards from a pack of cards the exhaustive number of cases is $^{52}C_2$, since 2 cards can be drawn out of 52 cards in $^{52}C_2$ ways

Trial	Random Experiment	Total number of trials	Sample Space
(1)	One pest is exposed to pesticide	21=2	{S,D}
(2)	Two pests are exposed to pesticide	22=4	{SS, SD, DS, DD}

(3)	Three pests are exposed to pesticide	23=8	{SSS, SSD, SDS, DSS, SDD, DSD,DDS, DDD}
(4)	One set of three seeds	41= 4	{0,1,2,3}
(5)	Two sets of three seeds	42=16	{0,1},{0,2},{0,3} etc

Favourable Events

The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event.

Example 2:

1. When a seed is sown if we observe non germination of a seed, it is a favourable event. If we are interested in germination of the seed then germination is the favourable event.

Mutually Exclusive Events

Events are said to be mutually exclusive (or) incompatible if the happening of any one of the events excludes (or) precludes the happening of all the others i.e.) if no two or more of the events can happen simultaneously in the same trial. (i.e.) The joint occurrence is not possible.

Example 3:

In observation of seed germination the seed may either germinate or it will not germinate. Germination and non germination are mutually exclusive events.

Equally Likely Events

Outcomes of a trial are said to be equally likely if taking in to consideration all the relevant evidences, there is no reason to expect one in preference to the others. (i.e.) Two or more events are said to be equally likely if each one of them has an equal chance of occurring.

Independent Events

Several events are said to be independent if the happening of an event is not affected by the happening of one or more events.

Example

When two seeds are sown in a pot, one seed germinates. It would not affect the germination or non germination of the second seed. One event does not affect the other event.

Dependent Events

If the happening of one event is affected by the happening of one or more events, then the events are called dependent events.

Example 4:

If we draw a card from a pack of well shuffled cards, if the first card drawn is not replaced then the second draw is dependent on the first draw.

Note: In the case of independent (or) dependent events, the joint occurrence is possible.

4.2 Definition of Probability

4.2.1 Mathematical (or) Classical (or) a-priori Probability

If an experiment results in 'n' exhaustive cases which are mutually exclusive and equally likely cases out of which 'm' events are favourable to the happening of an event 'A', then the probability 'p' of happening of 'A' is given by

$$p = P(A) = \frac{Favourable\ number\ of\ case}{Exhaustive\ number\ of\ cases} = \frac{m}{n}$$

Note

1. If m = 0 ⇒ P(A) = 0, then 'A' is called an impossible event. (i.e.) also by P(φ) = 0.

2. If m = n ⇒ P(A) = 1, then 'A' is called assure (or) certain event.
3. The probability is a non-negative real number and cannot exceed unity (i.e.) lies between 0 to 1.
4. The probability of non-happening of the event 'A' (i.e.) P($\bar{A}$) It is denoted by 'q'.

$$P(\bar{A}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(A)$$
$$\Rightarrow q = 1 - p$$
$$\Rightarrow p + q = 1$$
$$or\ P(A) + P(\bar{A}) = 1$$

4.2.2 Statistical (or) Empirical Probability (or) a-posteriori Probability

If an experiment is repeated a number (n) of times, an event 'A' happens 'm' times then the statistical probability of 'A' is given by

$$p = P(A) = \lim_{n\to\infty} \frac{m}{n}$$

4.2.3 Axioms for Probability

1. The probability of an event ranges from 0 to 1. If the event cannot take place its probability shall be '0' if it certain, its probability shall be '1'.

$$Let\ E_1, E_2, \ldots, E_n\ be\ any\ events, the\ P(E_i \geq 0)$$

2. The probability of the entire sample space is '1'. (i.e.) P(S) = 1.

$$, \text{Total Probability}, \sum_{i=1}^{n} P(E_i) = 1$$

3. If A and B are mutually exclusive (or)disjoint events then the probability of occurrence of either A (or) B denoted by P(AUB) shall be given by

$$P(A \cup B) = P(A) + P(B)$$
$$P(E1 \cup E2 \cup \ldots . \cup En) = P(E1) + P(E2) + \ldots\ldots + P(En)$$

If E1, E2, …., En are mutually exclusive events.

Example 5: Two dice are tossed. What is the probability of getting (i) Sum 6 (ii) Sum 9?

Solution

When 2 dice are tossed. The exhaustive number of cases is 36 ways.

(i) Sum 6 = {(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)}

$$favourable\ number\ of\ cases = 5, P(Sum\ 6) = \frac{5}{36}$$

(ii) Sum 9 = {(3,6), (4,5), (5,4), (6,3)}

∴ Favourable number of cases = 4

$$P(Sum\ 9) = \frac{4}{36} = \frac{1}{9}$$

Example 6: A card is drawn from a pack of cards. What is a probability of getting (i) a king (ii) a spade (iii) a red card (iv) a numbered card?

Solution

There are 52 cards in a pack. One can be selected in ${}^{52}C_1$ ways.

∴ Exhaustive number of cases is = ${}^{52}C_1 = 52$.

(i) A king

There are 4 kings in a pack.

One king can be selected in ${}^{4}C_1$ ways.

∴ Favourable number of cases is = ${}^{4}C_1 = 4$

Hence the probability of getting a king = $\frac{4}{52} = \frac{1}{13}$

(ii) A spade

There are 13 kings in a pack.

One spade can be selected in ${}^{13}C1$ ways.

∴ Favourable number of cases is = ${}^{13}C1 = 13$

Hence the probability of getting a spade = $\frac{13}{52} = \frac{1}{4}$

(iii) A red card

There are 26 kings in a pack.

One red card can be selected in $^{26}C1$ ways.

∴ Favourable number of cases is = $^{26}C1 = 26$

Hence the probability of getting a red card $=\frac{26}{52} = \frac{1}{2}$

(iv) A numbered card

There are 36 kings in a pack.

One numbered card can be selected in $^{36}C1$ ways.

∴ Favourable number of cases is = $^{36}C1 = 36$

Hence the probability of getting a numbered card = $\frac{36}{52}$

Example 7: What is the probability of getting 53 Sundays when a leap year selected at random?

Solution

A leap year consists of 366 days.

This has 52 full weeks and 2 days remained.

The remaining 2 days have the following possibilities.

(i) Sun. Mon
(ii) Mon, Tues
(iii) Tues, Wed
(iv) Wed, Thurs
(v) Thurs, Fri
(vi) Fri, Sat
(vii) Sat, Sun.

In order that a lap year selected at random should contain 53 Sundays, one of the 2 over days must be Sunday.

Exhaustive number of cases is = 7

∴ Favourable number of cases is = 2

∴ Required Probability is $=\frac{2}{7}$

4.3 Conditional Probability:

Two events A and B are said to be dependent, when B can occur only when A is known to have occurred (or vice versa). The probability attached to such an event is called the conditional probability and is denoted by P (A/B) (read it as: A given B) or, in other words, probability of A given that B has occurred.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}$$

If two events A and B are **dependent**, then the conditional probability of B given A is,

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(AB)}{P(A)}$$

There are two important theorems of probability namely,

1. The addition theorem on probability
2. The multiplication theorem on probability.

4.3.1 Addition Theorem on Probability

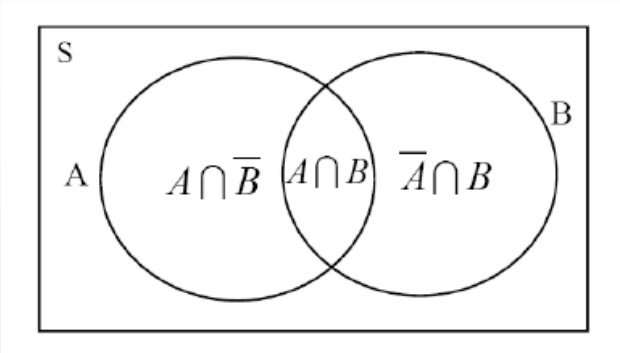
(i) Let A and B be any two events which are **not mutually exclusive**

$$\begin{aligned} P\,(A\ or\ B) &= P\,(A \cup B) = P\,(A + B) \\ &= P\,(A) + P\,(B) - P\,(A \cap B)(or) \\ &= P\,(A) + P\,(B) - P\,(AB) \end{aligned}$$

Proof

Let us take a random experiments with a sample space S of N sample points. The by the definition of probability,

$$\boldsymbol{P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N}}$$



From the diagram, using the axiom for the mutually exclusive events, we write

$$P(A \cup B) = \frac{n(A) + n(\bar{A} \cap B)}{N}$$

Adding and subtracting $n(A \cap B)$ in the numinator

$$= \frac{n(A) + n(\bar{A} \cap B) + n(A \cap B) - n(A \cap B)}{N}$$

$$\frac{n(A) + n(B) - n(A \cap B)}{N}$$

$$= \frac{n(A)}{N} + \frac{n(B)}{n} - \frac{n(A \cap B)}{N}$$

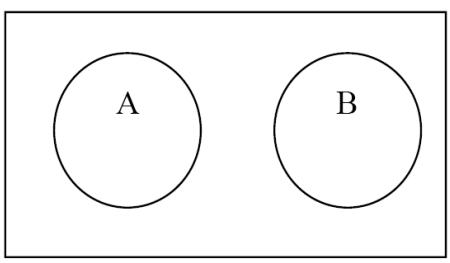
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(iii) Lat A and B be any two events which are mutually excusive

events then

$$P(A \text{ or } B) = P(A \cup B) = P(A + B) = P(A) + P(B)$$

Proof:



We know thaty, $n(A \cup B) = n(A) + n(B)$

$$P(A \cup B) = \frac{n(A \cup B)}{n}$$

$$\frac{n(A) + n(B)}{n}$$

$$= \frac{n(A)}{n} + \frac{n(B)}{n}$$

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

Note

(1) In the case of 3 events, (not mutually exclusive events)

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

(2) In the case of 3 events (mutually exclusive events)

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

Example : Using the additive law of probability we can find the probability that in one roll of a die, we will obtain either a one-spot or a six-spot. The probability of obtaining a onespot is 1/6. The probability of obtaining a six-

spot is also 1/6. The probability of rolling a die and getting a side that has both a one-spot with a six-spot is 0. There is no side on a die that has both these events. So substituting these values into the equation gives the following result

$$\frac{1}{6}+\frac{1}{6}-0=\frac{2}{6}=\frac{1}{3}=0.3333$$

Finding the probability of drawing a 4 of hearts or a 6 or any suit using the additive law of probability would give the following:

$$\frac{1}{52}+\frac{4}{52}-0=\frac{5}{52}=0.0962$$

There is only a single 4 of hearts, there are 4 sixes in the deck and there isn't a single card that is both the 4 of hearts and a six of any suit.

Now using the additive law of probability, you can find the probability of drawing either a king or any club from a deck of shuffled cards. The equation would be completed like this:

$$\frac{4}{52}+\frac{13}{52}-\frac{1}{52}=\frac{16}{52}=0.3077$$

There are 4 kings, 13 clubs, and obviously one card is both a king and a club. We don't want to count that card twice, so you must subtract one of it's occurrences away to obtain the result.

4.3.2 Multiplication Theorem on Probability

(i) If A and B be any two events which are not independent then

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

Where P(B/A) and P(A/B) are the conditional probability of B given A and A given B, respectively.

Proof

Let n is the total number of events

n (A) is the number of events in A

n (B) is the number of events in B

n (AU B) is the number of events in (AU B)

n (A∩B) is the number of events in (A∩B)

$$P(A \cap B) = \frac{n(A \cap B)}{n}$$

$$\frac{n(A \cap B)}{n} \times \frac{n(A)}{n(A)}$$

$$\frac{n(A)}{n} \times \frac{n(A \cap B)}{n(A)}$$

$$P(A \cap B) = P(A).P(B/A) \qquad (1)$$

$$P(A \cap B) = \frac{n(A \cap B)}{n}$$

$$\frac{n(A \cap B)}{n} \times \frac{n(B)}{n(B)}$$

$$\frac{n(B)}{n} \times \frac{n(A \cap B)}{n(B)}$$

$$P(A \cap B) = P(A).P(A/B) \qquad (2)$$

(ii) If A and B be any two events which are independent, then, P(B/A) = P(B) and P(A/B) = P(A)

$$P(A\ and\ B) = P(A \cap B) = P(AB) = P(A).P(B)$$

Note

(i) In the case of 3 events (dependent)

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

(ii) In the case of 3 events (independent)

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

Example 8:

So in finding the probability of drawing a 4 and then a 7 from a well shuffled deck of cards, this law would state that we need to multiply those separate probabilities together. Completing the equation above gives

$$P(4\ and\ 7) = \frac{4}{52} \times \frac{4}{52} = \frac{16}{2704} = 0.0059$$

Given a well shuffled deck of cards, what is the probability of drawing a Jack of Hearts, Queen of Hearts, King of Hearts, Ace of Hearts, and 10 of Hearts?

$$P(10, J, Q, K, A\ of\ hearts) = \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} = 0{,}000000026$$

In any case, given a well shuffled deck of cards, obtaining this assortment of cards, drawing one at a time and returning it to the deck would be highly unlikely (it has an exceedingly low probability).

4.4 Normal distribution

Continuous Probability distribution is normal distribution. It is also known as error law or Normal law or Laplacian law or Gaussian distribution. Many of the sampling distribution like student-t, f distribution and χ^2 distribution.

Definition

A continuous random variable x is said to be a normal distribution with parameters μ and σ^2, if the density function is given by the probability law

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

Note

The mean m and standard deviation s are called the parameters of Normal distribution. The normal distribution is expressed by $X \sim N(\mu, \sigma^2)$

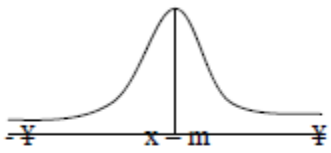
4.4.1 Condition of Normal Distribution

Normal distribution is a limiting form of the binomial distribution under the following conditions.

1. n, the number of trials is indefinitely large ie., $n \to \infty$ and
2. Neither p nor q is very small.
3. Normal distribution can also be obtained as a limiting form of Poisson distribution with parameter
4. Constants of normal distribution are mean = μ , variation =σ^2 , Standard deviation = σ .

4.4.2 Normal probability curve

The curve representing the normal distribution is called the normal probability curve. The curve is symmetrical about the mean (m), bell-shaped and the two tails on the right and left sides of the mean extends to the infinity. The shape of the curve is shown in the following figure.



4.4.3 Properties of normal distribution

1. The normal curve is bell shaped and is symmetric at $x = \mu$.

2. Mean, median, and mode of the distribution are coincide

i.e., Mean = Median = Mode = μ

3. It has only one mode at $x = \mu$ (i.e., unimodal)

4. The points of inflection are at $x = \mu \pm \sigma$

5. The maximum ordinate occurs at x = μ and its value is $=\frac{1}{\sigma\sqrt{2\pi}}$

6. Area Property

$P(\mu - \sigma < x < \mu + \sigma) = 0.6826$

$P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.9544$

$P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.9973$

4.4.4 Standard Normal distribution

Let X be random variable which follows normal distribution with mean m and variance σ^2 .The standard normal variate is defined as $Z = \frac{X-\mu}{\sigma}$which follows

standard normal distribution with mean 0 and standard deviation 1 i.e., Z ~ N(0,1). standard normal distribution is given by $\phi(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}$

The advantage of the above function is that it doesn't contain any parameter. This enables us to compute the area under the normal probability curve.

Note

Property of $\phi(Z)$

1. $\phi(-Z) = 1 - \phi(Z)$
2. $P(2 \leq Z \leq b) = \phi(b) - \phi(a)$

Example 9: In a normal distribution whose mean is 12 and standard deviation is 2. Find the probability for the interval from x = 9.6 to x = 13.8

Solution

Given that Z~ N (12, 4)

$$P(1.6 \leq Z \leq 13.8) = P\left(\frac{9.6 - 12}{2} \leq Z \leq \frac{13 - 8 - 12}{2}\right)$$

$$= P(-1.2 \leq Z \leq 0) + P(0 \leq Z \leq 0.9)$$

$$= P(0 \leq Z \leq 1.2) + P(0 \leq Z \leq 0.9) \; [by\ using\ symmetric\ property]$$

$$= 0.3849 + 0.3159$$

$$= 0.7008$$

When it is converted to percentage (ie) 70% of the observations are covered between 9.6 to 13.8.

