UNIT 1 INTRODUCTION TO STATISTICS

Structure

1.0 INTRODUCTION

The word statistics has different meaning to different persons. Knowledge of statistics is applicable in day to day life in different ways. In daily life it means general calculation of items, in railway statistics means the number of trains operating, number of passenger's freight etc. and so on. Thus statistics is used by people to take decision about the problems on the basis of different type of *quantitative and qualitative* information available to them.

However, in behavioural sciences, the word 'statistics' means something different from the common concern of it. Prime function of statistic is to draw statistical inference about population on the basis of available quantitative information. Overall, statistical methods deal with reduction of data to convenient descriptive terms and drawing some inferences from them. This unit focuses on the above aspects of statistics.

1.1 OBJECTIVES

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

- Define the term statistics;
- Explain the status of statistics;
- Describe the nature of statistics;
- State basic concepts used in statistics; and
- Analyse the uses and misuses of statistics.

1.2 MEANING OF STATISTICS

The word statistics has been derived from Latin word 'status' or Italian 'Statista' meaning statesman. Professor Gott Fried Achenwall used it in the 18th century. During early period, these words were used for political state of the region. The word 'Statista' was used to keep the records of census or data related to wealth of a state. Gradually, its meaning and usage extended and thereonwards its nature also changed.

The word statistics is used to convey different meanings in singular and plural sense. Therefore it can be defined in two different ways.

1.2.1 Statistics in Singular Sense

In singular sense, 'Statistics' refers to what is called statistical methods. It deals with the collection of data, their classification, analysis and interpretations of statistical data. Therefore, it is described as a branch of science which deals with classification, tabulation and analysis of numerical facts and make decision as well. Every statistical inquiry should pass through these stages.

1.2.2 Statistics in Plural Sense

'Statistics' used in plural sense means that quantitative information is available called 'data'. For example, information on population or demographic features, enrolment of students in Psychology programmes of IGNOU, and the like. According to Websters "Statistics are the classified facts representing the conditions of the people in a State specifically those facts which can be stated in number or in tables of number or classified arrangement".

Horace Secrist describes statistics in plural sense as follows : " By Statistics we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standard of accuracy , collected in a systematic manner for a pre-determined purpose and placed in relation to each other." Thus Secrist's definition highlights following features of statistics:

i) *Statistics are aggregate of facts:* Single or unrelated items are not considered as statistics.

ii) *Statistics are affected by multiplicity of causes:* In statistics the collected information are greatly influenced by a number of factors and forces working together.

iii) *Statistics are numerical facts:* Only numerical data constitute statistics.

iv) *Statistics are enumerated or estimated with a reasonable standard of accuracy:* While enumerating or estimating data, a reasonable degree of accuracy must be achieved.

v) *Statistics are collected in a systematic manner:* Data should be collected by proper planning by utilising tool/s developed by trained personnel.

vi) *Statistics are collected for a predetermined purpose :* It is necessary to define the objective of enquiry, before collecting the statistics. The objective of enquiry must be specific and well defined.

vii) *Statistics should be comparable:* Only comparable data will have some meaning. For statistical analysis, the data should be comparable with respect to time, place group, etc.

Thus, it may be stated that " All statistics are numerical statements of facts but all numerical statements of facts are not necessarily statistics ".

1.2.3 Definition of Statistics

In this unit emphasis is on the term statistics as a branch of science. It deals with classification, tabulation and analysis of numerical facts. Different statistician defined this aspect of statistics in different ways. For example.

A. L. Bowley gave several definitions of Statistics:

i) "Statistics may be called the science of counting" . This definition emphasises enumeration aspect only.

ii) In another definition he describes it as " Statistics may rightly be called the science of average".

iii) At another place Statistics is defined as, "Statistics is the science of measurement of social organism regarded as a whole in all its manifestations".

All three definitions given by Bowely seem to be inadequate because these do not include all aspects of statistics.

According to **Selligman** "Statistics is the science which deals with the methods of collecting, classifying, presenting , comparing and interpreting numerical data collected to throw some light on any sphere of enquiry".

Croxton and Cowden defined "statistics as the collection , presentation, analysis ,and interpretation of numerical data".

Among all the definitions , the one given by Croxton and Cowden is considered to be most appropriate as it covers all aspects and field of statistics.

These aspects are given below:

Collection of Data : Once the nature of study is decided , it becomes essential to collect information in form of data about the issues of the study. Therefore, the collection of data is the first basic step. Data may be collected either from primary source or secondary or from both the sources depending upon the objective/s of the investigation

Classification and Presentation : Once data are collected , researcher has to arrange them in a format from which they would be able to draw some conclusions. The arrangement of data in groups according to some similarities is known as classification.

Tabulation is the process of presenting the classified data in the form of table. A tabular presentation of data becomes more intelligible and fit for further statistical analysis. Classified and Tabulated data can be presented in diagrams and graphs to facilitate the understanding of various trends as well as the process of comparison of various situations.

Analysis of Data : It is the most important step in any statistical enquiry . Statistical analysis is carried out to process the observed data and transform it in such a manner as to make it suitable for decision making.

Interpretation of Data : After analysing the data, researcher gets information partly or wholly about the population. Explanation of such information is more useful in real life. The quality of interpretation depends more and more on the experience and insight of the researcher.

Self Assessment Questions

1) Complete the following statements
   i) The word statistics has been derived from Latin word .....
   ii) Statistics in plural means .....
   iii) Statistics in singular means .....
   iv) The first step in statistics is .....
   v) The last step in statistics is .....
2) Tick (√) the correct answer

   Statistical data are:
   i) Aggregates of facts
   ii) Unsystematic data
   iii) Single or isolated facts or figure
   iv) None of these
3) Which one of the following statement is true for statistics in singular sense?
   i) Statistics are aggregate of facts.
   ii) Statistics are numerical facts.
   iii) Statistics are collected in a systematic manner.
   iv) Statistics may be called the science of counting.

1.3 TYPES OF STATISTICS

After knowing the concept and definition of statistics, let us know the various types of statistics.

Though various bases have been adopted to classify statistics, following are the two major ways of classifying statistics: (i) on the basis of function and (ii) on the basis of distribution.

1.3.1 On the Basis of Functions

As statistics has some particular procedures to deal with its subject matter or data, three types of statistics have been described.

A) **Descriptive statistics:** The branch which deals with descriptions of obtained data is known as descriptive statistics. On the basis of these descriptions a particular group of population is defined for corresponding characteristics. The descriptive statistics include classification, tabulation measures of central tendency and variability. These measures enable the researchers to know about the tendency of data or the scores, which further enhance the ease in description of the phenomena.

B) **Correlational statistics:** The obtained data are disclosed for their inter correlations in this type of statistics. It includes various types of techniques to compute the correlations among data. Correlational statistics also provide description about sample or population for their further analyses to explore the significance of their differences.

C) **Inferential statistics:** Inferential statistics deals with the drawing of conclusions about large group of individuals (population) on the basis of observations of few participants from them or about the events which are yet to occur on the basis of past events. It provide tools to compute the probabilities of future behaviour of the subjects.

1.3.2 On the Basis of Distribution of Data

Parametric and nonparametric statistics are the two classifications on the basis of distribution of data. Both are also concerned to population or sample. By population we mean the total number of items in a sphere. In general it has infinite number therein but in statistics there is a finite number of a population, like the number of students in a college. According to Kerlinger (1968) "the term population and universe mean all the members of any well-defined class of people, events or objects." In a broad sense, statistical population may have three kinds of properties – (a) containing finite number of items and knowable, (b) having finite number of articles but unknowable, and (c) keeping infinite number of articles.

Sample is known as a part from population which represents that particular population's properties. As much as the sample selection will be unbiased and random, it will be more representing its population. "Sample is a part of a population selected (usually according to some procedure and with some purpose in mind) such that it is considered to be representative of the population as a whole".

Parametric statistics is defined to have an assumption of normal distribution for its population under study. "Parametric statistics refers to those statistical techniques that have been developed on the assumption that the data are of a certain type. In particular the measure should be an interval scale and the scores should be drawn from a normal distribution".

There are certain basic assumptions of parametric statistics. The very first characteristic of parametric statistics is that it moves after confirming its population's property of **normal distribution**. The normal distribution of a population shows its symmetrical spread over the continuum of –3 SD to +3 SD and keeping unimodal shape as its mean, median, and mode coincide. If the samples are from various populations then it is assumed to have same variance ratio among them. The samples are independent in their selection. The chances of occurrence of any event or item out of the total population are equal and any item can be selected in the sample. This reflects the randomized nature of sample which also happens to be a good tool to avoid any experimenter bias.

In view of the above assumptions, parametric statistics seem to be more reliable and authentic as compared to the nonparametric statistics. These statistics are more powerful to establish the statistical significance of effects and differences among variables. It is more appropriate and reliable to use parametric statistics

in case of large samples as it consist of more accuracy of results. The data to be analysed under parametric statistics are usually from interval scale.

However, along with many advantages, some disadvantages have also been noted for the parametric statistics. It is bound to follow the rigid assumption of normal distribution and further it narrows the scope of its usage. In case of small sample, normal distribution cannot be attained and thus parametric statistics cannot be used. Further, computation in parametric statistics is lengthy and complex because of large samples and numerical calculations. T-test, F-test, r-test, are some of the major parametric statistics used for data analysis.

Nonparametric statistics are those statistics which are not based on the assumption of normal distribution of population. Therefore, these are also known as distribution free statistics. They are not bound to be used with interval scale data or normally distributed data. The data with non-continuity are to be tackled with these statistics. In the samples where it is difficult to maintain the assumption of normal distribution, nonparametric statistics are used for analysis. The samples with small number of items are treated with nonparametric statistics because of the absence of normal distribution. It can be used even for nominal data along with the ordinal data. Some of the usual nonparametric statistics include chi-square, Spearman's rank difference method of correlation, Kendall's rank difference method, Mann-Whitney U test, etc.

Self Assessment Questions

1) State true/false for the following statements

   i) Parametric statistics is known as distribution free statistics (T/ F)

   ii) Nonparametric tests assume normality of distribution (T/F)

   iii) T test is an example of parametric test (T/F)

   iv) Nonparametric tests are not bound to be used with interval scale. (T/F)

   v) Parametric tests are bound to be used with either interval or ratio scale. (T/F)

   vi) In case of small sample where normal distribution can not be attained, the use of nonparametric test is more appropriate. (T/F)

2) Define the term sample and population with one example each.

   .....

   .....

   .....

   .....

   .....

   .....

   .....

1.4 SCOPE AND USE OF STATISTICS

Statistical applications have a wide scope. Some of the major ones are given below:

Policy planning: To finalise a policy, it requires some data from previous or expected environment that the policy can be effectively utilised with maximum favourable results. For example, in an organisation the previous sales data are analysed to develop future strategies in the field to obtain maximum benefit in terms of product sale.

Management: Statistics is very useful tool in an organisation to view various aspects of work and well being of the employees as well as keeping an eye on the progress trend of the organisation.

Behavioural and Social Sciences: In social sciences where both types (quantitative and qualitative) of information are used, statistics helps the researchers to alter the information in a comprehensive way to explain and predict the patterns of behaviour/ trend. Where the characteristics of the population being studied are normally distributed, the best and statistically important decision about variables being investigated is possible by using parametric statistics or nonparametric statistics to explain the pattern of activities.

Education: If education is intended to be well dispersed and effective in the interest of the population, the characteristics of students, instructor's contents and infrastructure are very important to understand and again statistics enable these characteristics being analysed in context of needs of the nation. Once the parameters of all components are analysed, areas needing more emphasis become obvious.