Example 10: For a normal distribution whose mean is 2 and standard deviation 3. Find the value of the variate such that the probability of the variate from the mean to the value is 0.4115

Solution:

Given that Z~ N (2, 9)

To find X1:

We have $P(2 \leq Z \leq X_1) = 0.4115$

$$P\left(\frac{2-2}{3} \leq \frac{X-\mu}{\sigma} \leq \frac{X_1-2}{3}\right) = 0.4115$$

$P(0 \leq Z \leq Z_1) = 0.4115$, where $Z_1 = \frac{X_1-2}{3}$

[From the normal table where 0.4115 lies is rthe value of Z_1]

Form the normal table we have Z_1=1.35

$$\therefore 1.35 = \frac{X_1 - 2}{3}$$

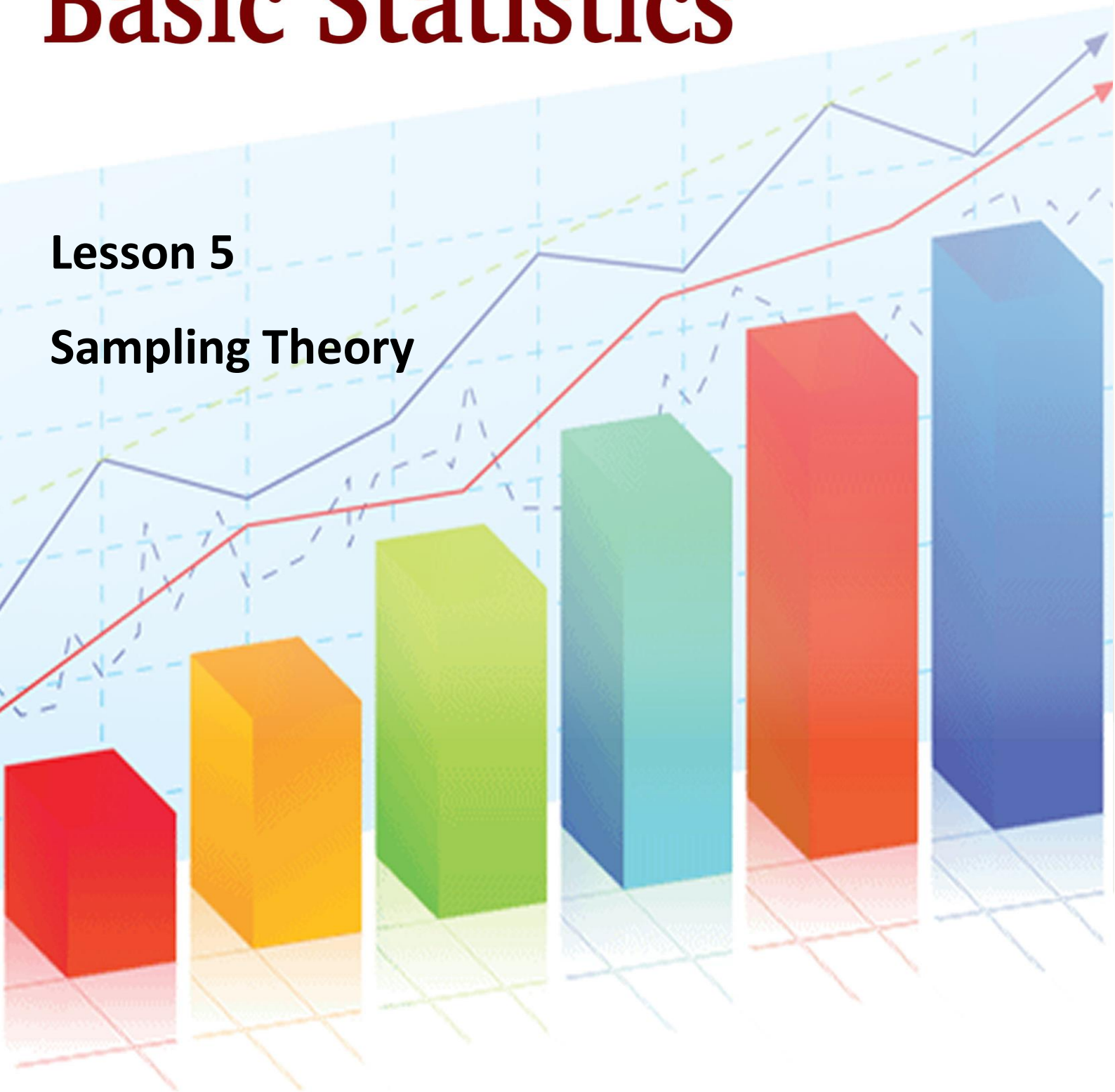
$$\Rightarrow 3(1.35) + 2 = X1$$

$$= X_1 = 6.05$$

(i.e) 41 % of the observation converged between 2 and 6.05

Basic Statistics

Basic Statistics



Lesson 5

Sampling Theory

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 5	**Sampling Theory**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Lesson-5

Objectives of the Lesson:

1. Sampling-basic concepts
2. Sampling methods
3. Simple random sampling
4. Stratified random sampling

Glossary of Terms: Sampling, Simple random Sampling, Stratified Sampling, Census Survey etc.

5.1 Basic Terminology of Sampling Theory:

Population (Universe)

Population means aggregate of all possible units. It need not be human population. It may be population of plants, population of insects, population of fruits, etc.

Finite population

When the number of observation can be counted and is definite, it is known as finite population

- No. of plants in a plot.
- No. of farmers in a village.
- All the fields under a specified crop.

Infinite population

When the number of units in a population is innumerably large, that we cannot count all of them, it is known as infinite population.

- The plant population in a region.
- The population of insects in a region.

Frame

A list of all units of a population is known as frame.

Parameter

A summary measure that describes any given characteristic of the population is known as parameter. Population are described in terms of certain measures like mean, standard deviation etc. These measures of the population are called parameter and are usually denoted by Greek letters. For example, population mean is denoted by μ, standard deviation by σ and variance by σ^2.

Sample

A portion or small number of unit of the total population is known as sample.

- All the farmers in a village(population) and a few farmers(sample)
- All plants in a plot is a population of plants.
- A small number of plants selected out of that population is a sample of plants

Statistic

A summary measure that describes the characteristic of the sample is known as statisitic. Thus sample mean, sample standard deviation etc is statistic. The statistic is usually denoted by roman letter.

$\bar{x} - \ sample\ mean$

$s - standard\ deviation$

The statistic is a random variable because it varies from sample to sample.

Sampling

The method of selecting samples from a population is known as sampling.

5.2 Sampling technique

There are two ways in which the information is collected during statistical survey.

They are

- Census survey
- Sampling survey

5.2.1 Census

It is also known as population survey and complete enumeration survey. Under census survey the information are collected from each and every unit of the population or universe.

5.2.2 Sample survey

A sample is a part of the population. Information are collected from only a few units of a population and not from all the units. Such a survey is known as sample survey. Sampling technique is universal in nature, consciously or unconsciously it is adopted in everyday life.

For example:

1. A handful of rice is examined before buying a sack.
2. We taste one or two fruits before buying a bunch of grapes.
3. To measure root length of plants only a portion of plants are selected from a plot.

5.3 Need for sampling

The sampling methods have been extensively used for a variety of purposes and in great diversity of situations.

In practice it may not be possible to collected information on all units of a population due to various reasons such as

1. Lack of resources in terms of money, personnel and equipment.
2. The experimentation may be destructive in nature. Eg- finding out

the germination percentage of seed material or in evaluating the efficiency of an insecticide the experimentation is destructive.

3. The data may be wasteful if they are not collected within a time limit. The census survey will take longer time as compared to the sample survey. Hence for getting quick results sampling is preferred. Moreover a sample survey will be less costly than complete enumeration.
4. Sampling remains the only way when population contains infinitely many number of units.
5. Greater accuracy.

5.4 Sampling methods

The various methods of sampling can be grouped under

1. Probability sampling or random sampling
2. Non-probability sampling or non random sampling

5.5 Random sampling

Under this method, every unit of the population at any stage has equal chance (or) each unit is drawn with known probability. It helps to estimate the mean, variance etc of the population.

Under probability sampling there are two procedures

1. Sampling with replacement (SWR)
2. Sampling without replacement (SWOR)

When the successive draws are made with placing back the units selected in the preceding draws, it is known as sampling with replacement. When such replacement is not made it is known as sampling without replacement.

When the population is finite sampling with replacement is adopted otherwise SWOR is adopted.

Mainly there are many kinds of random sampling. Some of them are.

1. Simple Random Sampling
2. Systematic Random Sampling
3. Stratified Random Sampling
4. Cluster Sampling

5.5 Simple Random sampling (SRS)

The basic probability sampling method is the simple random sampling. It is the simplest of all the probability sampling methods. It is used when the population is homogeneous.

When the units of the sample are drawn independently with equal probabilities. The sampling method is known as Simple Random Sampling (SRS). Thus if the population consists of N units, the probability 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 of size N. There are ^{N}Cn possible samples of size n. If all possible samples have an equal probability 1/NCn of being drawn, the sampling is said be simple random sampling.

There are two methods in SRS

1. Lottery method
2. Random no. table method

5.5.1 Lottery method

This is most popular method and simplest method. In this method all the items of the universe are numbered on separate slips of paper of same size, shape and color. They are folded and mixed up in a drum or a box or a container. A blindfold selection is made. Required number of slips is selected for the desired sample size. The selection of items thus depends on chance.

For example, if we want to select 5 plants out of 50 plants in a plot, we number the 50 plants first. We write the numbers from 1-50 on slips of the same size, role them and mix them. Then we make a blindfold selection of 5 plants. This method is also called unrestricted random sampling because units are selected from the population without any restriction. This method is mostly used in lottery draws. If the population is infinite, this method is inapplicable. There is a lot of possibility of personal prejudice if the size and shape of the slips are not identical.

5.5.2 Random number table method

As the lottery method cannot be used when the population is infinite, the alternative method is using of table of random numbers.

There are several standard tables of random numbers. But the credit for this technique goes to Prof. LHC. Tippet (1927). The random number table consists of 10,400 four-figured numbers. There are various other random numbers. They are fishers and Yates (19380 comprising of 15,000 digits arranged in twos. Kendall and B.B Smith (1939) consisting of 1, 00,000 numbers grouped in 25,000 sets of 4 digit random

numbers, Rand corporation (1955) consisting of 2, 00,000 random numbers of 5 digits each etc.,

5.5.3 Merits

1. There is less chance for personal bias.
2. Sampling error can be measured.
3. This method is economical as it saves time, money and labour.

5.5.4 Demerits

1. It cannot be applied if the population is heterogeneous.
2. This requires a complete list of the population but such up-to-date lists are not available in many enquires.
3. If the size of the sample is small, then it will not be a representative

of the population.

5.6 Stratified Sampling

When the population is heterogeneous with respect to the characteristic in which we are interested, we adopt stratified sampling.

When the heterogeneous population is divided into homogenous sub-population, the sub-populations are called strata. From each stratum a separate sample is selected using simple random sampling. This sampling method is known as stratified sampling.

We may stratify by size of farm, type of crop, soil type, etc.

The number of units to be selected may be uniform in all strata (or) may vary from stratum to stratum.

There are four types of allocation of strata

1. Equal allocation
2. Proportional allocation
3. Neyman's allocation
4. Optimum allocation

- If the number of units to be selected is uniform in all strata it is known as equal allocation of samples.
- If the number of units to be selected from a stratum is proportional to the size of the stratum, it is known as proportional allocation of samples.
- When the cost per unit varies from stratum to stratum, it is known as optimum allocation.
- When the costs for different strata are equal, it is known as Neyman's allocation.

5.6.1 Merits

1. It is more representative.
2. It ensures greater accuracy.
3. It is easy to administrate as the universe is sub-divided.

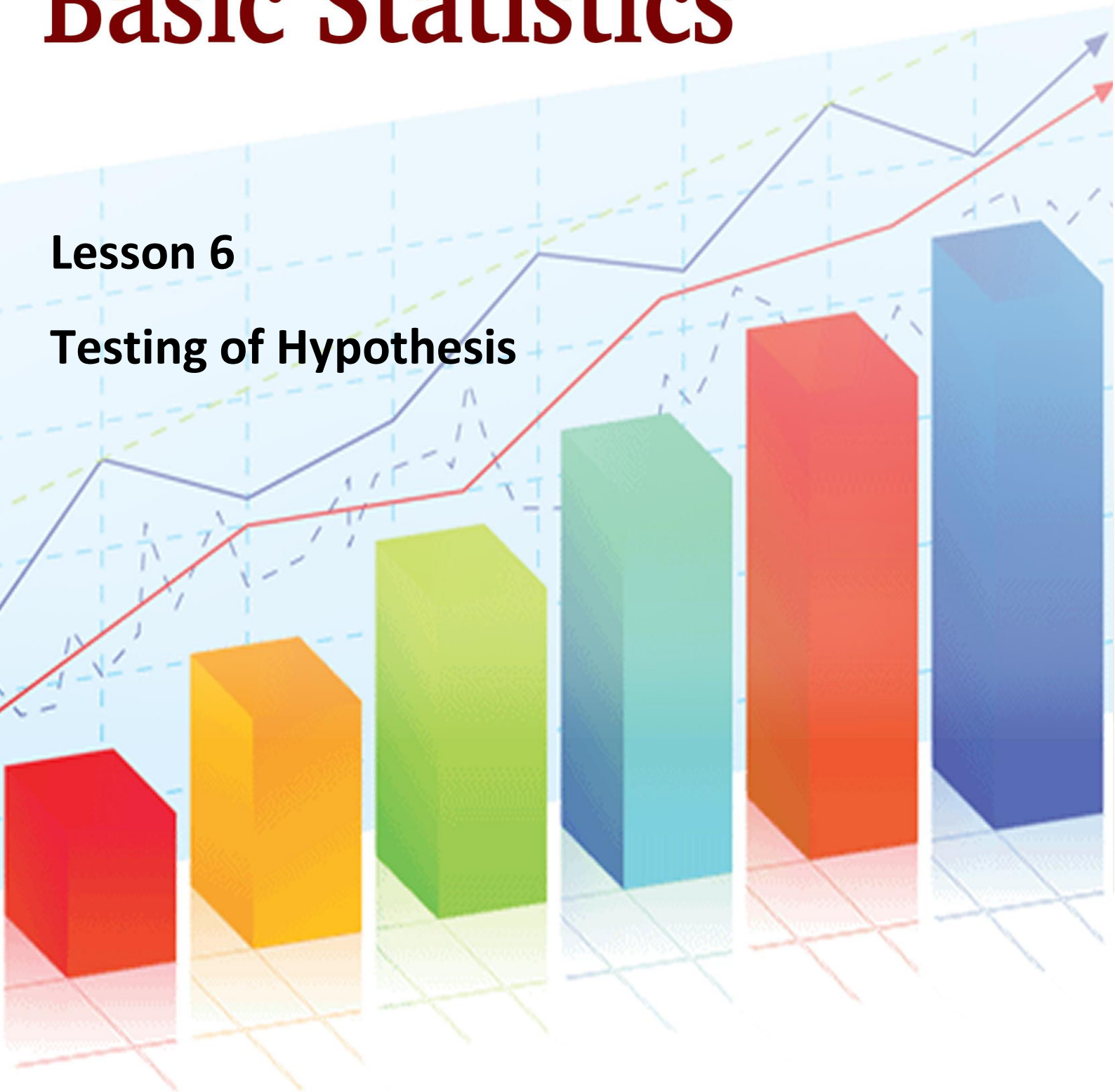
5.6.2 Demerits

1. To divide the population into homogeneous strata, it requires more money, time and statistical experience which is a difficult one.
2. If proper stratification is not done, the sample will have an effect of bias.

References:

1. Cochran, W.G. (1977), Sampling techniques, Wiley Eastern Limited.
2. Des Raj and Chandok. P. (1998), Sampling Theory. Narosa Publishing House. New Deihi.
3. Murthy, M.N. (1967), Sampling Theory and methods. Statistical Publishing Society. Calcutta.

Basic Statistics



Lesson 6

Testing of Hypothesis

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 6	**Testing of Hypothesis**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Lesson-6

Objectives of the Lesson:

1. Test of significance and Basic concepts
2. Null hypothesis, alternative hypothesis and level of significance
3. Standard error and its importance
4. Steps in testing of hypothesis with different tests

6.1 Sampling Distribution

By drawing all possible samples of same size from a population we can calculate the statistic, for example, $\bar{x}$ for all samples. Based on this we can construct a frequency distribution and the probability distribution of $\bar{x}$. Such probability distribution of a statistic is known a sampling distribution of that statistic. In practice, the sampling distributions can be obtained theoretically from the properties of random samples.

6.2 Standard Error

As in the case of population distribution the characteristic of the sampling distributions are also described by some measurements like mean & standard deviation. Since a statistic is a random variable, the mean of the sampling distribution of a statistic is called the expected valued of the statistic. The SD of the sampling distributions of the statistic is called standard error of the Statistic. The square of the standard error is known as the variance of the statistic. It may be noted that the standard deviation is for units whereas the standard error is for the statistic.

6.3 Theory of Testing Hypothesis

6.3.1 Hypothesis

Hypothesis is a statement or assumption that is yet to be proved.

6.3.2 Statistical Hypothesis

When the assumption or statement that occurs under certain conditions is formulated as scientific hypothesis, we can construct criteria by which a scientific hypothesis is either rejected or provisionally accepted. For this purpose, the scientific hypothesis is translated into statistical language. If the hypothesis in given in a statistical language it is called a statistical hypothesis.

For eg:-

The yield of a new paddy variety will be 3500 kg per hectare – scientific hypothesis.

In Statistical language if may be stated as the random variable (yield of paddy) is distributed normally with mean 3500 kg/ha.

6.3.3 Simple Hypothesis

When a hypothesis specifies all the parameters of a probability distribution, it is known as simple hypothesis. The hypothesis specifies all the parameters, i.e μ and σ of a normal distribution.

Eg:-

The random variable x is distributed normally with mean $\mu=0$ & SD=1 is a simple hypothesis. The hypothesis specifies all the parameters (μ & σ) of a normal distributions.

6.3.4 Composite Hypothesis

If the hypothesis specific only some of the parameters of the probability distribution, it is known as composite hypothesis. In the above example if only the μ is specified or only the σ is specified it is a composite hypothesis.

6.3.5 Null Hypothesis - Ho

Consider for example, the hypothesis may be put in a form 'paddy variety A will give the same yield per hectare as that of variety B' or there is no difference between the average yields of paddy varieties A and B. These hypotheses are in definite terms. Thus these hypothesis form a basis to work with. Such a working hypothesis in known as null hypothesis. It is called null hypothesis because if nullities the original hypothesis, that variety A will give more yield than variety B.

The null hypothesis is stated as 'there is no difference between the effect of two treatments or there is no association between two attributes (ie) the two attributes are independent. Null hypothesis is denoted by H_0.

Eg:-

There is no significant difference between the yields of two paddy varieties (or) they give same yield per unit area. Symbolically, H_0: $\mu_1=\mu_2$.

6.3.6 Alternative Hypothesis

When the original hypothesis is $\mu_1>\mu_2$ stated as an alternative to the null hypothesis is known as alternative hypothesis. Any hypothesis which is complementary to null hypothesis is called alternative hypothesis, usually denoted by H_1.

Eg:-

There is a significance difference between the yields of two paddy varieties.

Symbolically,

$$H_1: \mu_1 \neq \mu_2 \ (two\ sided\ or\ directionless\ alternative)$$

If the statement is that A gives significantly less yield than B (or) A gives significantly more yield than B. Symbolically,

$$H_1: \mu_1 < \mu_2 \ (one\ sided\ alternative - left\ tailed)$$

$$H_1: \mu_1 > \mu_2 \ (one\ sided\ alternative - right\ tailed)$$

6.4 Testing of Hypothesis

Once the hypothesis is formulated we have to make a decision on it. A statistical procedure by which we decide to accept or reject a statistical hypothesis is called testing of hypothesis.

6.5 Sampling Error

From sample data, the statistic is computed and the parameter is estimated through the statistic. The difference between the parameter and the statistic is known as the sampling error.

6.6 Test of Significance

Based on the sampling error the sampling distributions are derived. The observed results are then compared with the expected results on the basis of sampling distribution. If the difference between the observed and expected results is more than specified quantity of the standard error of the statistic, it is said to be significant at a specified probability level. The process up to this stage is known as test of significance.

6.7 Decision Errors

By performing a test we make a decision on the hypothesis by accepting or rejecting the null hypothesis Ho. In the process we may make a correct decision on Ho or commit one of two kinds of error.

- We may reject Ho based on sample data when in fact it is true. This error in decisions is known as Type I error.
- We may accept Ho based on sample data when in fact it is not true. It is known as Type II error.

	Accept Ho	Reject Ho
Ho is true	Correct Decision	Type I error
Ho is false	Type II error	Correct Decision

The relationship between type I & type II errors is that if one increases the other will decrease. The probability of type I error is denoted by α. The probability of type II error is denoted by β. The correct decision of rejecting the null hypothesis when it is false is known as the power of the test. The probability of the power is given by 1-β.

6.8 Critical Region

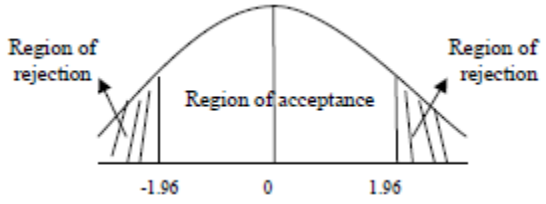
The testing of statistical hypothesis involves the choice of a region on the sampling distribution of statistic. If the statistic falls within this region, the null hypothesis is rejected: otherwise it is accepted. This region is called critical region.

Let the null hypothesis be H_0: $\mu_1 = \mu_2$ and its alternative be H_1: $\mu_1 \neq \mu_2$. Suppose Ho is true.

Based on sample data it may be observed that statistic $(\bar{x}_1 - \bar{x}_2)$ follows a normal distribution given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

We know that 95% values of the statistic from repeated samples will fall in the range $(\bar{x}_1 - \bar{x}_2) \pm 1.96\ times\ SE(\bar{x}_1 - \bar{x}_2)$. This is represented by a diagram



The border line value ±1.96 is the critical value or tabular value of Z. The area beyond the critical values (shaded area) is known as critical region or region of rejection. The remaining area is known as region of acceptance.

If the statistic falls in the critical region we reject the null hypothesis and, if it falls in the region of acceptance we accept the null hypothesis.

In other words if the calculated value of a test statistic (Z, t, χ^2 etc) is more than the critical value in magnitude it is said to be significant and we reject Ho and otherwise we accept Ho. The critical values for the t and are given in the form of readymade tables. Since the criticval values are given in the form of table it is commonly referred as table value. The table value depends on the level of significance and degrees of freedom.

Example: $Z_{cal} < Z_{tab}$ -We accept the H_0 and conclude that there is no significant difference between the means

Test Statistic

The sampling distribution of a statistic like Z, t, & χ2 are known as test statistic. Generally, in case of quantitative data

$$Test\ statistics = \frac{Statistic - Parameter}{Standard\ Error\ (Statistic)}$$

Note

The choice of the test statistic depends on the nature of the variable (ie) qualitative or quantitative, the statistic involved (i.e) mean or variance and the sample size, (i.e) large or small.

6.9 Level of Significance

The probability that the statistic will fall in the critical region is $\frac{\alpha}{2}+\frac{\alpha}{2}=\alpha$. This is nothing but the probability of committing type I error. Technically the probability of committing type I error is known as level of Significance.

6.10 One and two tailed test

The nature of the alternative hypothesis determines the position of the critical region. For example, if $H_1\ is\ \mu_1 \neq \mu_2$ it does not show the direction and hence the critical region falls on either end of the sampling distribution. If H_1 is $\mu_1 < \mu_2\ or\ \mu_1 > \mu_2$ the direction is known. In the first case the critical region falls on the left of the distribution whereas in the second case it falls on the right side.

6.10.1 One tailed test – When the critical region falls on one end of the sampling distribution, it is called one tailed test.

6.10.2 Two tailed test – When the critical region falls on either end of the sampling distribution, it is called two tailed test.

For example, consider the mean yield of new paddy variety (μ_1) is compared with that of a ruling variety (μ_2). Unless the new variety is more promising that the ruling variety in terms of yield we are not going to accept the new variety. In this case $H_1: \mu_1 > \mu_2$ for which one tailed test is used. If both the varieties are new our interest will be to choose the best of the two. In this case $H_1: \mu_1 \neq \mu_2$ for which we use two tailed test.

Degrees of freedom

The number of degrees of freedom is the number of observations that are free to vary after certain restriction have been placed on the data. If there are n observations in the sample, for each restriction imposed upon the original observation the number of degrees of freedom is reduced by one.

The number of independent variables which make up the statistic is known as the degrees of freedom and is denoted by (Nu)

Steps in testing of hypothesis

The process of testing a hypothesis involves following steps.

1. Formulation of null & alternative hypothesis.
2. Specification of level of significance.
3. Selection of test statistic and its computation.
4. Finding out the critical value from tables using the level of significance, sampling distribution and its degrees of freedom.
5. Determination of the significance of the test statistic.
6. Decision about the null hypothesis based on the significance of the test statistic.
7. Writing the conclusion in such a way that it answers the question on hand.

8.11 Large sample theory

The sample size n is greater than 30 ($n \geq 30$) it is known as large sample. For large samples the sampling distributions of statistic are normal (Z test). A study of sampling distribution of statistic for large sample is known as large sample theory.

6.12 Small sample theory

If the sample size n ils less than 30 ($n<30$), it is known as small sample. For small samples the sampling distributions are t, F and χ^2 distribution. A study of sampling distributions for small samples is known as small sample theory.

6.13 Test of Significance

The theory of test of significance consists of various test statistic. The theory had been developed

under two broad heading

1. Test of significance for large sample
   Large sample test or Asymptotic test or Z test (n≥30)
2. Test of significance for small samples(n<30)
   Small sample test or Exact test-t, F and χ2.

It may be noted that small sample tests can be used in case of large samples also.

6.13.1 Large sample test

Large sample test are

1. Sampling from attributes
2. Sampling from variables

6.13.2 Sampling from attributes

There are two types of test for attributes

1. Test for single proportion
2. Test for equality of two proportions

6.13.2.1 Test for single proportion

In a sample of large size n, we may examine whether the the sample would have come from a population having a specified proportion $P=P_0$. For testing

We may proceed as follows

1. **Null Hypothesis (H_0)**
   H_0: The given sample would have come from a population with specified proportion $P=P_0$
2. **Alternative Hypothesis(H_1)**
   H_1 : The given sample may not be from a population with specified proportion

$P \neq P_0$ (Two Sided)

$P > P_0$(One sided-right sided)

$P < P_0$(One sided-left sided)

3. **Test statistic**

$$Z = \frac{|p - P|}{\sqrt{\frac{PQ}{n}}}$$

It follows a standard normal distribution with $\mu=0$ and $\sigma^2=1$

4. **Level of Significance**

The level of significance may be fixed at either 5% or 1%

5. **Expected vale or critical value**

In case of test statistic Z, the expected value is

Ze = 1.96 at 5% level

2.58 at 1% level Two tailed test

Ze = 1.65 at 5% level

2.33 at 1% level One tailed test

6. **Inference**

If the observed value of the test statistic Z_o exceeds the table value Z_e we reject the Null Hypothesis Ho otherwise accept it.

6.13.2.2 Test for equality of two proportions

Given two sets of sample data of large size n_1 and n_2 from attributes. We may examine whether the two samples come from the populations having the same proportion. We may proceed as follows:

1. Null Hypothesis (Ho)

   Ho: The given two sample would have come from a population having the same proportion $P_1=P_2$

2. Alternative Hypothesis (H1)

   H_1 : The given two sample may not be from a population with specified proportion $P_1 \neq P2$ (Two Sided)

   $P_1>P_2$(One sided-right sided)

   $P_1<P_2$(One sided-left sided)

3. Test statistic

$$Z = \frac{|(p_1 - p_2) - (P_1 - P_2)|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

When $P_1$and P_2 are not known, then

$$Z = \frac{|p_1 - p_2|}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$

for heterogeneous population Where $q_1 = 1-p_1$ and $q_2 = 1-p_2$

$$Z = \frac{|p_1 - p_2|}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

for homogeneous population

p= combined or pooled estimate.

$$p = \frac{n_1p_1 + p_2n_2}{n_1 + n_2}$$

4. Level of Significance

   The level may be fixed at either 5% or 1%

5. Expected vale

The expected value is given by

Ze = 1.96 at 5% level

2.58 at 1% level Two tailed test

Ze = 1.65 at 5% level

2.33 at 1% level One tailed test

6. Inference

If the observed value of the test statistic Z exceeds the table value Ze we may reject the Null Hypothesis H_0 otherwise accept it.

6.13.3 Sampling from variable

In sampling for variables, the test are as follows

1. Test for single Mean
2. Test for single Standard Deviation
3. Test for equality of two Means
4. Test for equality of two Standard Deviation

6.13.3.1 Test for single Mean

In a sample of large size n, we examine whether the sample would have come from a population having a specified mean

1. **Null Hypothesis (H_0)**

H_0: There is no significance difference between the sample mean ie., $\mu=\mu_o$

or

The given sample would have come from a population having a specified mean ie., $\mu=\mu_0$

2. **Alternative Hypothesis(H_1)**

H_1 : There is significance difference between the sample mean is.

$\mu \neq \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_o$

3. Test statistics

$$Z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}}$$

When population variance is not known, it may be replaced by its estimate

$$Z = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}}, where\ s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

4. Level of Significance

The level may be fixed at either 5% or 1%

5. Expected vale

The expected value is given by

Ze = 1.96 at 5% level

2.58 at 1% level Two tailed test

Ze = 1.65 at 5% level

2.33 at 1% level One tailed test

6. **Inference**

If the observed value of the test statistic Z exceeds the table value Ze we may reject the Null Hypothesis H_0 otherwise accept it.

6.13.3.2 Test for equality of two Means

Given two sets of sample data of large size n_1 and n_2 from variables. We may examine whether the two samples come from the populations having the same mean. We may proceed as follows

1. **Null Hypothesis (H_0)**

   Ho: There is no significance difference between the sample mean ie., $\mu=\mu_0$

   or

   The given sample would have come from a population having a specified mean ie., $\mu_1=\mu_2$

2. **Alternative Hypothesis (H_1)**

   H1: There is significance difference between the sample mean ie., $\mu=\mu o$

   ie., $\mu_1\neq\mu_2$ or $\mu_1<\mu_2$ or $\mu_1>\mu_2$

3. **Test statistic**

   When the population variances are known and unequal (i.e) $\sigma_1^2 \neq \sigma_2^2$

$$Z = \frac{|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

   When $\sigma_1^2 = \sigma_2^2$

$$Z = \frac{|(\bar{x}_1 - \bar{x}_2)|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, where\ \sigma = \frac{(n_1\sigma_1^2 + n_2\sigma_2^2)}{(n_1 + n_2)}$$

   The equality of variances can be tested by using F test.