Commerce and Accounts: Where money matters are involved, it is essential to take extra care to manage the funds properly enabling efforts in various sectors. The cost and benefit analysis helps to decide putting money and regulating it for maximum benefit at minimum cost.

Industries: Statistics is a basic tool to handle daily matters not only in big organisations but also in small industries. It is required, at each level, to keep data with care and look at them in different perspectives to mitigate the expenditure and enable each employee to have his/ her share in the benefit. Psychologists/ personnel officers dealing with selection and training in industries also use statistical tools to differentiate among employees.

Pure sciences and Mathematics: Statistical tools are also instrumental to have precise measures in pure sciences and to see differences on different occasions in various conditions. Statistics itself is a branch of mathematics which helps them understand differences among properties of various applications in mathematics.

Problem solving: Knowing the useful difference between two or more variables enable the individual to find out the best applicable solution to a problem situation and it is possible because of statistics. During problem solving statistics helps the person analyse his/ her pattern of response and the correct solution thereby minimising the error factor.

Theoretical researches: Theories evolve on the basis of facts obtained from the field. Statistical analyses establish the significance of those facts for a particular paradigm or phenomena. Researchers are engaged in using the statistical measures to decide on the facts and data whether a particular theory can be maintained or challenged. The significance between the facts and factors help them to explore the connectivity among them.

1.5 LIMITATIONS OF STATISTICS

Although Statistics has a very wide application in everyday life as well as in Behavioural Sciences, Physical and Natural Sciences, it has certain limitations also. These limitations are as follow :

Statistics deals with aggregate of facts. It cannot deal with single observation. Thus statistical methods do not give any recognition to an object or a person or an event in isolation. This is a serious limitation of Statistics.

Since Statistics is a science dealing with numerical data, it is more applicable to those phenomenon which can be measured quantitatively. However, the techniques of statistical analysis can be applied to qualitative phenomenon indirectly by expressing them numerically with the help of quantitative standards.

Statistical conclusions are true only on the average . Thus, statistical inferences may not be considered as exact like inferences based on Mathematical laws.

1.6 DISTRUST AND MISUSE OF STATISTICS

Sometimes irresponsible, inexperienced people use statistical tools to fulfill their self motives irrespective of the nature and trend of the data. Because of such various misuses of statistical tools sometimes called an unscrupulous science. There are various misgivings about Statistics . These are as follows :

“Statistics can prove anything”

“Statistics is an unreliable science”

“There are three types of lies , namely, lies, damned lies, and statistics.”

“An ounce of truth will produce tons of Statistics “

Therefore care and precautions should be taken care for the interpretation of statistical data. “ Statistics should not be used as a blind man uses a lamp-post for support instead of illumination”

There are many other fields like, agriculture, space, medicine, geology, technology, etc. where statistics is extensively used to predict the results and find out precision in decision.

Self Assessment Question

1) Write three application of statistics in daily life.

.....

.....

.....

2) List atleast two misuses of statistics.

.....

.....

.....

.....

.....

.....

.....

.....

1.7 LET US SUM UP

In present era people must have some knowledge of statistics. In singular sense, it means statistical methods which include collection, classification, analysis and interpretation of data. In plural sense, it means quantitative information called data. Descriptive, correlational and inferential statistics are three different type of statistics on the basis of their functions. On the other hand, parametric and non parametric are other types of statistics on the basis of the nature of distribution. Statistics has application in almost in all branches of knowledge as well as all sphere of life. Inspite of its wide applicability, it has certain limitations too. Some times inexperienced people misuse statistics to fulfill their own motives.

1.8 UNIT END QUESTIONS

1) What do you mean by statistics? Define its various types with the help of examples of daily life.
2) "Statistical methods are most dangerous tools in the hand of inexpert." Discuss briefly
3) Define following concepts:
   i) Descriptive statistics
   ii) Inferential statistics
   iii) Parametric statistics
   iv) Non parametric statistics
4) Comments on the following statements in two or three lines with reasons:
   i) Statistics in singular sense implies statistical methods.
   ii) Statistics and statistic implies same thing.
   iii) Statistics may rightly be called the science of averages.
   iv) There are lies, damn lies and statistics. Give three examples of misuse of statistics.
5) Write a note on the limitations of statistics.

1.9 GLOSSARY

Statistics in singular sense : In singular sense, it means scientific methods for collection, presentation, analysis and interpretation of data.

Statistics in plural sense : In plural sense it means a set of numerical scores known as statistical data.

Correlational statistics : The statistics which speaks about one or more than one variable's positive or negative magnitude of relationship.

Descriptive statistics : The statistics which describes the tendency or variance of the scores in a distribution.

Inferential statistics : The statistics that enable the researchers to have some conclusions about population or events on the basis of past or observed observations.

Non parametric statistics : The statistics free from the assumptions of normal distribution.

Parametric statistics : The statistics based on assumptions of normal distribution

Statistics : The branch of mathematics that deals with inferring the chances of a particular pattern of population or events on the basis of observed patterns..

1.10 SUGGESTED READINGS

Asthana H.S, and Bhushan, B.(2007) *Statistics for Social Sciences* (with SPSS Applications). Prentice Hall of India

B.L.Aggrawal (2009). *Basic Statistics*. New Age International Publisher, Delhi.

Gupta, S.C.(1990) *Fundamentals of Statistics*. Himalaya Publishing House, Mumbai

UNIT 2 DESCRIPTIVE STATISTICS

Structure

2.0 INTRODUCTION

We have learned in the previous unit that looking at the functions of statistics point of view, statistics may be descriptive, correlational and inferential. In this unit we shall discuss the various aspects of descriptive statistics, particularly how to organise and discribe the data.

Most of the observations in this universe are subject to variability, especially observations related to human behaviour. It is a well known fact that Attitude, Intelligence, Personality, etc. differ from individual to individual. In order to make a sensible definition of the group or to identify the group with reference to their observations/ scores, it is necessary to express them in a precise manner. For this purpose observations need to be expressed as a single estimate which summarises the observations. Such single estimate of the series of data which summarises the distribution are known as parameters of the distribution. These parameters define the distribution completely. In this unit we will be focusing on descriptive statistics, the characteristic features and the various statistics used in this category.

2.1 OBJECTIVES

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

- Define the nature and meaning of descriptive statistics;
- Describe the methods of organising and condensing raw data;
- Explain concept and meaning of different measures of central tendency; and
- Analyse the meaning of different measures of dispersion.

2.2 MEANING OF DESCRIPTIVE STATISTICS

Let us take up a hypothetical example of two groups of students taking a problem solving test. One group is taken as experimental group in that the subjects in this group are given training in problem solving while the other group subjects do not get any training. Both were tested on problem solving and the scores they obtained were as given in the table below.

Table 2.1: Hypothetical performance scores of 10 students of experimental and control groups based on Problem solving experiment.

Experimental Condition (With Training)	Control Condition (Without Training)
4	2
8	4
12	10
8	6
7	3
9	4
15	8
6	4
5	2
8	3

The scores obtained by children in the two groups are the actual scores and are considered as raw scores. A look at the table shows that the control group subjects have scored rather lower as compared to that of the experimental group which had undergone training. There are some procedures to be followed and statistical tests to be used to describe the raw data and get some meaningful interpretation of the same. This is what Descriptive statistics is all about. Description of data performs two operations: (i) Organising Data and (ii) Summarising Data.

2.3 ORGANISING DATA

Univariate analysis involves the examination across cases of one variable at a time. There are four major statistical techniques for organising the data. These are:

- Classification
- Tabulation
- Graphical Presentation
- Diagrammatical Presentation

2.3.1 Classification

The classification is a summary of the frequency of individual scores or ranges of scores for a variable. In the simplest form of a distribution, we will have such value of variable as well as the number of persons who have had each value.

Once data are collected, researchers have to arrange them in a format from which they would be able to draw some conclusions.

The arrangement of data in groups according to similarities is known as classification. Thus by classifying data, the investigators move a step ahead to the scores and proceed forward concrete decision. Classification is done with following objectives:

- Presenting data in a condensed form
- Explaining the affinities and diversities of the data
- Facilitating comparisons
- Classification may be qualitative and quantitative
- Frequency distribution.

A much clear picture of the information of score emerges when the raw data are organised as a frequency distribution. Frequency distribution shows the number of cases following within a given class interval or range of scores. A frequency distribution is a table that shows each score as obtained by a group of individuals and how frequently each score occurred.

Frequency distribution can be with ungrouped data and grouped data

i) **An ungrouped frequency distribution** may be constructed by listing all score value either from highest to lowest or lowest to highest and placing a tally mark (/) besides each scores every times it occurs. The frequency of occurrence of each score is denoted by 'f' .

ii) **Grouped frequency distribution:** If there is a wide range of score value in the data, then it is difficult to get a clear picture of such series of data. In this case grouped frequency distribution should be constructed to have clear picture of the data. A group frequency distribution is a table that organises data into classes, into groups of values describing one characteristic of the data. It shows the number of observations from the data set that fall into each of the class.

Construction of Frequency Distribution

Before proceeding we need to know a few terminologies used in further discussion as for instance, a variable. A variable refers to the phenomenon under study. It may be the performance of students on a problem solving issue or it can be a method of teaching students that could affect their performance.

Here the performance is one variable which is being studied and the method of teaching is another variable that is being manipulated. Variables are of two kinds :

i) Continuous variable

ii) Discrete variable.

Those variables which can take all the possible values in a given specified range is termed as Continuous variable. For example, age (it can be measured in years, months, days, hours, minutes , seconds etc.) , weight (lbs), height(in cms), etc. On the other hand those variables which cannot take all the possible values within the given specified range are termed as discrete variables. For example, number of children, marks obtained in an examination (out of 200), etc.

Preparation of Frequency Distribution

To prepare a frequency distribution, we, first decide the range of the given data, that is, the difference between the highest and lowest scores. This will tell about the range of the scores. Prior to the construction of any grouped frequency distribution, it is important to decide the following

1) **The number of class intervals:** There is no hard and fast rules regarding the number of classes into which data should be grouped . If there are very few scores, it is useless to have a large number of class-intervals. Ordinarily, the number of classes should be between 5 to 30

2) **Limits of each class interval:** Another factor used in determining the number of classes is the size/ width or range of the class which is known as 'class interval' and is denoted by 'i'.

Class interval should be of uniform width resulting in the same-size classes of frequency distribution. The width of the class should be a whole number and conveniently divisible by 2, 3, 5, 10 or 20.

There are three methods for describing the class limits for distribution:

i) Exclusive method

ii) Inclusive method

iii) True or actual class method

i) **Exclusive method:** In this method of class formation, the classes are so formed that the upper limit of one class also becomes the lower limit of the next class. Exclusive method of classification ensures continuity between two successive classes. In this classification, it is presumed that score equal to the upper limit of the class is exclusive, i.e., a score of 40 will be included in the class of 40 to 50 and not in a class of 30 to 40

ii) **Inclusive method:** In this method classification includes scores, which are equal to the upper limit of the class. Inclusive method is preferred when measurements are given in whole numbers.

iii) **True or Actual class method:** In inclusive method upper class limit is not equal to lower class limit of the next class. Therefore, there is no continuity between the classes.

However, in many statistical measures continuous classes are required. To have continuous classes it is assumed that an observation or score does not just represent a point on a continuous scale but an internal unit length of which the given score is the middle point.

Thus, mathematically, a score is internal when it extends from 0.5 units below to 0.5 units above the face value of the score on a continuum. These class limits are known as true or actual class limits.

Types of frequency distributions: There are various ways to arrange frequencies of a data array based on the requirement of the statistical analysis or the study. A couple of them are discussed below:

Relative frequency distribution: A relative frequency distribution is a distribution that indicates the proportion of the total number of cases observed at each score value or internal of score values.