   When population variance is unknown, they may be replaced by their estimates $s_1^2\ and\ s_2^2$

$$Z = \frac{|(\bar{x}_1 - \bar{x}_2)|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, when\ s_1^2 \neq s_2^2$$

When $s_1^2 = s_2^2$

$$Z = \frac{|(\bar{x}_1 - \bar{x}_2)|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, where\ s = \frac{(n_1 s_1^2 + n_2 s_2^2)}{(n_1 + n_2)}$$

4. **Level of Significance**

   The level may be fixed at either 5% or 1%

5. **Expected vale**

   The expected value is given by

   The expected value is given by

   Ze = 1.96 at 5% level

   2.58 at 1% level Two tailed test

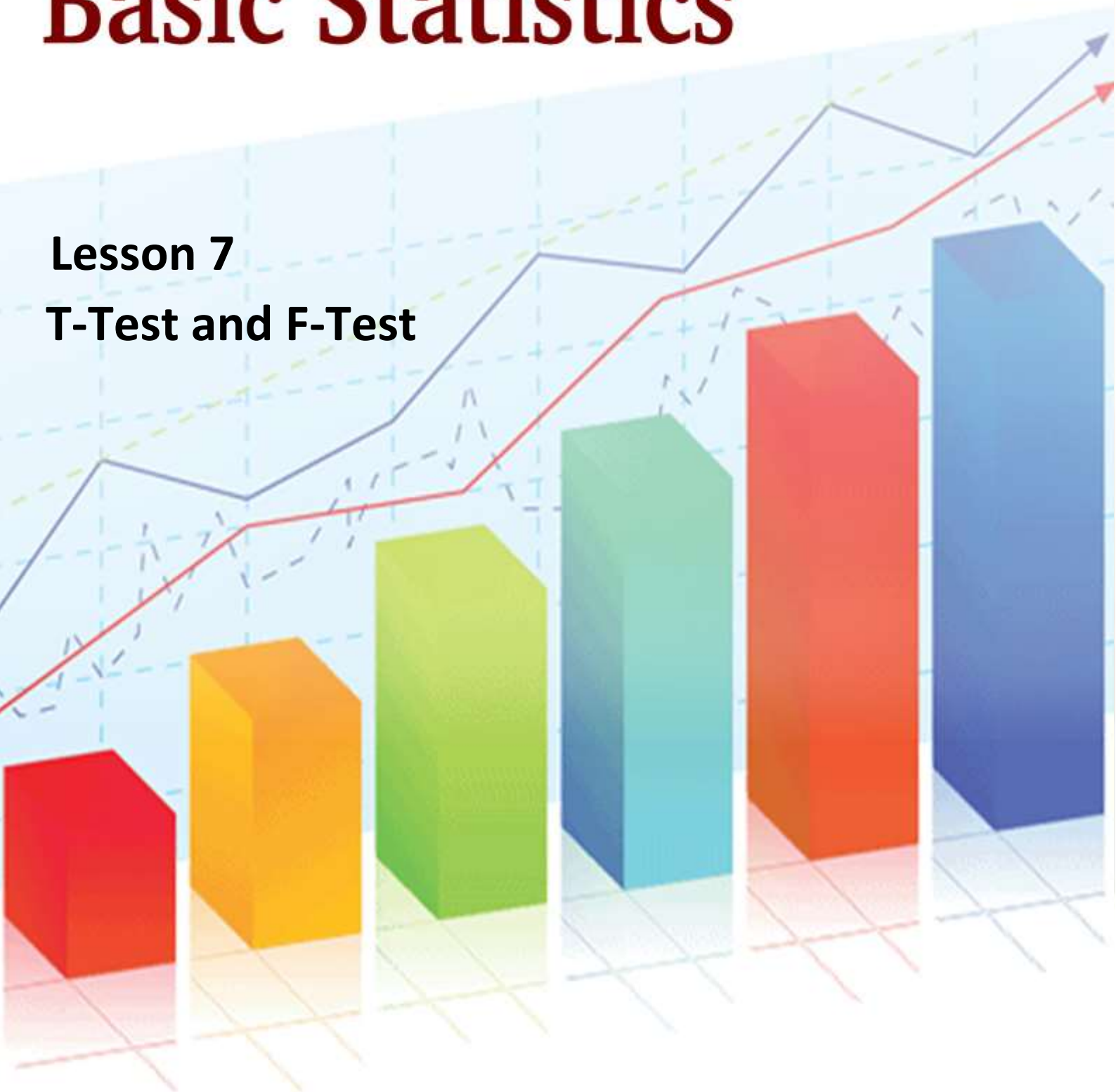
   Ze = 1.65 at 5% level

   2.34 at 1% level One tailed test

6. **Inference**

   If the observed value of the test statistic Z exceeds the table value Z we may reject the Null Hypothesis H_0 otherwise accept it.

Basic Statistics



Lesson 7

T-Test and F-Test

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 7	**T-Test and F-Test**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Objectives of the lesson:

1. Applications or uses of t-test
2. T-test for single mean
3. T-test for two means
4. T-test for paired sample
5. F-Test and its applications

Glossary of the lesson: Test-Statistic, Level of Significance, Independent Sample etc.

7.1 Student's t test

When the sample size is smaller, the ratio $Z = \frac{|\bar{x}-\mu|}{\frac{s}{\sqrt{n}}}$ will follow t distribution and not the standard normal distribution. Hence the test statistic is given as $t = \frac{|\bar{x}-\mu|}{\frac{s}{\sqrt{n}}}$ which follows normal distribution with mean 0 and unit standard deviation. This follows a t distribution with (n-1) degrees of freedom which can be written as t(n-1) d.f.

This fact was brought out by Sir William Gossest and Prof. R.A Fisher. Sir William Gossest published his discovery in 1905 under the pen name Student and later on developed and extended by Prof. R.A Fisher. He gave a test known as t-test.

7.2 Applications (or) uses

1 .To test the single mean in single sample case

2.To test the equality of two means in double sample case.

- Independent samples(Independent t test)
- Dependent samples (Paired t test)

2. To test the significance of observed correlation coefficient.
3. To test the significance of observed partial correlation coefficient.
4. To test the significance of observed regression coefficient.

7.2.1 Test for single Mean

1. **Form the null hypothesis**

Ho: $\mu=\mu$ (i.e) There is no significance difference between the sample mean and the population mean

2. **Form the Alternate hypothesis**

H1: $\mu\neq\mu_o$ (or $\mu>\mu_o$ or $\mu<\mu_o$) ie., There is significance difference between the sample mean and the population mean

3. **Level of Significance**

The level may be fixed at either 5% or 1%

4. **Test statistic**

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

Which follows t distribution with (n-1) degrees of freedom where

$$x = \frac{\sum x_i}{n} \text{ and } s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

where $\bar{x} = \frac{\sum x_i}{n}$

5. Find the table value of t corresponding to (n-1) d.f. and the specified level of significance.
6. **Inference**

If $t < t_{tab}$ we accept the null hypothesis H_0. We conclude that there is no significant difference sample mean and population mean

(or) if $t > t_{tab}$ we reject the null hypothesis H0. (ie) we accept the alternative hypothesis and conclude that there is significant difference between the sample mean and the population mean.

Example 1

Based on field experiments, a new variety of green gram is expected to given a yield of 12.0 quintals per hectare. The variety was tested on 10 randomly selected farmer's fields. The yield (quintals/hectare) were recorded as

14.3,12.6,13.7,10.9,13.7,12.0,11.4,12.0,12.6,13.1.

Do the results conform to the expectation?

Solution

1. **Null hypothesis**

H0: □=12.0

(i.e) the average yield of the new variety of green gram is 12.0 quintals/hectare.

2. **Alternative Hypothesis:**

H1:μ ≠ 12.0

(i.e) the average yield is not 12.0 quintals/hectare, it may be less or more than 12 quintals / hectare

3. Level of significance: 5 %

4. Test statistic:

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

From the given data

$$\sum x = 126.3 \, , \sum x^2 = 1605.77$$

$$\bar{x} = \frac{\sum x}{n} = \frac{126.3}{10} = 12.63$$

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{1605.77 - 1595.169}{9}} = \sqrt{\frac{10.601}{9}} = 1.0853$$

$$\frac{s}{\sqrt{n}} = \frac{1.0853}{\sqrt{10}} = 0.3432$$

$$\text{Now}, t = \left| \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \right|$$

$$t = \frac{12.63 - 12}{0.3432} = 1.836$$

Table value for t corresponding to 5% level of significance and 9 d.f. is 2.262

(two tailed test)

5. Inference

t < ttab

We accept the null hypothesis H_0

We conclude that the new variety of green gram will give an average yield of 12 quintals/hectare.

Note

Before applying t test in case of two samples the equality of their variances

has to be tested by using F-test

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1} d, f, \text{if } s_1^2 > s_2^2$$

or

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_2-1,n_1-1} d, f, \text{if } s_1^2 < s_2^2$$

where s_1^2 is the variance of the first sample whose size in n_1

s_2^2 is the variance of the second sample whose size is n_2

It may be noted that the numerator is always the greater variance. The critical value for F is read from the F table corresponding to a specified d.f. and level of significance Inference

F <Ftab

We accept the null hypothesis H0.(i.e) the variances are equal otherwise the variances are unequal.

7.2.2 Test for equality of two means (Independent Samples)

Given two sets of sample observation $x11, x12, x13 \ldots x1n$, and

$x21, x22, x23 \ldots x2n$ of sizes n_1 and n_2 respectively from the normal population.

1. **Using F-Test , test their variances**

(i) Variances are Equal

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$ $(\text{or } \mu_1 < \mu_2 \text{ or } \mu_1 > \mu_2)$

2. **Test statistics**

$$t = \frac{|(\bar{x}_1 - \bar{x}_2)|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where the combined variance $s^2 = \frac{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n_1}\right] + \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n_2}\right]}{n_1 + n_2 - 2}$

The test statistics t follows a t distribution with n1 + n2 -2 d.f.

ii) Variance are unequal and n1=n2

$$t = \frac{|(\bar{x}_1 - \bar{x}_2)|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

It follows a t distribution with $\left(\frac{n_1+n_2}{2}\right) - 1\ d.f.$

(i) Variances are unequal and n1≠n2

$$t = \left| \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \right|$$

This statistic follows neither t nor normal distribution but it follows

Behrens-Fisher d distribution. The Behrens – Fisher test is laborious one. An alternative simple method has been suggested by Cochran & Cox. In this method the critical value of t is altered as t_W (i.e) weighted t

$$t_x = \frac{t_1\left(\frac{s_1^2}{n_1}\right) + t_2\left(\frac{s_2^2}{n_2}\right)}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t1is the critical value for t with (n1-1) d.f. at a dspecified level of significance and t2 is the critical value for t with (n2-1) d.f. at a specified level of significance and

Example 2

In a fertilizer trial the grain yield of paddy (Kg/plot) was observed as follows

Under ammonium chloride 42,39,38,60 &41 kgs

Under urea 38, 42, 56, 64, 68, 69,& 62 kgs.

Find whether there is any difference between the sources of nitrogen?

Solution

H_o: $\mu_1 = \mu_2$ (i.e) there is no significant difference in effect between the sources of nitrogen.

H_1: $\mu_1 \neq \mu_2$ (i.e) there is a significant difference between the two sources

Level of significance = 5%

Before we go to test the means first we have to test their variances by using F-test. F-test

$$H_0: \sigma 1^2 = \sigma 2^2$$

$$H_1: \sigma 1^2 \neq \sigma 2^2$$

$$s_1^2 = \frac{\sum x_1^2 - \frac{(\sum x_1)^2}{n_1}}{n_1 - 1} = 82.5$$

$$s_2^2 = \frac{\sum x_2^2 - \frac{(\sum x_2)^2}{n_1}}{n_2 - 1} = 154.33$$

$\therefore$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_2-1, n_1-1} d, f, if\ s_1^2 < s_2^2$$

$$F = \frac{154.33}{32.5} = 1.8707$$

$F_{tab}(6,4)$ d.f. = 6.16

$F < F_{tab}$

We accept the null hypothesis H0. (i.e) the variances are equal. Use the test statistic

$$t = \frac{|(\bar{x}_1 - \bar{x}_2)|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s = \frac{\left[\sum x_1^2 - \frac{(\sum x_1)^2}{n_1}\right] + \left[\sum x_2^2 - \frac{(\sum x_2)^2}{n_2}\right]}{n_1 + n_2 - 1} = \frac{330 + 992}{10} = 125.6$$

$$t = \frac{|(44 - 57)|}{\sqrt{125.7\left(\frac{1}{7} + \frac{1}{75}\right)}} = 1.98$$

The degrees of freedom is 5+7-2= 10. For 5 % level of significance, table value of t is 2.228

Inference:

t <ttab

We accept the null hypothesis H_0

We conclude that the two sources of nitrogen do not differ significantly with regard to the grain yield of paddy.

7.3 F-Test:

A large number of research experiments are conducted to examine the effect of various factors on the production and quality attributes of milk and milk products. F-test is used either for testing the hypothesis about the equality of two population variances or the equality of two or more population means. The equality of two population means was dealt with t-test. Besides a t-test, we can also apply F-test for testing equality of two population means. Sir Ronald A. Fisher defined a statistic Z which is based upon ratio of two sample variances. In this lesson we will consider the distribution of ratio of two sample variances which was worked out by G.W. Snedecor.

7.3.1 F-Statistic:

Let X_{1i} (i=1,2,..,n_1) be a random sample of size $n_1$from the first population with variance σ_1^2 and X_{2j} (j=1,2,….,n_2) be another independent random sample of size n_2 from the second normal population with variance σ_2^2. The F- statistic is defined as the ratio of estimates of two variances as given below:

$$F = \frac{S_1^2}{S_2^2}$$

where, $S_1^2 > S_2^2$ and are unbiased estimates of population variances which are given by:

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$$

It follows Snedecor's F- distribution with $(n_1\text{-}1, n_2\text{-}1)$ d.f. i.e., $F \sim F(n_1 - 1, n_2 - 1)$. Further, if X is a χ^2-variate with n_1 d.f. and Y is another independent χ^2-variate with n_2 d.f., then F-statistic is defined as:

$$F = \frac{X/n_1}{Y/n_2}$$

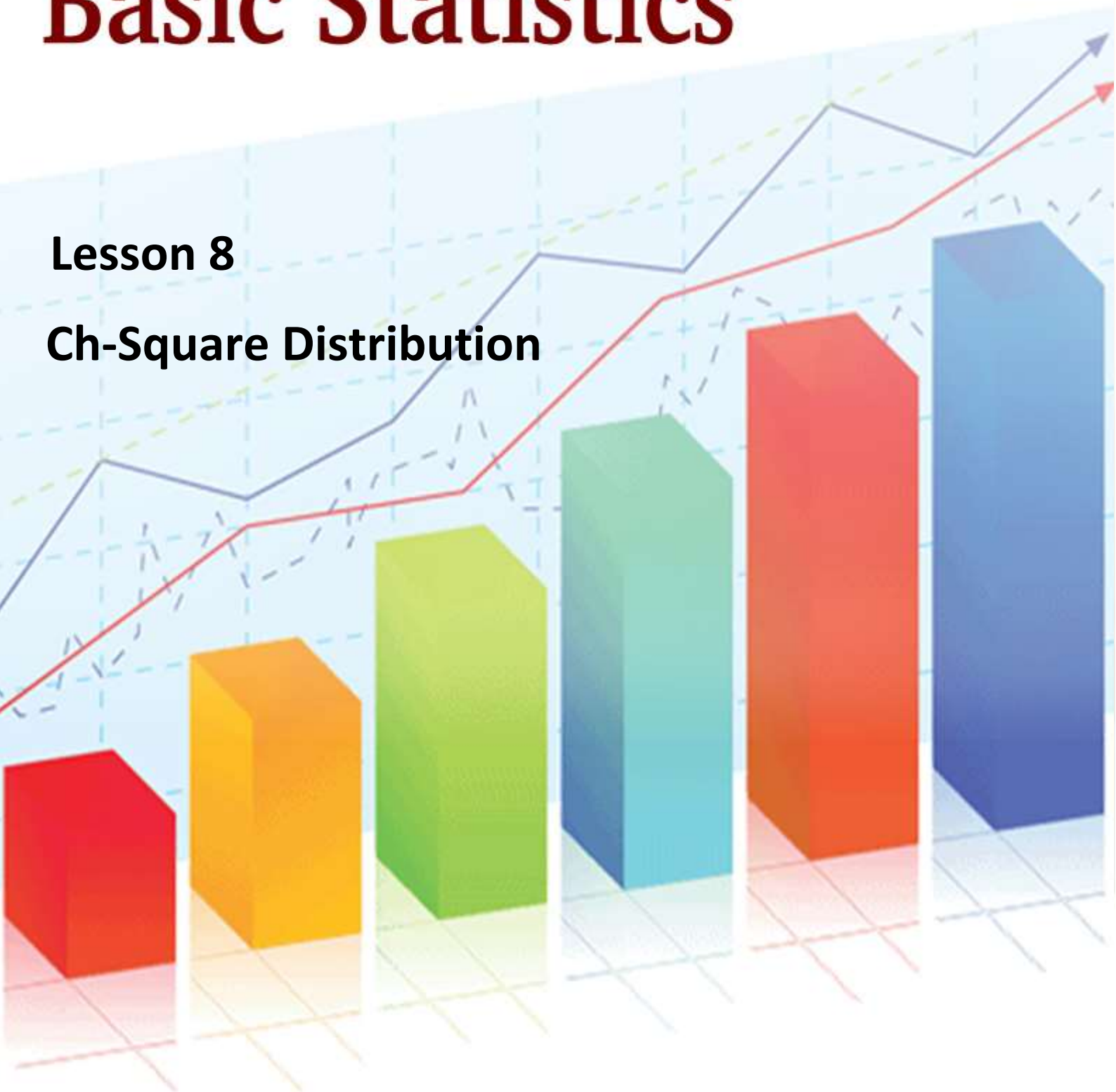
i.e. F-statistic is the ratio of two independent Chi-square variates divided by their respective degrees of freedom. This statistic follows G.W. Snedecor F distribution with (n_1, n_2) d.f. The sampling distribution of F-statistic does not involve any population parameter and depends only on the degrees of freedom n_1 and n_2.