Cumulative frequency distribution: Sometimes investigator is interested to know the number of observations less than a particular value. This is possible by computing the cumulative frequency. A cumulative frequency corresponding to a class-interval is the sum of frequencies for that class and of all classes prior to that class.

Cumulative relative frequency distribution: A cumulative relative frequency distribution is one in which the entry of any score of class interval expresses that score's cumulative frequency as a proportion of the total number of cases. Given below are ability scores of 20 students.

10, 14, 14, 13, 16, 17, 18, 20, 22, 23, 23, 24, 25, 18, 12, 13, 14, 16, 19, 20

Let us see how the above scores could be formed into a frequency distribution.

Scores	Frequency	Cum. Freq.	Rel. Cum.Freq.
10	1	1	1/20
12	1	2	2/20
13	2	4	4/20
14	3	7	7/20
16	2	9	9/20
17	1	10	10/20
18	2	12	12/20
19	1	13	13/20
20	2	15	15/20
22	1	16	16/20
23	2	18	18/20
24	1	19	19/20
25	1	20	20/20
Total	**20**		

Percentile: Cumulative frequency distribution are often used to find percentiles also. A percentile is the score at or below which a specified percentage of score in a distribution fall. For example, if the 40^{th} percentile on a exam is 75, it means that 40% of the scores on the examination are equal to or less than 75.

Frequency distribution can be either in the form of a table or it can be in the form of graph.

2.3.2 Tabulation

Tabulation is the process of presenting the classified data in the form of a table. A tabular presentation of data becomes more intelligible and fit for further statistical analysis. A table is a systematic arrangement of classified data in row and columns with appropriate headings and sub-headings.

Components of a Statistical Table

The main components of a table are :

Table number, Title of the table, Caption, Stub, Body of the table, Head note, Footnote, and Source of data

TITLE

Stub Head / Stub Entries	Caption			
	Column Head I		Column Head II	
	Sub Head	Sub Head	Sub Head	Sub Head
	MAIN BODY	OF	THE TABLE	
Total				

Footnote(s) :

Source :

Self Assessment Questions

1) Statistical techniques that summarise, organise and simplify data are called as:

   i) Population statistics ii) Sample statistics

   iii) Descriptive statistics iv) Inferential statistics

2) Which one of the alternative is appropriate for descriptive statistics?

   i) In a sample of school children, the investigator found an average weight was 35 Kg.

   ii) The instructor calculates the class average on their final exam. Was 76%

   iii) On the basis of marks on first term exam, a teacher predicted that Ramesh would pass in the final examination.

   iv) Both (i) and (ii)

3) Which one of the following statement is appropriate regarding objective/s of classification.

   i) Presenting data in a condensed form

   ii) Explaining the affinities and diversities of the data

   iii) Facilitating comparisons

   iv) All of these

4) Define the following terms

   i) Discrete variable

   ii) Continuous variable

   iii) Ungrouped frequency distribution

   iv) Grouped frequency distribution.

2.3.3 Graphical Presentation of Data

The purpose of preparing a frequency distribution is to provide a systematic way of "looking at" and understanding data. To extend this understanding, the information contained in a frequency distribution often is displayed in a graphic and/or diagrammatic forms. In graphical presentation of frequency distribution, frequencies are plotted on a pictorial platform formed of horizontal and vertical lines known as graph.

The graphs are also known as polygon, chart or diagram.

A graph is created on two mutually perpendicular lines called the X and Y–axes on which appropriate scales are indicated.

The horizontal line is called the abscissa and vertical the ordinate. Like different kinds of frequency distributions there are many kinds of graph too, which enhance the scientific understanding of the reader. The commonly used among these are bar graphs, line graphs, pie, pictographs, etc. Here we will discuss some of the important types of graphical patterns used in statistics.

Histogram: It is one of the most popular method for presenting continuous frequency distribution in a form of graph. In this type of distribution the upper limit of a class is the lower limit of the following class. The histogram consists of series of rectangles, with its width equal to the class interval of the variable on horizontal axis and the corresponding frequency on the vertical axis as its heights.

Frequency polygon: Prepare an abscissa originating from 'O' and ending to 'X'. Again construct the ordinate starting from 'O' and ending at 'Y'.

Now label the class-intervals on abscissa stating the exact limits or midpoints of the class-intervals. You can also add one extra limit keeping zero frequency on both side of the class-interval range.

The size of measurement of small squares on graph paper depends upon the number of classes to be plotted.

Next step is to plot the frequencies on ordinate using the most comfortable measurement of small squares depending on the range of whole distribution.

To obtain an impressive visual figure it is recommended to use the 3:4 ratio of ordinate and abscissa though there is no tough rules in this regard.

To plot a frequency polygon you have to mark each frequency against its concerned class on the height of its respective ordinate.

After putting all frequency marks a draw a line joining the points. This is the polygon. A polygon is a multi-sided figure and various considerations are to be maintained to get a smooth polygon in case of smaller N or random frequency distribution.

Frequency Curve : A frequency curve is a smooth free hand curve drawn through frequency polygon. The objective of smoothing of the frequency polygon is to eliminate as far as possible the random or erratic fluctuations that is present in the data.

Cumulative Frequency Curve or Ogive

The graph of a cumulative frequency distribution is known as cumulative frequency curve or ogive. Since there are two types of cumulative frequency distribution e.g., " less than" and " more than" cumulative frequencies. We can have two types of ogives.

i) **'Less than' Ogive:** In 'less than' ogive , the less than cumulative frequencies are plotted against the upper class boundaries of the respective classes. It is an increasing curve having slopes upwards from left to right.

ii) **'More than' Ogive:** In more than ogive , the more than cumulative frequencies are plotted against the lower class boundaries of the respective classes. It is decreasing curve and slopes downwards from left to right.

2.3.4 Diagrammatic Presentations of Data

A diagram is a visual form for the presentation of statistical data. They present the data in simple, readily comprehensible form. Diagrammatic presentation is used only for presentation of the data in visual form, whereas graphic presentation of the data can be used for further analysis. There are different forms of diagram e.g., Bar diagram, Sub-divided bar diagram, Multiple bar diagram, Pie diagram and Pictogram.

Bar Diagram: This is known as dimensional diagram also. Bar diagram is most useful for categorical data. A bar is defined as a thick line. Bar diagram is drawn from the frequency distribution table representing the variable on the horizontal axis and the frequency on the vertical axis. The height of each bar will be corresponding to the frequency or value of the variable.

Sub- divided Bar Diagram: Study of sub classification of a phenomenon can be done by using sub-divided bar digram. Corresponding to each sub-category of the data the bar is divided and shaded. There will be as many shades as there will sub portion in a group of data. The portion of the bar occupied by each sub-class reflect its proportion in the total.

Multiple Bar Diagram: This diagram is used when comparison are to be shown between two or more sets of interrelated phenomena or variables. A set of bars for person, place or related phenomena are drawn side by side without any gap. To distinguish between the different bars in a set, different colours, shades are used.

Pie Diagram: It is also known as angular diagram. A pie chart or diagram is a circle divided into component sectors corresponding to the frequencies of the variables in the distribution. Each sector will be proportional to the frequency of the variable in the group. A circle represent 3600. So 360 angle is divided in proportion to percentages. The degrees represented by the various component parts of given magnitude can be obtained by using this formula.

After the calculation of the angles for each component, segments are drawn in the circle in succession corresponding to the angles at the center for each segment. Different segments are shaded with different colour, shades or numbers.

Pictograms: It is known as cartographs also. In pictogram we used appropriate picture to represent the data. The number of picture or the size of the picture being proportional to the values of the different magnitudes to be presented. For showing population of human beings, human figures are used. We may represent 1 Lakh people by one human figure. Pictograms present only approximate values.

Self Assessment Questions

1) Explain the following terms.
   i) Histogram
   ii) Frequency polygon

iii) Bar diagram

iv) Pictogram

2) Ordinarily number of class should be:

(i) 5 to 10 (ii) 10 to 20 (iii) 5 to 30 (iv) None of these

3) Define the Inclusive and Exclusive method of classification

.....

.....

.....

.....

.....

4) Distinguish between relative frequency distribution and cumulative frequency distribution.

.....

.....

.....

.....

.....

2.4 SUMMARISING DATA

In the previous section we have discussed about tabulation of the data and its representation in the form of graphical presentation. In research, comparison between two or more series of the same type is needed to find out the trends of variables. For such comparison, tabulation of data is not sufficient and it is further required to investigate the characteristics of data. The frequency distribution of obtained data may differ in two ways, first in measures of central tendency and second, in the extent to which scores are spread over the central value. Both types of differences are the components of summary statistics.

2.4.1 Measures of Central Tendency

It is the middle point of a distribution and is also known as measures of location. Tabulation of data provides the data in a systematic order and enhances their understanding. However, most of the time, you may be interested to find out the differences between two or more classes. Generally, in any distribution values of the variables tend to cluster around a central value of the distribution. This tendency of the distribution is known as central tendency and measures devised to consider this tendency is know as measures of central tendency. In considering measures of central tendency the idea of representativeness is important. A measure of central tendency is useful if it represents accurately the distribution of scores on which it is based. A good measure of central tendency must possess the following characteristics:

It should be rigidly defined: The definition of a measure of central tendency should be clear and unambiguous so that it leads to one and only one information.

It should be readily comprehensible and easy to compute: The average should be such that even a non-mathematician can easily understand and compute it.

It should be based on all observations: A good measure of central tendency should be based on all the values of the distribution of scores.

It should be amenable for further mathematical treatment: If we are given two sets of data and a measure of central tendency for both of them, we should be able to calculate the measure for the combined data also.

It should be least affected by the fluctuation of sampling: If independent random samples of the same size are selected from a population, the value of average for each one of them should be sufficiently close to one another.

In Statistics there are three most commonly used measures of central tendency. These are:

1) Mean,
2) Median, and
3) Mode.

1) **Mean:** The arithmetic mean is most popular and widely used measure of central tendency. Whenever we refer to the average of data, it means we are talking about its arithmetic mean. This is obtained by dividing the sum of the values of the variable by the number of values.

   Merits and limitations of the arithmetic mean: The very first advantage of arithmetic mean is its universality, i.e., it remains in every data set. The arithmetic mean remains to be very clear and only single in a data set. It is also a useful measure for further statistics and comparisons among different data sets. One of the major limitations of arithmetic mean is that it cannot be computed for open-ended class-intervals.

2) **Median:** Median is the middle most value in a data distribution. It divides the distribution into two equal parts so that exactly one half of the observations is below and one half is above that point. Since median clearly denotes the position of an observation in an array, it is also called a position average. Thus more technically, median of an array of numbers arranged in order of their magnitude is either the middle value or the arithmetic mean of the two middle values. For example, the set of numbers 2, 3, 5, 7, 9, 12, 15 has the median 7.

   Thus, for ungrouped data median is $\left[\frac{n+1}{2}\right]^{th}$ value in case data are in their magnitude order, where n denotes the number of given observations.

   Merits: It is not affected by extreme values in the distribution. In other words, median is a better measure of central tendency in cases where very small or large items are present in the distribution. It can be calculated even in the case of open-ended classes.

3) **Mode:** Mode is the value in a distribution that corresponds to the maximum concentration of frequencies. It may be regarded as the most typical of a series value. In more simple words, mode is the point in the distribution comprising maximum frequencies therein.

Usually mode remains near the center of a distribution. In a unimodal type of distribution it coincides with mean and median. For ungrouped data it is defined as the datum value, which occurs most frequently.. When the scores and frequencies are presented as a simple frequency distribution, the mode is the score value that appears most often in frequency distribution.