7.4 Application of F- Distribution

F-distribution has a number of applications in statistics, some of which are given below

a) F-test for equality of population variances
b) F test for testing equality of several population means

Basic Statistics



Lesson 8

Ch-Square Distribution

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 8	**Ch-Square Distribution**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Objectives of the lesson:

1. Test for goodness of fit
2. Conditions for validity of chi-square test
3. Chi-Square (χ^2) test for independence of attributes
4. Yates correction for continuity

Glossary of the lesson: Test-Statistic, Level of Significance, Goodness of fit, Yates correction etc.

8.1 χ^2 distribution:

In case of attributes we cannot employ the parametric tests such as F and t. Instead we have to apply χ^2 test. When we want to test whether a set of observed values are in agreement with those expected on the basis of some theories or hypothesis then χ^2 statistic provides a measure of agreement between such observed and expected frequencies.

The χ^2 test has a number of applications. It is used to

1. Test the independence of attributes
2. Test the goodness of fit
3. Test the homogeneity of variances
4. Test the homogeneity of correlation coefficients
5. Test the equality of several proportions.
6. In genetics it is applied to detect linkage. Applications

8.2 χ^2 – test for goodness of fit

A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as "chi-square test of goodness of fit ".

If O_i $(i = 1,2, \ldots\ldots, n)$ is a set of observed (experimental frequencies) and E_i $(i = 1,2, \ldots\ldots, n)$ is the corresponding set of expected (theoretical or hypothetical) frequencies, then,

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

It follow a χ^2 distribution with n-1 d.f.. In case of χ^2 only one tailed test is used.

For Example : In plant genetics, our interest may be to test whether the observed segregation ratios deviate significantly from the mendelian ratios. In such situations we want to test the agreement between the observed and theoretical frequency, such test is called as test of goodness of fit.

8.3 Conditions for the validity of χ^2 –test

χ^2- test is an approximate test for large values of 'n' for the validity of χ^2-test of goodness of fit between theory and experiment, the following conditions must be satisfied.

1. The sample observations should be independent.
2. Constraints on the cell frequencies, if any, should be linear
   Example $\sum O_i = \sum E_i$
3. N, the total frequency should be reasonable large, say greater than (>) 50.
4. No theoretical cell frequency should be less than 5. If any theoretical 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 degree's of freedom lost in pooling

Example 1 :

The number of yeast cells counted in a haemocyto meter is compared to the theoretical value is given below. Does the experimental result support the theory?

No. of Yeast cells in the square	Obseved Frequency	Expected Frequency
0	103	106
1	143	141
2	98	93
3	42	41
4	8	14
5	6	5

Solution

H_0: the experimental results support the theory

H_1: the experimental results does not support the theory. Level of significance=5%

Test Statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 3.1779$$

Oi	Ei	Oi-Ei	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
103	106	-3	9	0.0849
143	141	2	4	0.0284
98	93	5	25	0.2688
42	41	1	1	0.0244
8	14	-6	36	2.5714
6	5	1	1	0.2000
400	400			**3.1779**

Table value

$$\chi^2_{6-1} = 5 \; at \; 5 \; \% \; level \; of \; significance = 10.070$$

Inference

$$As \; \chi^2_{cal} < \chi^2_{tab}$$

We accept the null hypothesis (i.e) there is a good correspondence between theory and experiment.

8.4 Chi-Square (χ^2) test for independence of attributes

At times we may consider two characteristics on attributes simultaneously. Our interest will be to test the association between these two attributes

Example:- An entomologist may be interested to know the effectiveness of different concentrations of the chemical in killing the insects. The concentrations of chemical form one attribute. The state of insects 'killed & not killed' forms another attribute. The result of this experiment can be arranged in the form of a contingency table. In general one attribute may be divided into m classes as $A_1, A_2, \ldots \ldots A_m$ and the other attribute may be divided into n classes as $B_1, B_2, \ldots \ldots B_n$. Then the contingency table will have m x n cells. It is termed as m x n contingency table

B\A	A_1	A_2	...	A_j	...	A_m	Row Total
B_1	O_{11}	O_{12}	...	O_{1j}	...	O_{1m}	r_1
B_2	O_{21}	O_{22}	...	O_{2j}	...	O_{2m}	r_2
.							
B_i	O_{i1}	O_{i2}		O_{ij}		O_{im}	r_i
.							
B_n	O_{n1}	O_{n2}		O_{ni}		O_{nm}	r_n
Col Total	c_1	c_2		c_2		c_n	$n = \sum r_i = \sum c_j$

where O_{ij} are observed frequencies.

The expected frequencies corresponding to O_{ij} is calculated as $\frac{r_i c_j}{n}$. *The* χ^2 is computed as

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

O_{ij} – observed frequencies

E_{ij} – Expected frequencies

n - number of rows

m - number of columns

It can be verified that $\sum O_{ij} = \sum E_{ij}$

This χ^2 is distributed as χ^2 with (n-1) d.f.

7.3.5 2x2 – contingency table:

When the number of rows and number of columns are equal to 2 it is termed as 2 x 2 contingency table. It will be in the following form

<table>
<tr><td></td><td>B_1</td><td>B_1</td><td colspan="2">Row Total</td></tr>
<tr><td>B_1</td><td>a</td><td>b</td><td>$a+b$</td><td>r_1</td></tr>
<tr><td>B_1</td><td>c</td><td>d</td><td>$c+d$</td><td>r_2</td></tr>
<tr><td rowspan="2">Col Tot</td><td>$a + c$</td><td>$b + d$</td><td rowspan="2">$a+b+c+d$</td><td rowspan="2">$= n$</td></tr>
<tr><td>c_1</td><td>a</td></tr>
</table>

Where a, b, c and d are cell frequencies c_1 and c_2 are column totals, r_1 and r_2 are row Totals and n is the total number of observations.

In case of 2 x 2 contingency table, χ^2 can be directly found using the short cut formula

$$\chi^2 = \frac{n(ad-bc)^2}{c_1 c_2 r_1 r_2}$$

The d.f. associated with χ^2 is (2-1)(2-1) = 1

8.5 Yates correction for continuity

If anyone of the cell frequency is < 5, we use Yates correction to make χ^2 as continuous. The Yates correction is made by adding 0.5 to the least cell frequency and adjusting the other cell frequencies so that the column and row totals remain same. Suppose, the first cell frequency is to be corrected then the contingency table will be as follows:

	B_1	B_1	Row Total
B_1	$a+0.5$	$b-0.5$	$a+b=r_1$
B_1	$c-0.5$	$d+0.5$	$c+d=r_2$
Col Tot	$a+c=c_1$	$b+d=c_2$	$a+b+c+d$

Then use the χ^2 – statistic as

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{c_1 c_2 r_1 r_2}$$

The d.f. associated with χ^2 is (2-1)(2-1) = 1

Example 2:

The severity of a disease and blood group were studied in a research projects. The findings are given in the following table, known as the m x n contingency table. Can this severity of the condition and blood group are associated. Severity of a disease classified by blood group in 1500 patients.

Condition	Blood Groups				Total
	O	A	B	AB	

Severe	51	40	10	9	110
Moderate	105	103	25	17	250
Mild	384	527	125	104	1140
Total	540	670	160	130	1500

Solution

H_0: The severity of the disease is not associated with blood group.

H_1: The severity of the disease is associated with blood group.

Condition	Blood Groups				Total
	O	A	B	AB	
Severe	51	40	10	9	110
Moderate	105	103	25	17	250
Mild	384	527	125	104	1140
Total	540	670	160	130	1500

Test statistics:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The d.f. associated with χ2 is (3-1)(4-1) = 6

Calculation of Expected frequencies

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
51	39.6	11.4	129.96	3.2818
40	49.1	-9.1	82.81	1.6866
10	11.7	-1.7	2.89	0.2470
9	9.5	-0.5	0.25	0.0263

105	90.0	15	225.00	2.5000
103	111.7	-8.7	75.69	0.6776
25	26.7	-1.7	2.89	0.1082
17	21.7	-4.7	22.09	1.0180
384	410.4	-26.4	696.96	1.6982
527	509.2	17.8	316.84	0.6222
125	121.6	3.4	11.56	0.0951
104	98.8	5.2	27.04	0.2737
Total				12.2347

Table value of χ 2 for 6 d.f. at 5% level of significance is 12.59

Inference

$\chi^2 < \chi^2_{tab}$

We accept the null hypothesis.

The severity of the disease has no association with blood group.

Example 3:

In order to determine the possible effect of a chemical treatment on the rate of germination of cotton seeds a pot culture experiment was conducted. The results are given below

Chemical treatment and germination of cotton seeds

	Germinated	Not germinated	Total
Chemically Treated	118	22	140
Untreated	120	40	160
Total	238	62	300

Does the chemical treatment improve the germination rate of cotton seeds?

Solution

H0:The chemical treatment does not improve the germination rate of cotton seeds.

H1: The chemical treatment improves the germination rate of cotton seeds.

Level of significance = 1%

Test statistic

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(n+d)} \; with \; 1 \; d.f.$$

$$\chi^2 = \frac{300(118 \times 40 - 22 \times 120)^2}{140 \times 160 \times 62 \times 238} = 3.927$$

Table value

$\chi^2(1) d.f. at \; 1 \; \% \; level of \; significance \; = \; 6.635$

Inference

$$\chi^2 < \chi^2_{tab}$$

We accept the null hypothesis.

The chemical treatment will not improve the germination rate of cotton seeds significantly.

Example 4

In an experiment on the effect of a growth regulator on fruit setting in muskmelon the following results were obtained. Test whether the fruit

setting in muskmelon and the application of growth regulator are independent at 1% level.

Aa	Fruit set	Fruit not set	Total
Treated	16	9	25
Control	4	21	25
Total	20	30	50

Solution

H_0:Fruit setting in muskmelon does not depend on the application of growth regulator.

H_1: Fruit setting in muskmelon depend on the application of growth regulator.

Level of significance = 1%

After Yates correction we have

	Fruit set	Fruit not set	Total
Treated	15.5	9.5	25
Control	4.5	20.5	25
Total	20	30	50

Test statistic

$$\chi^2 = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(n+d)}$$

$$\chi^2 = \frac{50\left(|15.5 \times 20.5 - 9.5 \times 4.5| - \frac{50}{2}|\right)^2}{25 \times 25 \times 20 \times 30} = 8.33$$

Table value

χ^2 (1) d.f. at 1 % level of significance is 6.635

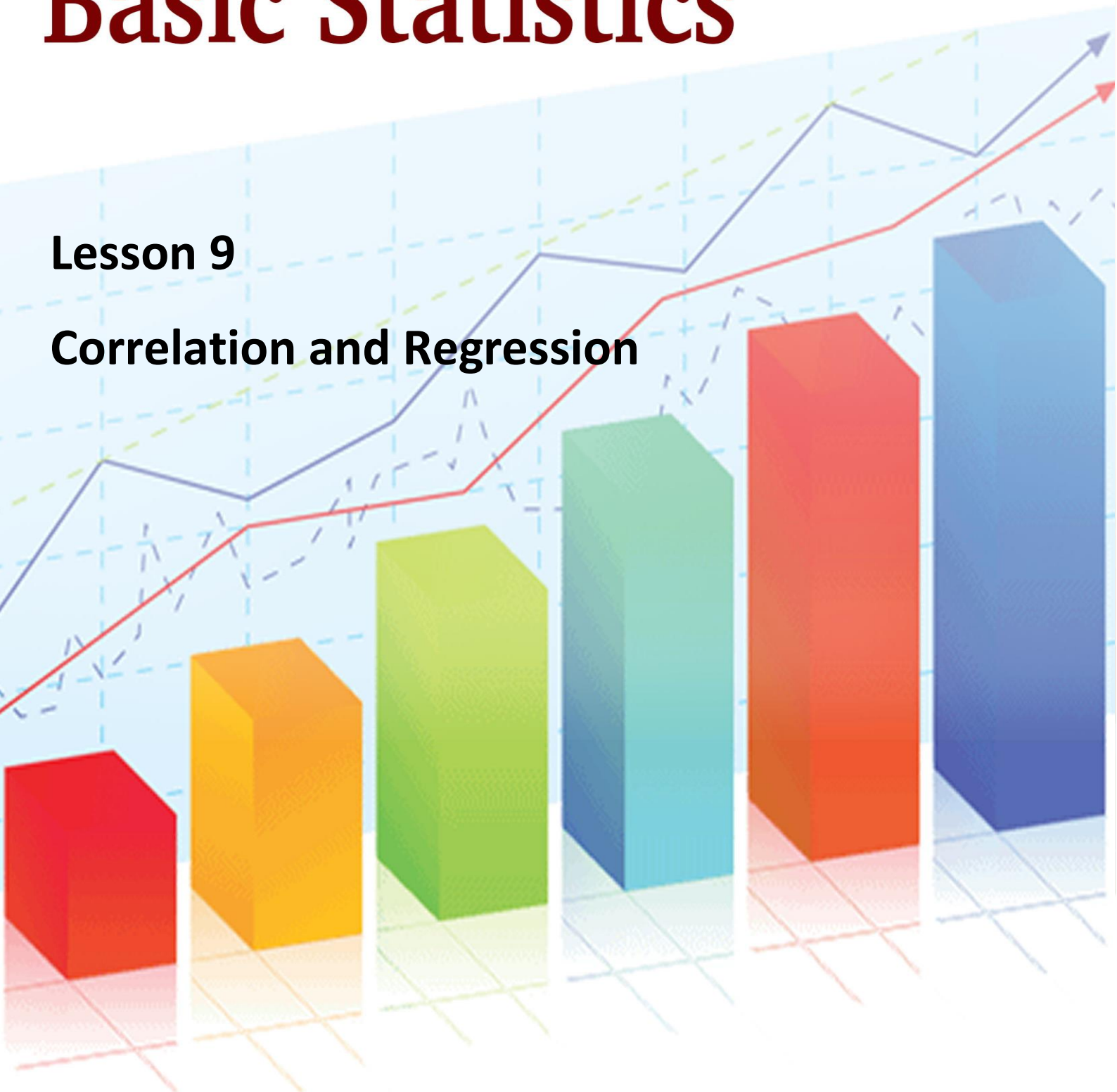
Inference

$$\chi^2 > \chi^2_{tab}$$

We reject the null hypothesis.

Fruit setting in muskmelon is influenced by the growth regulator. Application of growth regulator will increase fruit setting in musk melon.

Basic Statistics



Lesson 9

Correlation and Regression

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 9	**Correlation and Regression**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Lesson-9

Objectives of the Lecture:

1. Methods of Measuring the Correlation
2. Properties of Correlation
3. Spearman rank Correlation and its Properties
4. Regression and properties of regression coefficients

Glossary of terms: Correlation, Scatter Diagram, Regression, Regression Coefficient etc.

9.1 Introduction:

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only variable such as height, weight, ages, marks, wages, etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bivariate Analysis. Sometimes the variables may be inter-related. In health sciences we study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure, etc. The nature and strength of relationship may be examined by correlation and Regression analysis.

Thus Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each

other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

9.2 Uses of correlation:

1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

9.3 Scatter Diagram

To investigate whether there is any relation between the variables X and Y we use scatter diagram. Let $(x_1, y_1), (x_1, y_2) \ldots . (x_n, y_n)$ be n pairs of observations. If the variables X and Y are plotted along the X-axis and Y-axis respectively in the x-y plane of a graph sheet the resultant diagram of dots is known as scatter diagram. From the scatter diagram we can say whether there is any correlation between x and y and whether it is positive or negative or the correlation is linear or curvilinear.