Merits: It is readily comprehensible and easy to compute. It is not affected by extreme values. It can be easily calculated even in the case of open-end classes.

2.4.2 Measures of Dispersion

In the previous section we have discussed about measures of central tendency. By knowing only the mean, median or mode, it is not possible to have a complete picture of a set of data. Average does not tell us about how the score or measurements are arranged in relation to the center. It is possible that two sets of data with equal mean or median may differ in terms of their variability. Therefore, it is essential to know how far these observations are scattered from each other or from the mean. Measures of these variations are known as the 'measures of dispersion'. The most commonly used measures of dispersion are range, average deviation, quartile deviation, variance and standard deviation.

Range: Range is one of the simplest measures of dispersion. It is designated by 'R'. The range is defined as the difference between the largest score and the smallest score in the distribution. It is known as distance between the highest and the lowest scores in a distribution. It gives the two extreme values of the variable but no information about the values in between the extreme values. A large value of range indicates greater dispersion while a smaller value indicates lesser dispersion among the scores.

Merits: Range can be a good measure if the distribution is not much skewed. If the data are at the ordinal level, then range is only measure, which is technically meaningful.

Average Deviation: Average deviation refers to the arithmetic mean of the differences between each score and the mean. It is always better to find the deviation of the individual observations with reference to a certain value in the series of observation and then take an average of these deviations. This deviation is usually measured from mean or median. Mean, however, is more commonly used for this measurement. Average deviation is commonly denoted as AD. One of its prominent characteristics is that at the time of summing all the deviations from the mean, the positive or negative signs are not considered.

Merits: It is less affected by extreme values as compared to standard deviation. It provides better measure for comparison about the formation of different distributions.

Quartile Deviation

Quartile deviation is denoted as Q. It is also known as inter-quartile range. It avoids the problems associated with range. Inter-quartile range includes only 50% of the distribution. Quartile deviation is the difference between the 75% and 25% scores of a distribution. 75^{th} percentile is the score which keeps 75% score below itself and 25^{th} percentile is the score which keeps 25% scores below itself.

Merits and limitations: QD is a simple measure of dispersion. While the measure of central tendency is taken as median, QD is most relevant to find out the dispersion of the distribution. In comparison to range, QD is more useful because range speaks about the highest and lowest scores while QD speaks about the 50% of the scores of a distribution. As middle 50% of scores are used in QD there is no effect of extreme scores on computation, giving more reliable results. In case of open-end distribution QD is more reliable in comparison to other measures of dispersion. It is not recommended to use QD in further mathematical computations. It is not a complete reliable measure of distribution as it doesn't include all the scores. As QD is based on 50% scores, it is not useful to study in each and every statistical situation.

Standard deviation: Standard deviation is the most stable index of variability. In the computations of average deviation, the signs of deviation of the observations from the mean were not considered. In order to avoid this discrepancy, instead of the actual values of the deviations we consider the squares of deviations, and the outcome is known as variance. Further, the square root of this variance is known as standard deviation and designated as SD. Thus, standard deviation is the square root of the mean of the squared deviations of the individual observations from the mean. The standard deviation of the sample and population denoted by s and s, respectively.

Properties of SD

If all the score have an identical value in a sample, the SD will be 0 (zero).

In different samples drawn from the same population, SDs differ very less as compared to the other measures of dispersion.

For a symmetrical or normal distribution, the following relationship are true:

Mean ±1 SD covers 68.26 % cases

Mean ± 2 SD covers 95.45 % cases

Mean ± 3 SD covers 99.73 % cases

Merits: It is based on all observations. It is amenable to further mathematical treatments. Of all measures of dispersion, standard deviation is least affected by fluctuation of sampling.

Skewness and Kurtosis

There are two other important characteristics of frequency distribution that provide useful information about its nature. They are known as skewness and Kurtosis.

Skewness: Skewness is the degree of asymmetry of the distribution. In some frequency distributions scores are more concentrated at one end of the scale. Such a distribution is called a skewed distribution.

Thus, Skewness refers to the extent to which a distribution of data points is concentrated at one end or the other. Skewness and variability are usually related, the more the Skewness the greater the variability.

Skewness has both, direction as well as magnitude. In actual practice, frequency distributions are rarely symmetrical; rather they show varying degree of asymmetry or Skewness.

In perfectly symmetrical distribution, the mean, median and mode coincide, whereas this is not the case in a distribution that is asymmetrical or skewed. If

the frequency curve of a distribution has a longer tail to the right side of the origin, the distribution is said to be skewed positively (Fig.2.1).

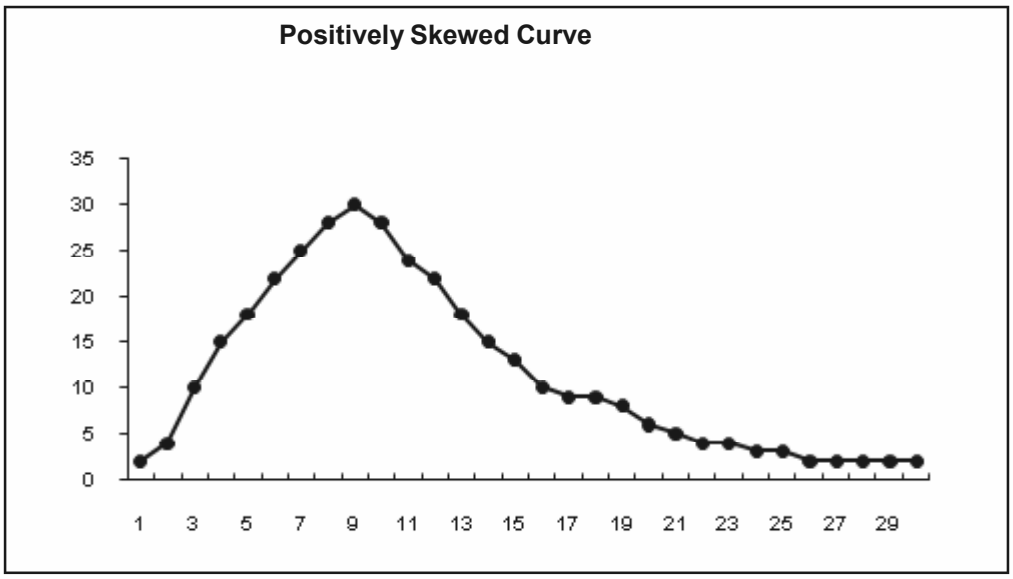


Fig.2.1: Positively Skewed Curve

In case the curve is having long tail towards left or origin, it is said to be negatively Skewed (Fig. 2.2).

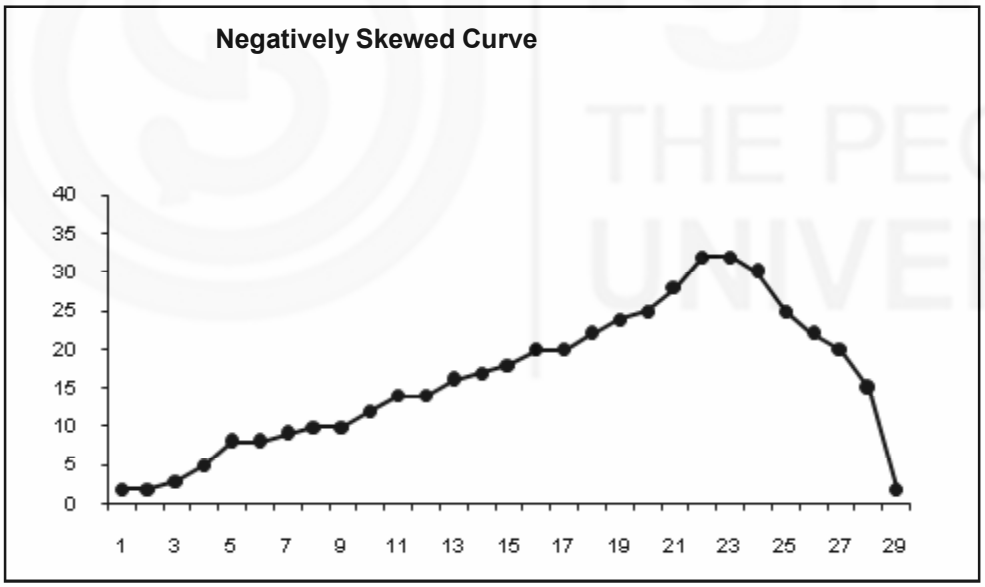


Fig.2.2: Negatively Skewed Curve

There are two measures of Skewness, i.e., SD and percentile. There are different ways to compute Skewness of a frequency distribution.

Kurtosis: The term 'kurtosis' refers to the 'peakedness' or flatness of a frequency distribution curve when compared with normal distribution curve. The Kurtosis of a distribution is the curvedness or peakedness of the graph as depicted in figure 2.3.

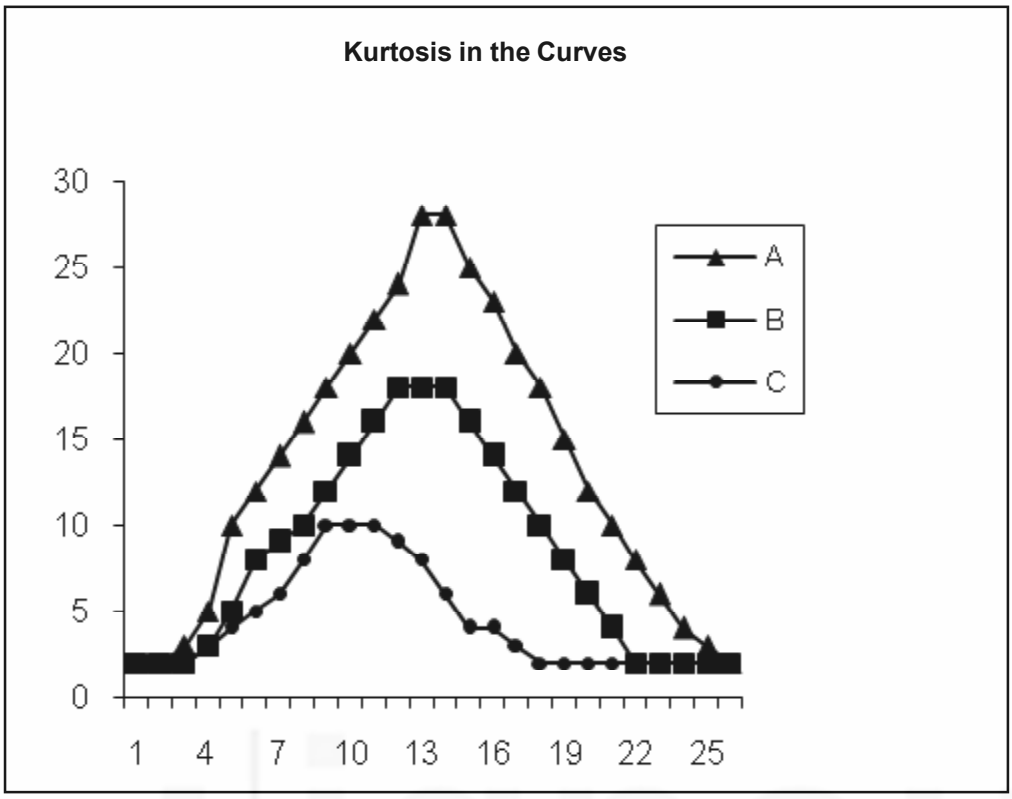


Fig. 2.3: Kurtosis in the Curves

Here, distribution A is more peaked than normal and is said to be leptokurtic. This kind of peakedness implies a thin distribution. On the other side B is more platykurtic than the normal. Platykurtic implies a flat distribution. A normal curve is known as Mesokurtic. Thus Kurtosis is the relative flatness of top and is measured by β_2. The normal curve is known as β_2 (= 3), Platykurtic curve is known as β_2 (< 3), and Leptokurtic curves are known as β_2 (>3). Platykurtic distribution produces the value of Kurtosis, which remains less than 3 and for Leptokurtic, the Kurtosis is greater than 3, and in Mesokurtic, it is equal to 3.