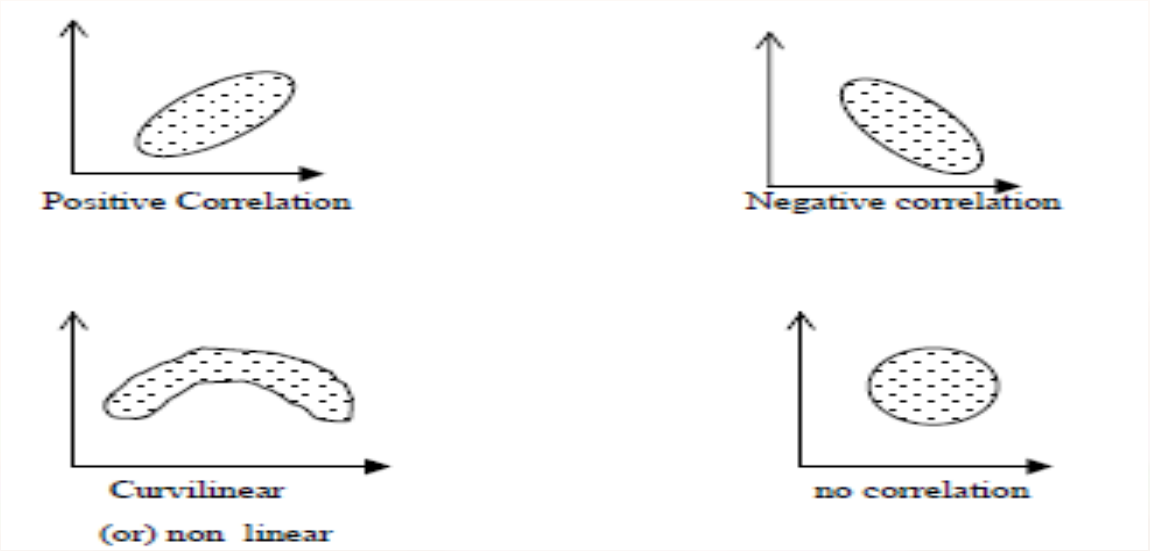
9.3.1 Types of Correlation

Positive correlation: Both variables tend to increase (or decrease) together.

Negative correlation: The two variables tend to change in opposite directions, with one increasing while the other decreases.

No correlation: There is no apparent (linear) relationship between the two variables.

Nonlinear relationship: The two variables are related, but the relationship results in a scatter diagram that does not follow a straight-line pattern



9.4 Classification of Correlation:

Correlation is classified into various types. The most important ones are

i) Positive and negative.
ii) Linear and non-linear.
iii) Partial and total.
iv) Simple and Multiple.

9.4.1 Positive and Negative Correlation:

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (i.e) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or

direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.

If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.

9.4.2 Linear and Non-linear correlation:

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

X	2	4	6	8	10	12
Y	3	6	9	12	15	18

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvilinear (or) non-linear correlation. The graph will be a curve.

9.4.3 Simple and Multiple correlations:

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlations.

9.4.4 Partial and total correlation:

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

9.5 Computation of correlation:

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure

of correlation (or) correlation coefficient and it is denoted by 'r'.

Co-variation:

The covariance between the variables x and y is defined as-

$$Cov\,(XY) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N}$$

$$= \frac{\sum xy}{N}$$

Where, X bar is the mean of X and Y bar is the mean of Y. x and y are deviations from its mean.

Karl Pearson's Coefficient of Correlation:

It is most widely used method in practice and it is known as Pearsonian Coefficient of Correlation. It is denoted by 'r'. The formula for calculating 'r' is-

$$r = \frac{Cov\,(x,y)}{\sigma x.\sigma y};\ Where\ \sigma_x = SD\ of\ X\ and\ \sigma_y = SD\ of\ Y$$

$$r = \frac{\sum xy}{N.\sigma x.\sigma y}$$

$$r = \frac{\sum xy}{\sqrt{\sum x2.\sum y2}}$$

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series respectively.

9.6 Properties of Correlation Coefficient:

Property 1: Correlation coefficient lies between –1 and +1.

Property 2: 'r' is independent of change of origin and scale.

Property 3: It is a pure number independent of units of measurement.

Property 4: Independent variables are uncorrelated but the converse is not true.

Property 5: Correlation coefficient is the geometric mean of two regression coefficients.

Property 6: The correlation coefficient of x and y is symmetric. $r_{xy} = r_{yx}$.

9.7 Limitations:

1. Correlation coefficient assumes linear relationship regardless of the assumption is correct or not.
2. Extreme items of variables are being unduly operated on correlation coefficient.
3. Existence of correlation does not necessarily indicate cause- effect relation.

9.8 Interpretation:

The following rules helps in interpreting the value of ' r' .

1. When r = 1, there is perfect +ve relationship between the variables.
2. When r = -1, there is perfect –ve relationship between the variables.
3. When r = 0, there is no relationship between the variables.
4. If the correlation is +1 or –1, it signifies that there is a high degree of correlation. (+ve or –ve) between the two variables.

If r is near to zero (ie) 0.1,-0.1, (or) 0.2 there is less correlation.

Example 1:

Find Karl Pearson's coefficient of correlation from the following data between height of father (x) and son (y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Comment on the result.

Solution:

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
64	66	-3	9	-2	4	6
65	67	-2	4	-1	1	2
66	65	-1	1	-3	9	3
67	68	0	0	0	0	0
68	70	1	1	2	4	2
69	68	2	4	0	0	0

70	72	3	9	4	16	12
469	476	0	28	0	34	25

$$Mean\ of\ X = \frac{\sum X}{N} = \frac{469}{7} = 67;$$

$$Mean\ of\ Y = \frac{\sum Y}{N} = \frac{476}{7} = 68.$$

Hence, Karl Pearson's Coefficient of Correlation,

$$r = \frac{\sum xy}{\sqrt{\sum x2.\sum y2}} = \frac{25}{\sqrt{28 \times 34}} = \frac{25}{\sqrt{952}} = \frac{25}{30.85} = 0.81$$

$Since\ r = +0.81$, the variables are highly positively correlated i. e., tall fathers have tall sons.

Example:

Calculate coefficient of correlation from the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Example:

Calculate Pearson's Coefficient of correlation.

9.9 Rank Correlation

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc. The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. It is defined as-

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N}$$

where, ρ (rho) = rank correlation coefficient;

$\sum D^2$ = sum of squares of differences between the pairs of ranks; and

N = number of pairs of observations.

The value of ρ lies between –1 and +1. If ρ = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If ρ = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5^{th} rank, the common rank to be assigned to each item is

$$= \frac{5+6}{2} = 5.5$$

which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is $\frac{1}{12}(m^3 - m)$. A slight formula is used when there is more than one item having the same value. The formula is-

$$\rho = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)]}{N^3 - N}$$

Where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

Example 2:

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price.

Pric	8	9	9	7	6	7	5
e of	8	0	5	0	0	5	0

tea							
Price of coffee	120	134	150	115	110	140	100

Price of tea	Rank	Price of coffee	Rank	D	D^2
88	3	120	4	1	1
90	2	134	3	1	1
95	1	150	1	0	0
70	5	115	5	0	0
60	6	110	6	0	0
75	4	140	2	2	4
50	7	100	7	0	0
					$\sum D^2 = 6$

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 6}{7^3 - 7} = 1 - \frac{36}{336} = 1 - 0.1071 = 0.8929$$

The relation between price of tea and coffee is positive at 0.89.

Based on quality the association between price of tea and price of coffee is highly positive.

Example 3:

In an evaluation of answer script the following marks are awarded by the examiners.

1	8	9	7	9	5	8	7	8

st	8	5	0	60	0	0	5	5
2nd	84	90	88	55	48	85	82	72

Do you agree the evaluation by the two examiners is fair?

Solution:

x	R1	y	R2	D	D^2
88	2	84	4	2	4
95	1	90	1	0	0
70	6	88	2	4	16
60	7	55	7	0	0
50	8	48	8	0	0
75	5	82	5	0	0
80	4	85	3	1	1
85	3	75	6	3	9
					30

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 30}{8^3 - 8} = 1 - \frac{180}{504} = 1 - 0.357 = 0.643$$

ρ = 0.643 shows fair in awarding marks in the sense that uniformity has arisen in evaluating the answer scripts between the two examiners.

Example 4:

Rank Correlation for tied observations. Following are the marks obtained by 10 students in a class in two tests.

Students										

Test 1										
Test 2										

Calculate the rank correlation coefficient between the marks of two tests.

Solution:

Student	Test 1	R1	Test 2	R2	D	D^2
A	70	3	65	5.5	-2.5	6.25
B	68	4	65	5.5	-1.5	2.25
C	67	5	80	1.0	4.0	16.00
D	55	10	60	8.5	1.5	2.25
E	60	8	68	4.0	4.0	16.00
F	60	8	58	10.0	-2.0	4.00
G	75	1	75	2.0	-1.0	1.00

H	63	6	62	7.0	-1.0	1.00
I	60	8	60	8.5	0.5	0.25
J	72	2	70	3.0	-1.0	1.00
						$\sum D^2$ = 50.00

60 is repeated 3 times in test 1. 60, 65 is repeated twice in test 2. m = 3; m = 2; m = 2.

$$\rho = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)]}{N^3 - N}$$

$$\rho = 1 - \frac{6[50 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{10^3 - 10}$$

$$\rho = 1 - \frac{6[50 + 2 + 0.5 + 0.5]}{990} = 1 - \frac{6 \times 53}{990} = 0.68$$

Interpretation: There is uniformity in the performance of students in the two tests.

9.10 Regression

Regression is the functional relationship between two variables and of the two variables one may represent cause and the other may represent effect. The variable representing cause is known as independent variable and is denoted by X. The variable X is also known as predictor variable or repressor. The variable representing effect is known as dependent variable and is denoted by Y. Y is also known as predicted variable. The relationship

between the dependent and the independent variable may be expressed as a function and such functional relationship is termed as regression. When there are only two variables the functional relationship is known as simple regression and if the relation between the two variables is a straight line I is known a simple linear regression. When there are more than two variables and one of the variables is dependent upon others, the functional relationship is known as multiple regression. The regression line is of the form y=a+bx where a is a constant or intercept and b is the regression coefficient or the slope. The values of 'a' and 'b' can be calculated by using the method of least squares. An alternate method of calculating the values of a and b are by using the formula:

The regression equation of y on x is given by $y = a + bx$

The regression coefficient of y on x is given by

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

$$and\ a = \bar{y} - b\bar{x}$$

The regression line indicates the average value of the dependent variable Y associated with a particular value of independent variable X.

9.11 Assumptions

1. The x's are non-random or fixed constants
2. At each fixed value of X the corresponding values of Y have a normal distribution about a mean.
3. For any given x, the variance of Y is same.
4. The values of y observed at different levels of x are completely independent.

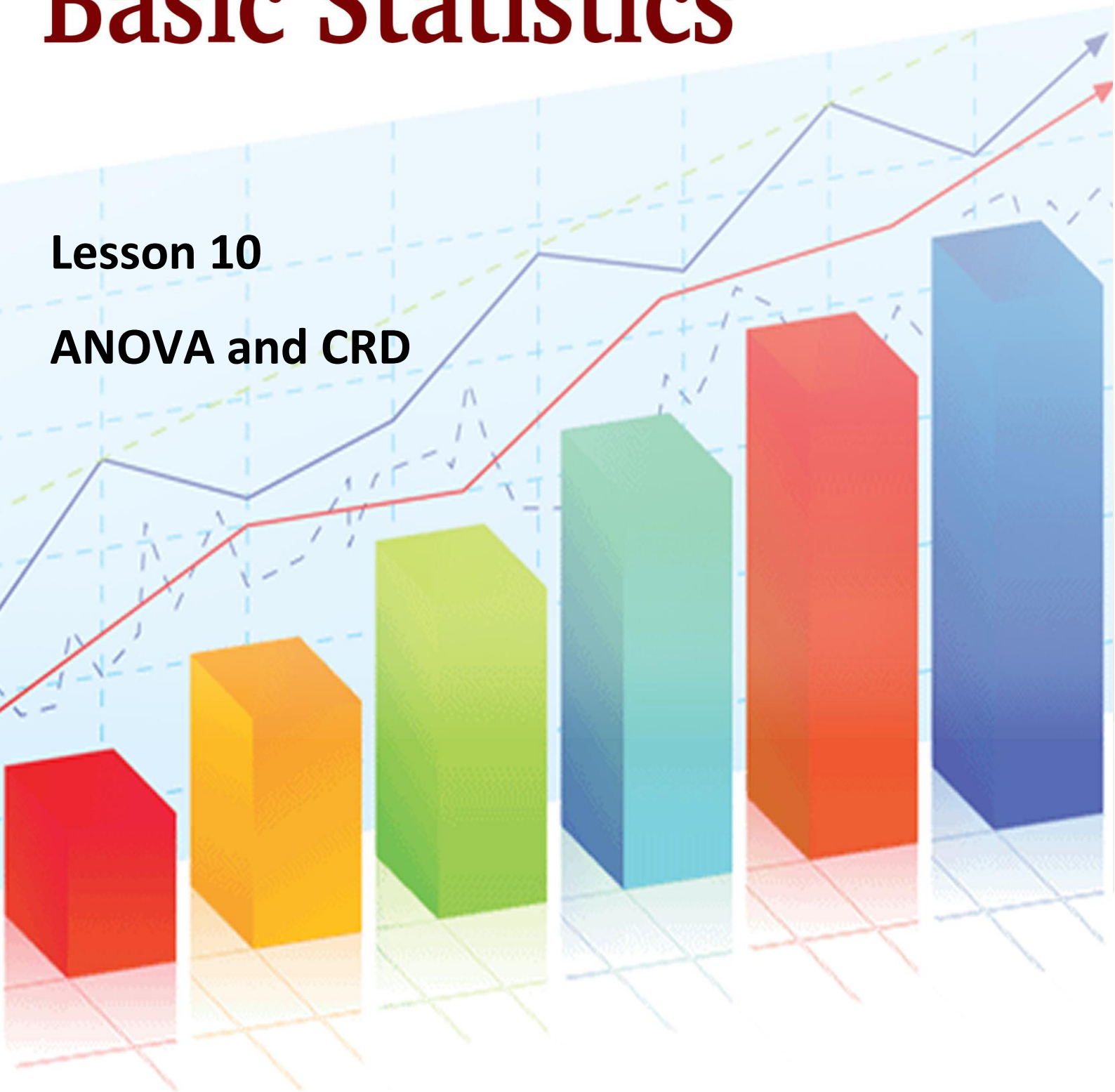
9.12 Properties of Regression coefficients

1. The correlation coefficient is the geometric mean of the two regression coefficients

2. Regression coefficients are independent of change of origin but not of scale.
3. If one regression coefficient is greater than unit, then the other must be less than unit but not vice versa. ie. both the regression coefficients can be less than unity but both cannot be greater than unity, ie. if $b_1>1$ then $b_2<1$ and if $b_2>1$, then $b_1<1$.
4. Also if one regression coefficient is positive the other must be positive (in this case the correlation coefficient is the positive square root of the product of the two regression coefficients) and if one regression coefficient is negative the other must be negative (in this case the correlation coefficient is the negative square root of the product of the two regression coefficients). ie.if $b_1>0$, then $b_2>0$ and if $b_1<0$, then $b_2<0$.
5. If θ is the angle between the two regression lines then it is given by

$$\tan\theta = \frac{(1-r^2)\sigma_x\sigma_y}{r\left(\sigma_x^2+\sigma_y^2\right)}$$

Basic Statistics



Lesson 10

ANOVA and CRD

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 10	**ANOVA and CRD**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Lesson-10

Objectives of the lesson:

1. Assumption of the ANOVA
2. One way Classification of ANOVA
3. Two way Classification of ANOVA

Glossary of the lesson: Test-Statistic, Variance, ANOVA, Source of variation etc.