Self Assessment Questions

1) Which one is the most frequently used measures of central tendency

   i) Arithmetic mean ii) Geometric mean

   iiii) Mode iv) Moving average

2) Which measures of central tendency is concerned with position ?

   i) Mode, ii) Median, iiii) Arithmetic mean, iv) None of these

3) State whether the following statements are true (T) or false (F)

   i) Arithmetic mean is not affected by extreme values. ( )

   ii) Mode is affected by extreme values ( )

   iii) Mode is useful in studying qualitative facts such as intelligence ( )

   iv) Median is not affected by extreme values ( )

   v) Range is most unstable measures of variability ( )

   vi) Standard deviation is most suitable measures of dispersion ( )

   vii) Skewness is always negative ( )

2.5 USE OF DESCRIPTIVE STATISTICS

Descriptive statistics are used to describe the basic features of the data in a study.

- They provide simple summaries about the sample and the measures.
- Together with simple graphical analysis, they form the basis of virtually every quantitative analysis of data.
- With descriptive statistics one is simply describing what is in the data or what the data shows.
- Descriptive Statistics are used to present quantitative descriptions in a manageable form.
- In any analytical study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simplify large amounts of data in a sensible way.
- Each descriptive statistic reduces lots of data into a simple summary.
- Every time you try to describe a large set of observations with a single indicator, you run the risk of distorting the original data or losing important detail.
- Even given these limitations, descriptive statistics provide a powerful summary that may enable you to make comparisons across people or other units.

2.6 LET US SUM UP

Descriptive statistics are used to describe the basic features of the data in investigation. Such statistics provide summaries about the sample and measures. Data description comprises two operations : organising data and describing data. Organising data includes : classification, tabulation , graphical and diagrammatic presentation of raw scores. Whereas, measures of central tendency and measures of dispersion are used in describing the raw scores.

2.7 UNIT END QUESTIONS

1) What do you mean by Descriptive statistics? Discuss its importance briefly.
2) Define the following terms:

   i) Class interval ii) Upper limit of class interval

   iii) lower limit of class interval, iv) Midpoint of class interval
3) What do you mean by organisation of data ? Describe various methods for organising data.
4) How can you describe the data? State the various types of measures of central tendency and their respective uses.
5) What do you mean by measures of dispersion? Explain why the range is relatively unstable measures of variability.

2.8 GLOSSARY

Abscissa	:	X axis
Array	:	A rough grouping of data.
Classification	:	A systematic grouping of data
Cumulative frequency distribution	:	A classification, which shows the cumulative frequency below, the upper real limit of the corresponding class interval.
Data	:	Any sort of information that can be analysed.
Discrete data	:	When data are counted in a classification.
Exclusive classification	:	The classification system in which the upper limit of the class becomes the lower limit of next class.
Frequency distribution	:	Arrangement of data values according to their magnitude.
Inclusive classification	:	When the lower limit of a class differs the upper limit of its successive class.
Secondary data	:	Information gathered through already maintained records about a variable.
Mean	:	The ratio between total and numbers of scores.
Median	:	The mid point of a score distribution.
Mode	:	The maximum occurring score in a score distribution.
Central Tendency	:	The tendency of scores to bend towards center of distribution.
Arithmetic mean	:	Mean for stable scores.
Dispersion	:	The extent to which scores tend to scatter from their mean and from each other.
Standard Deviation	:	The square root of the sum of squared deviations of scores from their mean.
Skewness	:	Tendency of scores to polarize on either side of abscissa.
Kurtosis	:	Curvedness of a frequency distribution graph.
Platykurtic	:	Curvedness with flat tendency towards abscissa.
Mesokurtik	:	Curvedness with normal distribution of scores.
Leptokurtic	:	Curvedness with peak tendency from abscissa.
Range	:	Difference between the two extremes of a score distribution.

2.9 SUGGESTED READINGS

Asthana, H. S. and Bhushan, B. (2007). *Statistics for Social Sciences* (with SPSS Application). Prentice Hall of India, New Delhi.

Yale, G. U., and M.G. Kendall (1991). *An Introduction to the Theory of Statistics*. Universal Books, Delhi.

Garret, H. E. (2005). *Statistics in Psychology and Education.* Jain Publishing, India.

Nagar, A. L., and Das, R. K. (1983). *Basic Statistics*. Oxford University Press, Delhi.

Elhance, D. N., and Elhance, V. (1988). *Fundamentals of Statistics*. Kitab Mahal, Allahabad.

UNIT 3 INFERENTIAL STATISTICS

Structure

3.0 INTRODUCTION

Before conducting any study, investigators it must be decided as to whether he/she will depend on census details or sample details. On the basis of the information contained in the sample we try to draw conclusions about the population. This process is known as statistical inference. Statistical inference is widely applicable in behavioural sciences, especially in psychology. For example, before the Lok sabha or vidhan sabha election process starts or just before the declaration of election results print media and electronic media conduct exit poll to predict the election result. In this process all voters are not included in the survey, only a portion of voters i.e. sample is included to infer about the population. This is called inferential statistics and the present unit deals with the same in detail.

3.1 OBJECTIVES

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

- define inferential statistics;
- state the concept of estimation;

- distinguish between point estimation and interval estimation; and
- explain the different concepts involved in hypothesis testing

3.2 CONCEPT AND MEANING OF INFERENTIAL STATISTICS

In the previous unit we have discussed about descriptive statistics. Descriptive statistics is used to describe data. Organising and summarizing data is only one step in the process of analysing the data. Behavioural scientists are interested in estimating population parameters from the descriptive statistics of a sample.

The reason being that quantitative research in psychology and behavioural sciences aims to test theories about the nature of the world in general (or some part of it) based on samples of "subjects" taken from the world (or some part of it). When we examine the effect of frustration on children's aggression, our intent is to create theories that apply to all children who are frustrated, or perhaps to all children in cultures having similar frustrating situations. We, of course, cannot study all children, but we can study samples of children that, hopefully, will generalise back to the populations from which the samples were taken.

Inferential statistics deals with drawing conclusions about large group of individuals (population) on the basis of observation of a few participants from among them or about the events which are yet to occur on the basis of past events. It provides tools to compute the probabilities of future behaviour of the subjects. Inferential statistics throws light on how generalisation from sample to population can be made. The fundamental question is: can we infer the population's characteristics from the sample's characteristics? Descriptive statistics remains local to the sample describing its central tendency and variability, while inferential statistics focuses on making statements about the population.

3.3 INFERENTIAL PROCEDURES

There are two types of inferential procedures : (1) Estimation , (2) Hypothesis testing

3.3.1 Estimation

In estimation a sample is drawn and studied and inference is made about the population characteristics on the basis of what is discovered about the sample. There may be sampling variations because of chance fluctuations, variations in sampling techniques, and other sampling errors. We, therefore, do not expect our estimate of the population characteristics to be exactly correct. We do, however , expect it to be close. The real question in estimation is not whether our estimate is correct or not but how close is it to be the true value.

Our first interest is in using the sample mean ($\bar{X}$) to estimate the population mean (μ).

Characteristics of X as an estimate of (μ).

The sample mean ($\bar{X}$) often is used to estimate a population mean (μ). For example, the sample mean of 45.0 from the Academic Anxiety Test may be used to estimate the mean Academic Anxiety of population of college students. Using this sample would lead to an estimate of 45.0 for the population mean. Thus, sample mean is an unbiased and consistent estimator of population mean.

Unbiased Estimator : An unbiased estimator is one which , if we were to obtain an infinite number of random samples of a certain size, the mean of the statistic would be equal to the parameter. The sample mean, ($\bar{X}$) is an unbiased estimate of (μ) because if we look at possible random samples of size N from a population, mean of the sample would be equal to μ.

Consistent Estimator : A consistent estimator is one that as the sample size increases, the probability that estimate has a value close to the parameter also increase. Because it is a consistent estimator, a sample mean based on 20 scores has a greater probability of being closer to (μ) than does a sample mean based upon only 5 scores. Better estimates of a population mean should be more probable from large samples.

Accuracy of Estimation: The sample mean is an unbiased and consistent estimator of (μ) . But we should not overlook the fact that an estimate is just a rough or approximate calculation. It is unlikely in any estimate that ($\bar{X}$) will be exactly equal to (μ). Whether or not $\bar{X}$ is a good estimate of (μ) depends upon the representativeness of sample, the sample size, and the variability of scores in the population.

3.3.2 Point Estimation

We have indicated that x obtained from a sample is an unbiased and consistent estimator of the population mean (μ) . Thus , if a researcher obtains Academic Anxiety Score from 100 students and wanted to estimate the value of (μ) for the population from which these scores were selected, researcher would use the value of $\bar{X}$as an estimate of (μ). If the obtained value of $\bar{X}$ was 45.0, this value would be used as estimate of (μ).

This form of estimate of population parameters from sample statistic is called point estimation. Point estimation is estimating the value of a parameter as a single point, for example, (μ) = 45.0 from the value of the statistic $\bar{X}$ = 45.0

3.3.3 Interval Estimation

A point estimate of the population mean almost is assured of being in error, the estimate from the sample will not equal to the exact value of the parameter. To gain confidence about the accuracy of this estimate we may also construct an interval of scores that is expected to include the value of the population mean. Such intervals are called confidence interval. A confidence interval is a range of scores that is expected to contain the value of (μ). The lower and upper scores that determine the interval are called confidence limits. A level of confidence can be attached to this estimate so that the researcher can be 95% or 99% confidence level that encompasses the population mean.

Self Assessment Questions

1) What is statistical inference?

.....

.....

.....

.....

2) Explain with illustrations the concept of (i) estimation, (ii) point estimation, and, (iii) interval estimation.

.....

.....

.....

.....

.....

3) What is the distinction between statistic and parameter?

.....

.....

.....

.....

.....

4) State the procedures involved in statistical inference.

.....

.....

.....

.....

.....

3.4 HYPOTHESIS TESTING

Inferential statistics is closely tied to the logic of hypothesis testing.In hypothesis testing we have a particular value in mind. We hypothesize that this value characterise the population of observations. The question is whether that hypothesis is reasonable in the light of the evidence from the sample. In estimation no particular population value need to be stated. Rather, the question is: What is the population value? For example, Hypothesis testing is one of the important areas of statistical analyses. Sometimes hypothesis testing is referred to as statistical decision-making process. In day-to-day situations we are required to take decisions about the population on the basis of sample information. For example, on the basis of sample data, we may have to decide whether a new method of teaching is better than the existing one, whether new medicine is more effective in curing the disease than the previously available medicine, and so forth.

3.4.1 Statement of Hypothesis

A statistical hypothesis is defined as a statement, which may or may not be true about the population parameter or about the probability distribution of the parameter that we wish to validate on the basis of sample information. Most of the times, experiments are performed with random samples instead of the entire

population and inferences drawn from the observed results are then generalised over the entire population. But before drawing inferences about the population, it should always be kept in mind that the observed results might have come due to chance factor. In order to have an accurate or more precise inference, the chance factor should be ruled out. The probability of chance occurrence of the observed results is examined by the **null hypothesis** (H_0). Null hypothesis is a statement of no differences. The other way to state null hypothesis is that the two samples came from the same population. Here, we assume that population is normally distributed and both the groups have equal means and standard deviations.