10.1 Introduction:

In hypothesis testing, we test the significance of difference between two sample means. For this, one test statistic employed was the t-test where we assumed that the two populations from which the samples were drawn had the same variance. But in real life, there may be situations when instead of comparing two sample means, a researcher has to compare three or more than three sample means (specifically, more than two). A researcher may have to test whether the three or more sample means computed from the three populations are equal. In other words, the null hypothesis can be that three or more population means are equal as against the alternative hypothesis that these population means are not equal. For example, suppose that a researcher wants to measure work attitude of the employees in four organizations. The researcher has prepared a questionnaire consisting of 10 questions for measuring the work attitude of employees. A five-point rating scale is used with 1 being the lowest score and 5 being the highest score. So, an employee can score 10 as the minimum score and 50 as the maximum score. The null hypothesis can be set as all the means are equal (i.e., there is no difference in the degree of work attitude of the employees) as against the alternative hypothesis that at least one of the means is different from the others (there is a significant difference in the degree of work attitude of the employees). In this situation, analysis of variance technique is used. "In its simplest form analysis of variance may be regarded as an extension or development of the $t-test$." Analysis of variance technique makes use of

F distribution (F-statistic). Some more examples are presented below where we are required to test the equality of means of three or more populations. For example, whether:

(1) The average life of light bulbs being produced in three different plants is the same.
(2) All the three varieties of fertilizers have the same impact on the yield of rice.
(3) The level of satisfaction among the participants in all IIMs is the same.
(4) The impact of training on salesmen trained in three institutes is the same.
(5) The service time of a transaction is the same on four different counters in a service unit.
(6) The average price of different commodities in four different retail outlets is the same.
(7) Performance of salesmen in four zones is the same.

The word 'analysis of variance' is used since the technique involves first finding out the total variation among the observations in the collected data, then assigning causes or components of variation to various factors and finally drawing conclusions about the equality of means. Thus Analysis of variance or ANOVA can be defined as a technique of testing hypotheses about the significant difference in several population means. This statistical technique was developed by R.A.Fisher. The main purpose of analysis of variance is to detect the difference among various population means based on the information gathered from the samples (sample means) of the respective populations.

10.2 Assumptions of Analysis of Variance

Analysis of variance is based on some assumptions.

(i) Each sample is a simple random sample.

(ii) Populations from which the samples are selected are normally distributed. However, in case of large samples this assumption is not required.

(iii) Samples are independent.

(iv) The population variances are identical, i.e.,$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$

10.3 Computation of Test Statistic:

The beauty of the technique of analysis of variance is that it performs the test of equality of more than two population means by actually analyzing the variance. In simple terms, ANOVA decomposes the total variation into two components of variation, namely, variation between the samples known as the mean square between samples and variation within the samples known as the mean square within samples. The variance ratio denoted by F is given by:

$$F = \frac{Mean\ square\ between\ samples\ or\ groups}{Mean\ square\ within\ samples\ or\ groups}$$

If the calculated value of F is greater than the critical value of F, we must reject the null hypothesis. In case the calculated value of F is less than the critical value of F, we are to retain or accept the null hypothesis.

10.3.1 Analysis of variance table (ANOVA):

The table showing the sources of variation, the sum of squares, degrees of freedom, mean squares and the formula for the F statistic is known as ANOVA table.

10.4 Classification of Analysis of Variance

ANOVA is mainly carried on under the following two classifications:

(i) One-way classification

(ii) Two- way classification

Variance and its different components may be obtained in each of the two types of classification by:

(a) Direct Method, (b) Short-cut Method

10.4.1 One-Way Classification

Many business applications involve experiments in which different populations (or groups) are classified with respect to only one attribute of interest such as (i) percentage of marks secured by students in a course, (ii) flavor preference of ice-cream by customers, (iii) yield of crop due to varieties of seeds, and so on. In all such cases observations in the sample data are classified into several groups based on a single attribute (i.e., criterion) and are termed one-way classification of sample data.

Under this one-way classification we set up the null hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$where $\mu_1, \mu_2, \ldots, \mu_k$are the arithmetic means of the populations from which the k samples are drawn.

The alternative hypothesis is $H_0: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k$

After formulation of the null hypothesis and alternative hypothesis one needs to calculate the following by using any one of the above mentioned methods.

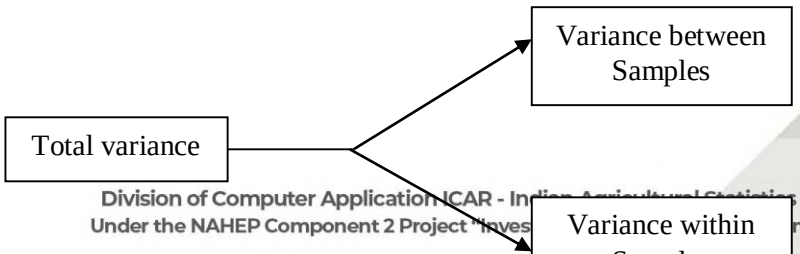
(i)Variance between the samples.

(ii)Variance within the samples.

(iii)Total variance by summing (i) and (ii) which may also be calculated directly for verification of calculations.

(iv)F-ratio (or F-statistic)

Thus in this process, the total variance can be divided into two additive and independent parts as shown:



After calculating the test statistic F, it should be compared with the critical value of F at a specified level of significance α for $(k-1, n-k)$ degrees of freedom and on the basis of this comparison; accordingly the decision to accept or reject the null hypothesis is taken.

(a) Direct Method

I. Calculation of variance between samples

It is the sum of squares of the deviations of the means of various samples from the grand mean. The procedure of calculating the variance between the samples is as shown below:

Observations		Number of samples			
	1	2	……j	……	k
1	x_{11}	x_{12}	… x_{1j}	……	x_{1k}
2	x_{21}	x_{22}	… x_{2j}	……	x_{2k}
.					
.	.	.	.	.	.
i	x_{i1}	x_{i2}	… x_{ij}	……	x_{ik}
.					
.					
N	x_{n1}	x_{n2}	… x_{nj}	……	x_{nk}
Total	T_1	T_2	… T_j	……	T_k
A.M.	$\bar{x}_1$	$\bar{x}_2$	… $\bar{x}_j$	……	$\bar{x}_k$

$T_j = \sum_{i=1}^{n} x_{ij}$, $T = \sum_{j=1}^{k} T_j$, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{k} T_j$ and $\bar{\bar{x}} = \frac{1}{nk}\sum_{j=1}^{k} \bar{x}_j$

The steps in calculating variance between samples are given by:

(i) In the first step, we need to calculate the mean of each sample i.e., $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$f all the k samples.

(ii) Next, the grand mean is calculated by using the formula

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} = \frac{T}{K}$$

where T =Grand total of all observations.

N =Total number of observations in all k samples.

(iii) In step 3, the difference between the mean of each sample and grand mean is calculated, that is we calculate $\bar{x}_1 - \bar{\bar{x}}, \bar{x}_2 - \bar{\bar{x}}, \ldots, \bar{x}_k - \bar{\bar{x}}$

(iii) In step 4, we square the deviations obtained in step (iii) and multiply by the number of items in the corresponding sample and then add the total. This total gives the sum of the squares of the deviations between the samples (or between the columns) and it is denoted by SSB or SSC

$$Thus\ SSB = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

where k is the number of groups or samples being compared, n_j the number of observations in group j, $\bar{x}_j$the sample mean of group j, and $\bar{\bar{x}}$ the grand mean.

(iv) In the last step, the total obtained in step (iv) is divided by the degrees of freedom. The degrees of freedom is one less than the total number of samples. If there are k samples, the degrees of freedom will be $v = k - 1$ When the sum of squares obtained in step (iv) is divided by the number of degrees of freedom, the result is called mean square which is denoted by MSB or MSC. MSB indicates the degree of explained variance due to sampling fluctuations.

$$Thus\ MSB = \frac{SSB}{k-1}$$

II. Calculation of Variance within the samples:

This is usually referred to as the sum of squares within samples. The variance within samples measures the difference within the samples due to chance. This is usually denoted by SSE. The steps involved are the following:

(i) The first step is to calculate the mean values $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ of all k samples,

(ii) Calculate the deviations of the various observations of k samples from the mean values of the respective samples,

(iii) Square all the deviations obtained in (ii) and find the total of these squared deviations. This total gives the sum of the squares of deviations within the samples or the sum of squares due to error. It is denoted by SSW or SSE.

(v) $Thus\ SSW\ or\ SSE = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \bar{x}_j)$

where x_{ij} is the i[th] observation in the j[th] group ,$\bar{x}_j$ the sample mean of j[th] group, k the number of groups being compared, and N the total number of observations in all the groups.

(iv) As the last step, divide the total squared deviations obtained in step (iii) by the degrees of freedom and obtain the mean square. The number of degrees of freedom can be calculated as the difference between the total number of observations and the number of samples. If there are N observations and k samples then the degrees of freedom is $\nu = N - k$.

$$Thus\ MSW = \frac{SSE}{N-k}$$

III. Calculate total sum of squares: The total variation is equal to the sum of the squared difference between each observation (sample value) and the grand mean $\bar{\bar{x}}$ and is often referred to as SST.

Thus SST can be calculated as:

$$SST = SSB + SSE$$

III. Calculation of the Test Statistic F:

When the null hypothesis is true, both mean squares MSB (MSC) and MSW (MSE) are the independent unbiased estimates of the same population variance σ^2. Hence, the test statistic is

$$F = \frac{MSB}{MSW} \; or \; F = \frac{MSC}{MSE}$$

which follows F-distribution with degrees of freedom $(k-1, N-k)$

F is the ratio between the greater variance to the smaller variance. Generally, variance between the samples (MSB or MSC) is greater than the variance within the samples (MSW or MSE). But if MSW>MSB then the reverse ratio of F will be used i.e.,

$$F = \frac{MSW}{MSB} \; or, F = \frac{MSE}{MSC}$$

IV. Conclusion:

To compare the calculated value of F with the critical value (tabulated) of F for $(k-1, N-k)$ degrees of freedom at the specified level of significance, usually 5% or 1% level. If the calculated value is greater than the tabulated value, the null hypothesis H_0 is rejected and conclude that all population means are not equal. Otherwise, the null hypothesis is accepted.

For analyzing variance in case of one way classification the following table known as the Analysis of Variance Table (or ANOVA Table) is constructed.

ANOVA Table

Sources of Variation	Sum of squares (SS)	Degrees of Freedom (df)	Mean Squares(MS)	Test Statistic
Between samples	SSB	$k-1$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within samples	SSW	$N-k$	$MSW = \frac{SSW}{N-k}$	
Total	SST	$N-1$		

(b) Short-cut Method: The calculation of F-statistic (variance ratio) by using the direct Method is very time consuming. In practice, a short-cut Method based on sum of squares of the individual values (observations) is usually

used. The computational work is much minimized in this method. The method is more convenient when some or all the sample means and the grand mean are fractional. The various steps involved in the calculations of variance ratio are the following:

(i) Calculate the grand total of all observations in samples, T

$$T = \sum x_1 + \sum x_2 + \cdots + \sum x_k$$

(ii) Calculate the correction factor $CF = \frac{T^2}{N}$; N= Total observations in samples.

(iv) Find the sum of the squares of all observations in samples from each of k samples and subtract CF from this sum to obtain the total sum of the squares of deviations SST:

$$SST = (\sum x_1^2 + \sum x_2^2 + \ \ldots + \sum x_k^2 - CF$$

$SSB = \frac{(\sum x_j)^2}{n_j} - CF$, where n_j is the size of the j[th] sample.

And $SSE = SST - SSB$

(v) MSB, MSW and F are obtained as in the Direct method and decision either to accept or to reject is exactly same as taken in case of Direct method.

Example 1: The following data give the yield on 12 plots of land in three samples of 4 plots each, under three varieties of fertilizers A, B and C

A	B	C
25	20	24
22	17	26
24	16	30
21	16	20

Test whether there is any significant difference in the average yields of land under three varieties of fertilizers.

Solution: Here first we set up null and alternative hypotheses

Null hypothesis H_0 :There is no significant difference in the average yields under the three varieties.

Alternative hypothesis H_1 :There is significant difference in the average yield under the three varieties.

We calculate sample means, the variance between the samples and the variance within the samples by using the Direct Method.

	Sample I X_1	Sample II X_2	Sample III X_3
	25	20	24
	22	17	26
	24	16	30
	21	16	20
Total	92	72	100

$$\bar{x}_1 = \frac{\sum X_i}{n_1} = \frac{92}{4} = 23, \bar{X}_2 = \frac{\sum X_2}{n_2} = \frac{72}{4} = 18, \bar{x}_3 = \frac{\sum X_3}{n_3} = \frac{100}{4} = 25$$

$$Grand\ mean\ (\bar{\bar{X}}) = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3}{3} = \frac{23 + 18 + 25}{3} = 22$$

$$SSB = n_1(X_1 - \bar{\bar{X}})^2 + n_2(X_2 - \bar{\bar{X}})^2 + n_3(X_3 - \bar{\bar{X}})^2$$
$$= 4(23 - 22)^2 + 4(8 - 22)^2 + 4(25 - 22)^2$$
$$= 104$$

Degrees of freedom, $\nu_1 = k - 1 = 2$

Now MSW=Mean square between the samples

$$= \frac{SSB}{\nu_1} = \frac{104}{12} = 52$$

Calculation for SSW

Sample I		Sample II		Sample III	
X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$	X_3	$(X_3 - \bar{X}_3)^2$
25	4	20	4	24	1
22	1	17	1	26	1
24	1	16	4	30	25

21	4	19	1	20	25
	10		10		52

SSW = Sum of squares within the samples

$$= \sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2 + \sum(X_3 - \bar{X}_3)^2$$

$$= 10 + 10 + 52 = 72$$

The degrees of freedom, $\nu_2 = N - k = 9$

$$MSW = \frac{72}{9} = 8$$

Now we prepare the Analysis of Variance Table

Analysis of Variance Table (ANOVA)

Sources of Variation	Sum of squares (SS)	Degrees of Freedom (df)	Mean Squares(MS)	Test Statistic
Between samples	104	2	52	F=52/8=6.5
Within samples	72	9	8	
Total	SST=176	N-1=11		

The critical value of F for $\nu_1 = 2$ and $\nu_2 = 9$ at 5% level i.e., $F_{0.05}(2,9)$ is 4.26.

Since the calculated value of F (i.e., 6.5) is greater than the critical value $F_{0.05}(2,9)$, herefore we reject the null hypothesis at 5% level and conclude that the difference in the average yields under the three varieties is significant.

Solution of the above problem by Short cut Method:

After formulating the null and alternative hypotheses as in the Direct Method one can proceed with the Short cut Method as follows:

Calculation of T, SST, SSB, SSW

Sample	Sample I	Sample I

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
25	625	20	400	24	576
22	484	17	289	26	676
24	576	16	256	30	900
21	441	19	361	20	400
92	2126	72	1306	100	2552

T = Sum of all the values in the three samples

$$= \sum X_1 + \sum X_2 + \sum X_3 = 92 + 72 + 100 = 264$$

$$Correction\ factor\ = \frac{T^2}{N} = \frac{(264)^2}{12} = 5808$$

SST = Total sum of squares

$$= (\sum X_1^2 + \sum X_2^2 + \sum X_3^2) - \frac{T^2}{N}$$

$$(2126 + 1306 + 2552) - 5808 = 176$$

SSB = Sum of squares between the samples

$$\frac{\sum X_1^2}{n_1} + \frac{\sum X_2^2}{n_2} + \frac{\sum X_3^2}{n_3} - \frac{T^2}{N}$$

$$= 2116 + 1296 + 2500 - 5808$$

$$= 104$$

$Degrees\ of\ freedom\ = k - 1 = 2$

$SSW\ = Sum\ of\ squares\ within\ the\ samples.$

$= SST - SSB\ = 176 - 104\ = 72$

$Degrees\ of\ freedom\ = N - k = 12 - 3 = 9$

Now proceeding with the ANOVA table as in the Direct Method, one can arrive at the same conclusion

10.5 Design of Experiments

Choice of treatments, method of assigning treatments to experimental units and arrangement of experimental units in different patterns are known as designing an experiment. We study the effect of changes in one variable on another variable. For example how the application of various

doses of fertilizer affects the grain yield. Variable whose change we wish to study is known as response variable. Variable whose effect on the response variable we wish to study is known as factor.