Since the null hypothesis is a testable proposition, there is counter proposition to it known as **alternative hypothesis** and denoted by H_1. In contrast to null hypothesis H_1 proposes that the two samples belong to two different populations, that their means are estimates of two different parametric means of the respective population, and there is a significant difference between their sample means. The alternative hypothesis is not directly tested statistically; rather its acceptance or rejection is determined by the rejection or retention of the null hypothesis. The probability 'p' of the null hypothesis being correct is assessed by a statistical test. If probability 'p' is too low, H_0 is rejected and H_1 is accepted. It is inferred that the observed difference is significant. If probability 'p' is high, H_0 is accepted and it is inferred that the difference is due to the chance factor and not due to the variable factor.

3.4.2 Level of Significance

The level of significance (α) is that probability of chance occurrence of observed results up to and below which the probability 'p' of the null hypothesis being correct is considered too low and the results of the experiment are considered significant ($p \leq \alpha$). On the other hand, if p exceeds α, the null hypothesis (H_0) cannot be rejected because the probability of it being correct is considered quite high and in such case, observed results are not considered significant ($p > \alpha$). The selection of level of significance depends on the choice of the researcher. Generally level of significance is taken to be 5% or 1%, i.e., $\alpha = .05$ or $\alpha = .01$). If null hypothesis is rejected at .05 level, it means that the results are considered significant so long as the probability 'p' of getting it by mere chance of random sampling works out to be 0.05 or less ($p < .05$). In other words, the results are considered significant if out of 100 such trials only 5 or less number of the times the observed results may arise from the accidental choice in the particular sample by random sampling.

3.4.3 One-tail and Two-tail Test

Depending upon the statement in alternative hypothesis (H_1), either a one-tail or two-tail test is chosen for knowing the statistical significance. A one-tail test is a directional test. It is formulated to find the significance of both the magnitude and the direction (algebraic sign) of the observed difference between two statistics. Thus, in two-tailed tests researcher is interested in testing whether one sample mean is significantly higher (alternatively lower) than the other sample mean. Here, the entire rejection region (α) of the null hypothesis distribution is on a single tail, i.e., either positive or negative tail. The probability 'p' of the H_0 being correct is given by the fractional area is a single tail (see Figure 3.1)

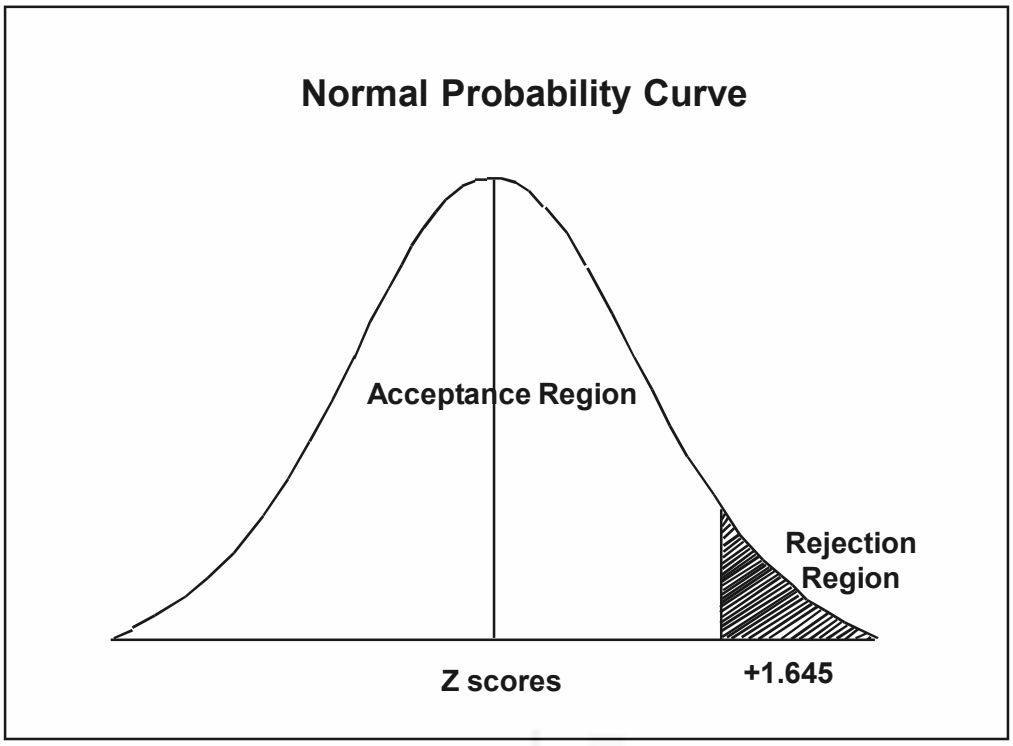


Fig. 3.1: Rejection region of the null hypothesis in a one-tailed test

A two-tail test is a non-directional statistical test for finding out significance of the magnitude of the observed difference between the statistics of two samples. In a two-tailed test hypothesis, we reject the null hypothesis if the sample mean is significantly higher or lower than the population mean. In two-tail test the rejection region of the null hypothesis involves both the tails (see Figure 3.2) , amounting to p/2 in each tail. Thus, the total factorial area is both the tails give the probability 'p' of the null hypothesis being correct.

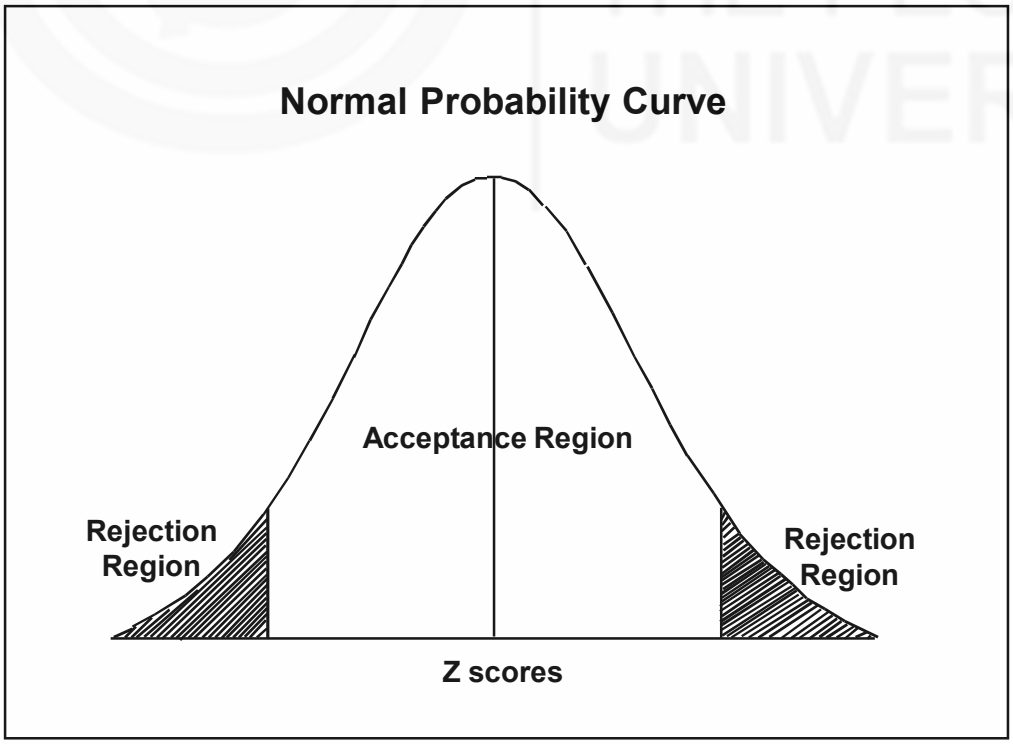


Fig. 3.2: Rejection region of the null hypothesis in a two-tailed test

For a two-tailed test with the p chosen to be .05, each tail of the H_0 distribution ends with a rejection or critical region of area .025 extending beyond the critical Z score of 1.96 in the tail. If the computed Z score lies between –1.96 to +1.96, then the observed difference falls within the rejection region and consequently null hypothesis is rejected. But in a one tail test with chosen p = .05, if the computed Z score is equal to or greater than 1.645 then the observed difference falls within the rejection region. Hence, the null hypothesis is rejected. It is clear that with an identical p, an observed difference may be significant in a one-tail test though it may fail to be significant in a two-tail test.

3.4.4 Errors in Hypothesis Testing

In hypothesis testing, there would be no errors in decision making as long as a null hypothesis is rejected when it is false and also a null hypothesis is accepted when it is true. But the decision to accept or reject the null hypothesis is based on sample data. There is no testing procedure that will ensure absolutely correct decision on the basis of sampled data. There are two types of errors regarding decision to accept or to reject a null hypothesis.

Type I error– When the null hypothesis is true, a decision to reject it is an error and this kind of error is known as type I error in statistics. The probability of making a type I error is denoted as 'α' (read as alpha). The null hypothesis is rejected if the probability 'p' of its being correct does not exceed the p. The higher the chosen level of p for considering the null hypothesis, the greater is the probability of type I error.

Type II error– When null hypothesis is false, a decision to accept it is known as type II error. The probability of making a type II error is denoted as 'β' (read as beta). The lower the chosen level of significance p for rejecting the null hypothesis, the higher is the probability of the type II error. With a lowering of p, the rejection region as well as the probability of the type I error declines and the acceptance region (1-p) widens correspondingly.

The goodness of a statistical test is measured by the probability of making a type I or type II error. For a fixed sample size n, α and β are so related that reduction in one causes increase in the other, and therefore, simultaneous reductions in α and β are not possible. If n is increased, it is possible to decrease both α and β.

3.4.5 Power of a Test

The probability of committing type II error is designated by β. Therefore, 1-β is the probability of rejecting null hypothesis when it is false. This probability is known as the power of a statistical test. It measures how well the test is working. The probability of type II error depends upon the true value of the population parameter and sample size n.

Self Assessment Questions

1) Fill in the blanks

   i) Null hypothesis is a statement of ..... difference.

   ii) Null hypothesis is denoted by .....

   iii) Alternative hypothesis is ..... directly tested statistically.

iv) ..... is that probability of chance of occurrence of observed results.

v) Level of significance is denoted by .....

vi) When the null hypothesis is true, a decision to reject is known as .....

vii) When a null hypothesis is false, a decision to accept is known as .....

3.5 GENERAL PROCEDURE FOR TESTING A HYPOTHESIS

- Set up a null hypothesis suitable to the problem.
- Define the alternative hypothesis.
- Calculate the suitable test statistics.
- Define the degrees of freedom for the test situation.
- Find the probability level 'p' corresponding to the calculated value of the test statistics and its degree of freedom. This can be obtained from the relevant tables.
- Reject or accept null hypothesis on the basis of tabulated value and calculated value at practical probability level.

Theses are the some situations in which inferential statistics is carried out to test the hypothesis and draw conclusion about the population.

3.5.1 Test of Hypothesis About a Population Mean

If researcher is interested in testing hypothesis about the value of population mean, then Z test is most appropriate statistics. Z test is more effective under following conditions:

The population mean and standard deviation are known.

The sampling distribution of mean is normally distributed. This requires that either the sample size n should be large (n>30) or the parent population itself should be normally distributed.

3.5.2 Testing Hypothesis about a Population Mean (Small Sample)

In smaller sample size the assumption of normal approximation does not work effectively. In such situations, other sampling distribution, such as 't' is used.

The 't' distribution is appropriate whenever the population standard deviation is unknown and estimated from the sample data. The 't' distribution resembles normal distribution except that 't' has heavier tails than normal.