Treatment: Objects of comparison in an experiment are defined as treatments. Examples are Varieties tried in a trail and different chemicals.

Experimental unit: The object to which treatments are applied or basic objects on which the experiment is conducted is known as experimental unit.

Example: piece of land, an animal, etc

Experimental error: Response from all experimental units receiving the same treatment may not be same even under similar conditions. These variations in responses may be due to various reasons. Other factors like heterogeneity of soil, climatic factors and genetic differences, etc also may cause variations (known as extraneous factors). The variations in response caused by extraneous factors are known as experimental error.

Our aim of designing an experiment will be to minimize the experimental error.

10.6 Basic principles

To reduce the experimental error we adopt certain principles known as basic principles of experimental design. The basic principles are

1) Replication,
2) Randomization and
3) Local control

10.6.1 Replication

Repeated application of the treatments is known as replication. When the treatment is applied only once we have no means of knowing about the variation in the results of a treatment. Only when we repeat several times we can estimate the experimental error.

With the help of experimental error we can determine whether the obtained differences between treatment means are real or not. When the number of replications is increased, experimental error reduces.

10.6.2 Randomization

When all the treatments have equal chance of being allocated to different experimental units it is known as randomization.

If our conclusions are to be valid, treatment means and differences among treatment means should be estimated without any bias. For this purpose we use the technique of randomization.

10.6.3 Local Control

Experimental error is based on the variations from experimental unit to experimental unit. This suggests that if we group the homogenous experimental units into blocks, the experimental error will be reduced considerably. Grouping of homogenous experimental units into blocks is known as local control of error.

In order to have valid estimate of experimental error the principles of replication and randomization are used.

In order to reduce the experimental error, the principles of replication and local control are used.

In general to have precise, valid and accurate result we adopt the basic principles.

10.7 Completely Randomized Design (CRD)

CRD is the basic single factor design. In this design the treatments are assigned completely at random so that each experimental unit has the same chance of receiving any one treatment. But CRD is appropriate only when the experimental material is homogeneous. As there is generally large variation among experimental plots due to many factors CRD is not preferred in field experiments.

In laboratory experiments and greenhouse studies it is easy to achieve homogeneity of experimental materials and therefore CRD is most useful in such experiments.

10.7.1 Layout of a CRD

Completely randomized Design is the one in which all the experimental units are taken in a single group which are homogeneous as far as possible. The randomization procedure for allotting the treatments to various units will be as follows.

Step 1: Determine the total number of experimental units.

Step 2: Assign a plot number to each of the experimental units starting from left to right for all rows.

Step 3: Assign the treatments to the experimental units by using random numbers.

The statistical model for CRD with one observation per unit

$$Y_{ij} = \mu + t_i + e_{ij}$$

μ = overall mean effect

t_i = true effect of the ith treatment

e_{ij} = error term of the jth unit receiving ith treatment

The arrangement of data in CRD is as follows:

Treatments

	T_1	T_2	T_i	T_K	
	y_{11}	y_{21}	y_{i1}	Y_{K1}	
	y_{12}	y_{22}	y_{i2}	Y_{K2}	
	$y_{1r}1$	y_{2r2}	y_{iri}	Y_k	
Total	Y_1	Y_2	Y_i	T_k	GT

(GT – Grand total)

The null hypothesis will be

$H_o : \mu_1 = \mu_2 = \cdots = \mu_k$ or There is no significant difference between the treatments

And the alternative hypothesis is

$H\_1: \mu_1 \neq \mu_2 \neq \ldots\ldots\ldots\ldots \neq \mu_k$. There is significant difference between the treatments

The different steps in forming the analysis of variance table for a CRD are:

$$C.F. = \frac{(GT)^2}{n}, where\ n = Total\ number\ of\ observations$$

$$Total\ SS = TSS = \sum_{i=1}^{k}\sum_{j=1}^{v} y_{ij}^2 - C.F.$$

$$Treatment\ Ss = TrSS = \frac{Y_1^2}{r_1} + \frac{Y_2^2}{r_2} + \ \ldots + \frac{Y_k^2}{r_k} - C.F.$$

$$= \sum_{i=1}^{k} \frac{Y_i^2}{r_i} - C.F.$$

$$Error\ SS = ESS = \sum_{i=1}^{k}\sum_{j=1}^{v} y_{ij}^2 - \sum_{i=1}^{k} \frac{Y_i^2}{r_i}$$

$$= TSS - TrSS$$

Form the following ANOVA table and calculate F value

Source of variation	d.f.	SS	MS	F
Treatment	t-1	TrSS	$TrMS = \frac{TrSS}{t-1}$	$\frac{TrMS}{EMS}$
Error	n-t	ESS	$EMS = \frac{ESS}{n-t}$	
Total	n-1			

6. Compare the calculated F with the critical value of F corresponding to treatment degrees of freedom and error degrees of freedom so that acceptance or rejection of the null hypothesis can be determined.

7. If null hypothesis is rejected that indicates there is significant differences between the different treatments.

8. Calculate C D value.

$$C.D. = SE(d) \times t, where\ S.E.(d) = \sqrt{EMS\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$$

r_i = number of replications for treatment i
r_j = number of replications for treatment j and
t is the critical t value for error degrees of freedom at specified level of significance, either 5% or 1%.

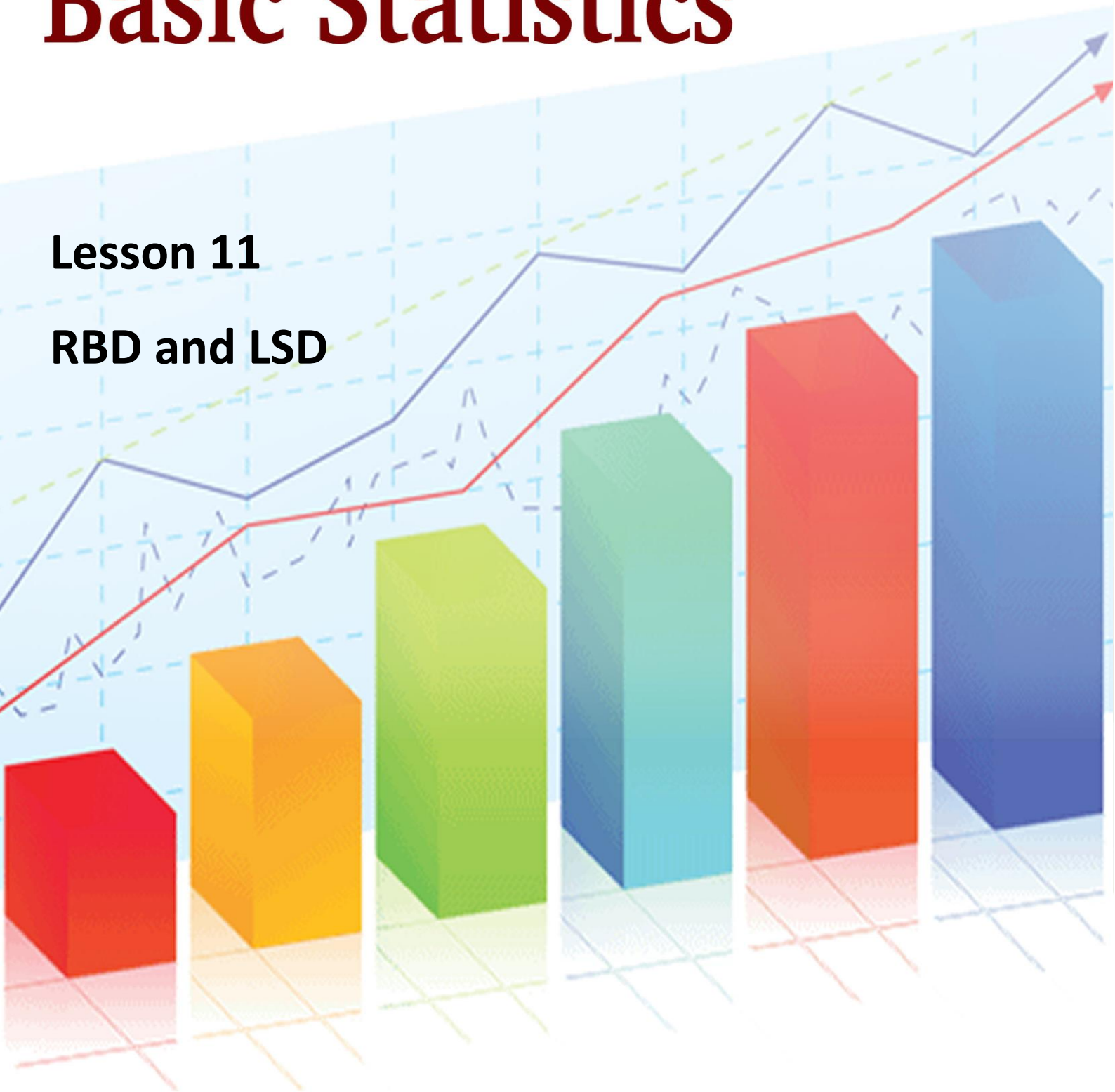
10.7.2 Advantages of a CRD

1. Its layout is very easy.
2. There is complete flexibility in this design i.e. any number of treatments and replications for each treatment can be tried.
3. Whole experimental material can be utilized in this design.
4. This design yields maximum degrees of freedom for experimental error.
5. The analysis of data is simplest as compared to any other design.
6. Even if some values are missing the analysis can be done.

10.7.3 Disadvantages of a CRD

1. It is difficult to find homogeneous experimental units in all respects and hence CRD is seldom suitable for field experiments as compared to other experimental designs.
2. It is less accurate than other designs.

Basic Statistics



Lesson 11

RBD and LSD

Content

DESIGNED AND DEVELOPED UNDER THE AEGIS OF
NAHEP Component-2 Project "Investments In ICAR Leadership In Agricultural Higher Education"
Division of Computer Applications

Course Name	**Basic Statistics**
Lesson 11	**RBD and LSD**
Content Creator Name	**Dr. Vinay Kumar**
University/College Name	**Chaudhary Charan Singh Haryana Agricultural University,Hisar**
Course Reviewer Name	**Dr Dhaneshkumar V Patel**
University/college Name	**Unagadh Agricultural University,Junagadh**

Lesson-11

Objectives of the lesson:

1. Layout of RBD
2. Analysis, Merits, Demerits of RBD
3. Layout of LSD
4. Analysis, Merits, Demerits of LSD

Glossary of the lesson: RBD, LSD, Variance, ANOVA, Source of variation etc.

11.1 Randomized Blocks Design (RBD)

When the experimental material is heterogeneous, the experimental material is grouped into homogenous sub-groups called blocks. As each block consists of the entire set of treatments a block is equivalent to a replication.

If the fertility gradient runs in one direction say from north to south or east to west then the blocks are formed in the opposite direction. Such an arrangement of grouping the heterogeneous units into homogenous blocks is known as randomized blocks design. Each block consists of as many experimental units as the number of treatments. The treatments are allocated randomly to the experimental units within each block independently such that each treatment occurs once. The number of blocks is chosen to be equal to the number of replications for the treatments.

The analysis of variance model for RBD is

$$Y_{ij} = \mu + t_i + r_j + e_{ij}$$

where

$\mu = the\ overall\ mean$

$t_i = the\ ith\ treatment\ effect$

$r_j = the\ jth\ replication\ effect$

$e_{ij} = the\ error\ term\ for\ i^{th}\ treatment\ and\ j^{th}\ replication$

11.2 Analysis of RBD

The results of RBD can be arranged in a two way table according to the replications (blocks) and treatments. There will be r x t observations in total where r stands for number of replications and t for number of treatments. The data are arranged in a two way table form by representing treatments in rows and replications in columns.

Treatment	Replication					Total
	1	2	3	.....	r	
1	y_{11}	y_{12}	y_{13}	.....	y_{1r}	T_1
2	y_{21}	y_{22}	y_{23}	.....	y_{2r}	T_2
3	y_{31}	y_{32}	y_{33}	.....	y_{3r}	T_3
t	y_{t1}	y_{t2}	y_{t3}	.....	y_{tr}	T_t
Total	R_1	R_2	R_3		R_r	G.T

In this design the total variance is divided into three sources of variation viz., between replications, between treatments and error

$$CF = \frac{(GT)^2}{n}$$

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

$$Replication\ SS = RSS = = \Sigma R_j^2 - CF$$

$$Treatments\ SS = TrSS = \Sigma T_i^2 - CF$$

$$Error\ SS = ESS = Total\ SS - Replication\ SS - Treatment\ SS$$

The skeleton ANOVA table for RBD with t treatments and r replications

Sources of variation	d.f.	SS	MS	F- Value
Replication	r-1	RSS	RMS	$\frac{RMS}{EMS}$
Treatment	t-1	TrSS	TrMS	$\frac{TrMS}{EMS}$
Error	(r-1) (t-1)	ESS	EMS	
Total	rt-1	TSS		

$$CD = SE(d).t \; where \; S.E(d) = \sqrt{\frac{2EMS}{r}}$$

t = critical value of t for a specified level of significance and error degrees of freedom

Based on the CD value various treatment means can be compared

11.2.1 Advantages of RBD

The precision is more in RBD. The amount of information obtained in RBD is more as compared to CRD. RBD is more flexible. Statistical analysis is simple and easy. Even if some values are missing, still the analysis can be done by using missing plot technique.

11.2.2 Disadvantages of RBD

When the number of treatments is increased, the block size will increase. If the block size is large maintaining homogeneity is difficult and hence when more number of treatments is present this design may not be suitable.

11.3 Latin Square Design

When the experimental material is divided into rows and columns and the treatments are allocated such that each treatment occurs only once in each row and each column, the design is known as L S D.

In LSD the treatments are usually denoted by A B C D etc.

For a 5 x 5 LSD the arrangements may be

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 1

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

Square 2

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

Square 3

11.3.1 Statistical Analysis:

The ANOVA model for LSD is

$$Y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk}$$

r_i is the ith row effect

c_j is the jth column effect

t_k is the kth treatment effect and

e_{ijk} is the error term

The analysis of variance table for LSD is as follows:

Sources of Variation	**d.f.**	**SS**	**MS**	**F**
Rows	t-1	RSS	RMS	$\frac{RMS}{EMS}$
Columns	t-1	CSS	CMS	$\frac{CMS}{EMS}$
Treatments	t-1	TrSS	TrMS	$\frac{TrMS}{EMS}$
Error	(t-1)(t-2)	ESS	EMS	
Total	t2-1	TSS		

F table value

$F_{[(t-1),(t-1)(t-2)]}$ degrees of freedom at 5% or 1% level of significance

Steps to calculate the above Sum of Squares are as follows:

$$Correction\ Factor\ (CF) = \frac{(GT)^2}{(t)^2}$$

$$Total\ Sum\ of\ Squares\ (TSS) = \sum(y_{ijk})^2 - CF$$

$$Row\ sum\ of\ squares\ (RSS) = \frac{1}{t}\sum_{i=1}^{t}(R_i)^2 - CF$$

$$Column\ sum\ of\ squares(CSS) = \frac{1}{t}\sum_{j=1}^{t}\left(C_j\right)^2 - CF$$

$$Treatment\ sum\ of\ squares\ (TrSS = \frac{1}{t}\sum_{k=1}^{t}(T_k)^2 - CF$$

$$Error\ Sum\ of\ Squares\ = \ TSS - RSS - CSS - TrSS$$

These results can be summarized in the form of analysis of variance table.

Calculation of SE, SE (d) and CD values

$$SE = \frac{\sqrt{EMS}}{r}$$

where r is the number of rows

$$SE(d) = \sqrt{2} \times SE$$

$$CD = \ SE\ (d) \times \ t$$

where t = table value of t for a specified level of significance and error degrees of

freedom

Using CD value the bar chart can be drawn and the conclusion may be written.

11.3.2 Advantages

- LSD is more efficient than RBD or CRD. This is because of double grouping that will result in small experimental error.
- When missing values are present, missing plot technique can be used and analysed.

11.3.3 Disadvantages

- This design is not as flexible as RBD or CRD as the number of treatments is limited to the number of rows and columns. LSD is seldom used when the number of treatments is more than 12. LSD is not suitable for treatments less than five.

Because of the limitations on the number of treatments, LSD is not widely used in agricultural experiments.

Note: The number of sources of variation is two for CRD, three for RBD and four for LSD.