The 't' distribution is characterised by the degree of freedom (denoted by df). Degrees of freedom relate to the sample size, so that larger samples allow more degrees of freedom. As the degree of freedom increases, the 't' distribution comes closer to resembling a normal distribution. A 't' distribution with infinite degree of freedom is identical to the normal distribution.

3.6 't' TEST FOR SIGNIFICANCE OF DIFFERENCE BETWEEN MEANS

In experiments using small samples (n<30) drawn at random from the population the scores are distributed in the form of 't' distribution. Therefore, to test the significance of difference between the means of the two small samples that difference is converted to 't' score.

3.6.1 Assumption for 't' Test

The dependent variable should be continuous.

The variable has normal distribution in the population.

Each score of the dependent variable occur at random and independent of all other scores in the sample.

The sample comes from population having identical variance.

3.6.2 't' test for Independent Sample

Two or more random samples, used in an independent group experiment, are drawn from the population independent of each other so that such sample consist of separate group of individuals and may or may not be identical in size. One of these random samples serves as the control group (the subjects are not given any independent treatment) while the other constitute the experimental group (the subjects are treated with independent variable). After such treatment, the dependent variable being investigated is measured in both the groups. The difference between two such group's mean may be estimated by the 't' test in different ways according to the nature of samples of groups. These are given below:

a) *For independent samples of small and unequal sizes*:

b) *For both small and large independent samples of equal size*:

c). *For large samples of unequal sizes*: When both the sample sizes are large (more than 30), but not identical, the 't' score is computed for the difference between sample means, using SDs of the individual sample.

3.6.3 't' Test for Paired Observation by Difference Method

In paired observation, single group scores first as the control group and subsequently as the experimental group. In control condition, group is measured for dependent variable without induction of the independent variable. In experimental condition, the same group is treated with independent variables followed by the measurement of the dependent variable. 't' test is used to find out the significance of difference between means of paired scores of a small group (n<30) in such a single group experiment.

3.7 LET US SUM UP

This unit is intended to aware the learner to the basic concepts and general procedure involved in statistical inference. Inferential statistics is about inferring or drawing conclusions from the sample to population. This process is known as statistical inference. There are two types of inferential procedures : estimation

and hypothesis testing. An estimate of unknown parameter could be either point or interval. Sample mean is usually taken as a point estimate of population mean. Whereas in interval estimation we construct upper and lower limits around the sample mean.

Hypothesis is a statement about a parameter. There are two types of hypotheses: null and alternative hypotheses. Important concepts involved in the process of hypothesis testing e.g., level of significance, one tail test, two tail test, type I error, type II error, power of a test are explained. General procedure for hypothesis testing is also given.

3.8 UNIT END QUESTIONS

1) Explain the importance of inferential statistics.
2) Describe the important properties of good estimators.
3) What do you mean by statement of hypothesis?
4) Discuss the different types of hypothesis formulated in hypothesis testing .
5) Discuss the errors involved in hypothesis testing.
6) Explain the concept of level of significance, one tail test, two tail test and power of a test.
7) Explain the various steps involved in hypothesis testing.

3.9 GLOSSARY

Confidence Level : It gives the percentage (probability) of samples where the population mean would remain within the confidence interval around the sample mean.

Estimation : It is a method of prediction about parameter value on the basis Statistic.

Hypothesis testing : The statistical procedures for testing hypotheses.

Independent sample : Samples in which the subjects in the groups are different individuals and not deliberately matched on any relevant characteristics.

Level of significance : The probability value that forms the boundary between rejecting and not rejecting the null hypothesis.

Null hypothesis : The hypothesis that is tentatively held to be true (symbolized by Ho)

One-tail test : A statistical test in which the alternative hypothesis specifies direction of the departure from what is expected under the null hypothesis.

Parameter : It is a measure of some characteristic of the population.

Population : The entire number of units of research inerest

Power of a test	:	An index that reflects the probability that a statistical test will correctly reject the null hypothesis relative to the size of the sample involved.
Sample	:	A sub set of the population under study
Statistical Inference	:	It is the process of concluding about an unknown population from known sample drawn from it
Statistical hypothesis	:	The hypothesis which may or may not be true about the population parameter.
t-test	:	It is a parametric test for the significance of differences between means.
Type I error	:	A decision error in which the statistical decision is to reject the null hypothesis when it is actually true.
Type II error	:	A decision error in which the statistical decision is not to reject the null hypothesis when it is actually false.
Two-tail test	:	A statistical test in which the alternative hypothesis does not specify the direction of departure from what is expected under the null hypothesis.

3.10 SUGGESTED READINGS

Asthana, H. S. and Bhushan, B. (2007). *Statistics for Social Sciences* (with SPSS Application). Prentice Hall of India, New Delhi.

Yale, G. U., and M.G. Kendall (1991). *An Introduction to the Theory of Statistics*. Universal Books, Delhi.

Garret, H. E. (2005). *Statistics in Psychology and Education*. Jain publishing, India.

Nagar, A. L., and Das, R. K. (1983). *Basic Statistics*. Oxford University Press, Delhi.

Elhance, D. N., and Elhance, V. (1988). *Fundamentals of Statistics*. Kitab Mahal, Allahabad.

Sani, F., and Todman, J. (2006). *Experimental Design and Statistics for Psychology*. A first course book. Blackwell Publishing.

Howell, D. C. (2002). *Statistical Method for Psychology*. Pacific Grove, CA.

UNIT 4 FREQUENCY DISTRIBUTION AND GRAPHICAL PRESENTATION

Structure

4.0 INTRODUCTION

Data collected either from Primary or Secondary source need to be systemetically presented as these are invariably in unsystematic or rudimentary form. Such raw data fail to reveal any meaningful information. The data should be rearranged and classified in a suitable manner to understand the trend and message of the collected information. This unit therefore, deals with the method of getting the data organised in all respects in a tabular form or in graphical presentation.

4.1 OBJECTIVES

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

- Explain the methods of organising and condensing statistical data;

- Define the concepts of frequency distribution and state its various types;
- Analyse the different methods of presenting the statistical data;
- Explain how to draw tables and graphs diagrams, pictograms etc; and
- describe the uses and misuses of graphical techniques.

4.2 ARRANGEMENT OF DATA

After data collection, you may face the problem of arranging them into a format from which you will be able to draw some conclusions. The arrangement of these data in different groups on the basis of some similarities is known as classification. According to Tuttle, " A classification is a scheme for breaking a category into a set of parts, called classes, according to some precisely defined differing characteristics possessed by all the elements of the category"

Thus classification, is the process of grouping data into sequences according to their common characteristics, which separate them into different but related parts. Such classification facilitates analysis of the data and consequently prepares a foundation for absolute interpretation of the obtained scores. The prime objective of the classification of data is concerned with reducing complexities with raw scores by grouping them into some classes. This will provides a comprehensive insight into the data.

The classification procedure in statistics enables the investigators to manage the raw scores in such a way that they can proceed with ease in a systematic and scientific manner. There are different ways of organising and presenting the raw data. Let us discuss one by one.

4.2.1 Simple Array

The simple array is one of the simplest ways to present data. It is an arrangement of given raw data in ascending or descending order. In ascending order the scores are arranged in increasing order of their magnitude. For example, numbers 2,4,7,8,9,12, are arranged in ascending order. In descending order the scores are arranged in decreasing order of their magnitude. For example, numbers 12, 9, 8, 7, 4,2, are arranged in descending order. Simple array has several advantages as well as disadvantages over raw data. Using Simple array, we can easily point out the lowest and highest values in the data and the entire data can be easily divided into different sections. Repetition of the values can be easily checked, and distance between succeeding values in the data can be observed on the first look. But sometimes a data array is not very helpful because it lists every observation in the array. It is cumbersome for displaying large quantities of data.

4.2.2 Discrete Frequency Distribution

Here different observations are not written as in simple array. Here we count the number of times any observation appears which is known as frequency. The literary meaning of frequency is the number or occurrence of a particular event/ score in a set of sample. According to Chaplin (1975) "frequency distribution shows the number of cases falling within a given class interval or range of scores." A frequency distribution is a table that organises data into classes, i.e., into groups of values describing one characteristic of the data. It shows the number of observations from the data set that fall into each of the classes. An example is presented in the table below.

Table 4.1: Frequency distribution of persons in small scale industry according to their wages per month.

Wages per month (Rs.)	500	550	700	750
No. of Persons	21	25	18	20

When the number of observations is large , the counting of frequency is often done with the help of tally bars vertical strokes (I). A bunch of four marks is crossed by fifth to make counting simpler (𝍸)

Table 4.2: Frequency distribution of number of persons and their wages per month

Wages per month (Rs.)	Tally Sheet	Frequency
500	𝍸 𝍸 𝍸 𝍸 I	21
550	𝍸 𝍸 𝍸 𝍸 𝍸	25
700	𝍸 𝍸 𝍸 III	18
750	𝍸 𝍸 𝍸 𝍸	20
Total		**84**

4.2.3 Grouped Frequency Distribution

The quantitative phenomena under study is termed as Variable. Variables are of two kinds : (i) continuous variable, and (ii) discrete variable. Those variables which can take all the possible values in a given specified range are termed as Continuous variable. For example, age (it can be measured in years, months, days, hours, minutes , seconds etc.) , weight (lbs), height(in cms), etc.

On the other hand, those variables which cannot take all the possible values within the given specified range, are termed as discrete variables. For example, number of children, marks obtained in an examination (out of 200), etc.

To prepare a grouped frequency distribution, first we decide the range of the given data, i.e., the difference between the highest and lowest scores. This will tell about the range of the scores. Prior to the construction of any grouped frequency distribution, it is important to decide following things:

1) What would be the number of class intervals ?
2) What would be the limits of each class interval ?
3) How would the class limits be designated ?

1) What would be the number of class intervals? There is no specific rules regarding the number of classes into which data should be grouped.

   If there are very few scores, it is useless to have a large number of class-intervals. Ordinarily, the number of classes should be between 5 to 30. The number of classes would also depend on the number of observations. The larger the number of observations, the more will be the classes. With less number of classes accuracy is lost and with more number, the computation becomes tiresome.

   Usually the formula to determine the number of classes is given by

   $$\text{Number of classes} = 1 + 3.322 \times \log 10\ N$$

Where N is the total number of observations.

In this example scores of 30 students are given below. Let us prepare the frequency distribution by using exclusive method of classification.

3, 30, 14, 30, 27, 11, 25, 16, 18, 33, 49, 35, 18, 10, 25, 20, 14, 18, 9, 39, 14, 29, 20, 25, 29, 15, 22, 20, 29, 29

In above example of raw data of 30 students, the number of classes can be calculated as under :

$$\text{Number of Classes} = 1 + 3.322 \times \log 10\,(30) = 1 + 3.322 \times 1.4771 = 1 + 4.9069262$$

$$1 + 5 = 6$$

2) What would be the limits of each class interval ? Another factor used in determining the number of classes is the size/ width or range of the class which is known as 'class interval' and is denoted by 'i'. Class interval should be of uniform width resulting in the same-size classes of frequency distribution. The width of the class should be a whole number and conveniently divisible by 2, 3, 5, 10, or 20.

The width of a class interval (i) = Largest Observation(OL – OS) / I (class interval)

After deciding the class interval, the range of scores should be decided by subtracting the highest value to the lowest value of the data array.

Now, the next step is to decide from where the class should be started. There are three methods for describing the class limits for distribution

- Exclusive method
- Inclusive method
- True or actual class method

Exclusive method: In this method of class formation, the classes are so formed that the upper limit of one class also becomes the lower limit of the next class. Exclusive method of classification ensures continuity between two successive classes. In this classification, it is presumed that score equal to the upper limit of the class is exclusive, i.e., a score of 40 will be included in the class of 40 to 50 and not in a class of 30 to 40.

Finally we count the number of scores falling in each class and record the appropriate number in frequency column. The number of scores falling in each class is termed as class frequency. Tally bar is used to count these frequencies.

Example: Scores of 30 students are given below. Prepare the frequency distribution by using exclusive method of classification.

3, 30, 14, 30, 27, 11, 25, 16, 18, 33, 49, 35, 18, 10, 25, 20, 14, 18, 9, 39, 14, 29, 20, 25, 29, 15, 22, 20, 29, 29

The above ungrouped data do not provide any useful information about observations rather it is difficult to understand.

Solution:

Step 1: First of all arrange the raw scores in ascending order of their magnitude.

3,9,10,11,14,14,14,15,16,18,18,18,20,20,20,22,25,25,25,27,29,29,29,29,30,30,33,35,39,49

Step 2: Determine the **range** of scores by adding 1 to the difference between

largest and smallest scores in the data array. For above array of data it is 49–3 = 46+1= 47.

Step 3: Decide the number of classes. Say 5 for present array of data.

Step 4: To decide the approximate size of class interval, divide the range with the decided number of classes (5 for this example) . If the quotient is in fraction, accept the next integer. For examples, 47/5 = 9.4. Take it as 10

Step 5: Find the lower class-limit of the lowest class interval and add the width of the class interval to get the upper class-limit. (e.g. 3 – 12)

Step 6: Find the class-limits for the remaining classes.(13-22), (23-32), (33-42), (43-52)

Step 7: Pick up each item from the data array and put the tally mark (I) against the class to which it belongs. Tallies are to mark in bunch of five, four times in vertical and fifth in cross-tally on the first four. Count the number of observations, i.e., frequency in each class. (an example is given)

Table 4.3: Representation of preparing class-interval by marking the tallies for data frequencies in the exclusive method.

Class Interval	Tallies	Frequency
40-50	I	1
30-40	𝍸 I	6
20-30	𝍸 𝍸 I	11
10-20	𝍸 𝍸	10
0-10	II	2
		30

***Note*:** The tallying of the observations in frequency distribution may be checked out for any omitted or duplicated one that the sum of the frequencies should equal to the total number of scores in the array.

Inclusive method: In this method classification includes scores, which are equal to the upper limit of the class. Inclusive method is preferred when measurements are given in the whole numbers. Above example may be presented in the following form by using inclusive method of classification.(Refer to table below)

Table 4.4: An illustration of preparing the class interval by marking the tallies for data frequencies in an inclusive method.

Class Interval	Tallies	Frequency
40-49	I	1
30-39	𝍸 I	6
20-29	𝍸 𝍸 I	11
10-19	𝍸 𝍸	10
0-9	II	2
		30

True or Actual class method: In inclusive method upper class limit is not equal to lower class limit of the next class. Therefore, there is no continuity between the classes. However, in many statistical measures continuous classes are required. To have continuous classes, it is assumed that an observation or score does not just represent a point on a continuous scale but an internal of unit length of which the given score is the middle point. Thus, mathematically, a score is internal when it extends from 0.5 units below to 0.5 units above the face value of the score on a continuum. These class limits are known as true or actual class limits.

Table 4.5: A representation of preparing the class interval in an exact method.

Exclusive Method	Inclusive Method	True or Exact Method
70-80	70-79	69.5-79.5
60-70	60-69	59.5-69.5
50-60	50-59	49.5-59.5
40-50	40-49	39.5-49.5
30-40	30-39	29.5-39.5
20-30	20-29	19.5-29.5

4.2.4 Types of Grouped Frequency Distributions

There are various ways to arrange frequencies of a data array based on the requirement of the statistical analysis or the study. A few of them are discussed below.

i) **Open End Frequency Distribution:** Open end frequency distribution is one which has at least one of its ends open. Either the lower limit of the first class or upper limit of the last class or both are not specified. Example of such frequency distribution is given in Table below.

Table 4.6: Open end class frequency

Class	Frequency
Below 10	5
10- 20	7
20-30	10
30 -40	9
40-50	4

ii) **Relative frequency distribution**: A relative frequency distribution is a distribution that indicates the proportion of the total number of cases observed at each score value or internal of score values.

iii) **Cumulative frequency distribution**: Sometimes investigator is interested to know the number of observations less than a particular value. This is possible by computing the cumulative frequency. A cumulative frequency corresponding to a class-interval is the sum of frequencies for that class and of all classes prior to that class.

iv) **Cumulative relative frequency distribution**: A cumulative relative frequency distribution is one in which the entry of any score of class-interval

expresses that score's cumulative frequency as a proportion of the total number of cases. The following table 4.7 shows frequency distributions for ability scores of 100 students.

Table 4.7: A representation of different kinds of frequency distributions

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cum. Relative Frequency
95-99	5	.05	100	1.00
90-94	3	.03	95	.95
85-89	7	.07	92	.92
80-84	4	.04	85	.85
75-79	4	.04	81	.81
70-74	7	.07	77	.77
65-69	9	.09	70	.70
60-64	8	.08	61	.61
55-59	4	.04	53	.53
50-54	9	.09	49	.49
45-49	13	.13	40	.40
40-44	12	.12	27	.27
35-39	5	.05	15	.15
30-34	10	.10	10	.10
	100	**1.00**		

Self Assessment Questions

1) In exclusive series
   i) both the class limits are considered
   ii) the lower limit is excluded
   iii) both the limits are excluded
   iv) the upper limit is excluded
2) For discrete variable, more appropriate class intervals are:
   i) Exclusive ii) Inclusive iii) both
3) When both the lower and upper limits are considered , such classes are called:
   i) exclusive ii) inclusive iii) cumulative
4) In "less than" cumulative frequency distribution, the omitted limit is
   i) lower ii) upper iii) last iv) none of these

4.3 TABULATION OF DATA

Tabulation is the process of presenting the classified data in the form of a table. A tabular presentation of data becomes more intelligible and fit for further

statistical analysis. A table is a systematic arrangement of classified data in row and columns with appropriate headings and sub-headings.

4.3.1 Components of a Statistical Table

The main components of a table are given below:

Table number: When there are more than one tables in a particular analysis, a table should be marked with a number for their reference and identification. The number should be written in the center at the top of the table.

Title of the table: Every table should have an appropriate title, which describes the content of the table. The title should be clear, brief, and self-explanatory. Title of the table should be placed either centrally on the top of the table or just below or after the table number.

Caption: Captions are brief and self-explanatory headings for columns. Captions may involve headings and sub-headings. The captions should be placed in the middle of the columns. For example, we can divide students of a class into males and females, rural and urban, high SES and Low SES etc.

Stub: Stubs stand for brief and self-explanatory headings for rows. A relatively more important classification is given in rows. Stub consist of two parts : (i) Stub head : It describes the nature of stub entry (ii) Stub entry : It is the description of row entries.

Body of the table: This is the real table and contains numerical information or data in different cells. This arrangement of data remains according to the description of captions and stubs.

Head note: This is written at the extreme right hand below the title and explains the unit of the measurements used in the body of the tables.

Footnote: This is a qualifying statement which is to be written below the table explaining certain points related to the data which have not been covered in title, caption, and stubs.

Source of data : The source from which data have been taken is to be mentioned at the end of the table. Reference of the source must be complete so that if the potential reader wants to consult the original source they may do so.

4.3.2 General Rules for Preparing a Table

There are no hardcore rules for preparing a table. But tabulation requires a lot of skills and common sense of the researcher. Though specific rules have not been provided for tabulation, some general rules are in fashion/ tradition. They are as follows:

- Table should be compact and readily comprehensible being complete and self-explanatory.
- It should be free from confusion.
- It should be arranged in a given space. It should neither be too small or too large.

- Items in the table should be placed logically and related items should be placed nearby.
- All items should be clearly stated.
- If item is repeated in the table, its full form should be written.
- The unit of measurement should be explicitly mentioned preferably in the form of a head note.
- The rules of forming a table is diagrammatically presented in the table below.

TITLE

Stub Head	**Caption**			
Stub Entries	**Column Head I**		**Column Head II**	
	Sub Head	**Sub Head**	**Sub Head**	**Sub Head**
	MAIN BODY	**OF**	**THE TABLE**	
Total				

Footnote(s) :

Source :

4.3.3 Importance of Tabulation

Tabulation is the process of condensation of data for convenience in statistical processing, presentation and interpretation of the information combined therein. Importance of tabulation is given below.

It simplifies complex data: If data are presented in tabular form, these can be readily understood. Confusions are avoided while going through the data for further analysis or drawing the conclusions about the observation.

It facilitates comparison: Data in the statistical table are arranged in rows and columns very systematically. Such an arrangement enables you to compare the information in an easy and comprehensive manner.

Tabulation presents the data in true perspective: With the help of tabulation, the repetitions can be dropped out and data can be presented in true perspective highlighting the relevant information.

Figures can be worked-out more easily: Tabulation also facilitates further analysis and finalization of figures for understanding the data.

Self Assessment Questions

1) What points are to be kept in mind while taking decision for preparing a frequency distribution in respect of (a) the number of classes and (b) width of class interval.

.....

.....

.....

.....

2) Differentiate between following pairs of statistical terms

   i) Column and row entry

   ii) Caption and stub head

   iii) Head note and foot note

   .....

   .....

   .....

   .....

3) State briefly the importance of tabulation in statistical analysis.

   .....

   .....

   .....

   .....

   .....

4.4 GRAPHICAL PRESENTATION OF DATA

The frequency distribution in itself again provides a somewhat rough picture of observations. The viewers remain unable to locate the contour of the actual spread of score distributed among different classes. Through frequency distribution it is not possible to have a physical image of the scores of a sample and it merely reflects the counts among the classes. To elaborate the data array after constructing the frequency distribution, it is scientific tradition to plot the frequencies on a pictorial platform formed of horizontal and vertical lines, known as 'graph'. The graphs are also known as polygon, chart or diagram. A graph is created on two mutually perpendicular lines called the X and Y–axes on which appropriate scales are indicated. The horizontal line is called abscissa and the vertical ordinate. Like different kinds of frequency distributions, there are many kinds of graph too, which enhance the scientific understanding to the reader. The commonly used among these are bar graphs, line graphs, pie, pictographs, etc. Here we will discuss some of the important types of graphical patterns used in statistics.

4.4.1 Histogram

It is one of the most popular method for presenting continuous frequency distribution in a form of graph. In this type of distribution the upper limit of a class is the lower limit of the following class. The histogram consists of series of rectangles, with its width equal to the class interval of the variable on horizontal axis and the corresponding frequency on the vertical axis as its heights. The steps in constructing a histogram are as follows:

Step 1: Construct a frequency distribution in table form.

Step2: Before drawing axes, decide on a suitable scale for horizontal axis then determine the number of squares (on the graph paper) required for the width of the graph.

Step 3: Draw bars equal width for each class interval. The height of a bar corresponds to the frequency in that particular interval . The edge of a bar represents both the upper real limit for one interval and the lower real limit for the next higher interval.

Step 4: Identify class intervals along the horizontal axis by using either real limit or midpoint of class interval. In case of real limits, these will be placed under the edge of each bar. On the other hand , if you use midpoint of class interval, it will be placed under the middle of each bar.

Step 5: Label both axes and decide appropriate title to the histogram.

Table 4.8: Results of 200 students on Academic achievement test.

Class Interval	Frequency
10- 20	12
20- 30	10
30- 40	35
40- 50	55
50- 60	45
60- 70	25
70- 80	18

Let us take a simple example to demonstrate the construction of histogram based on the above data.

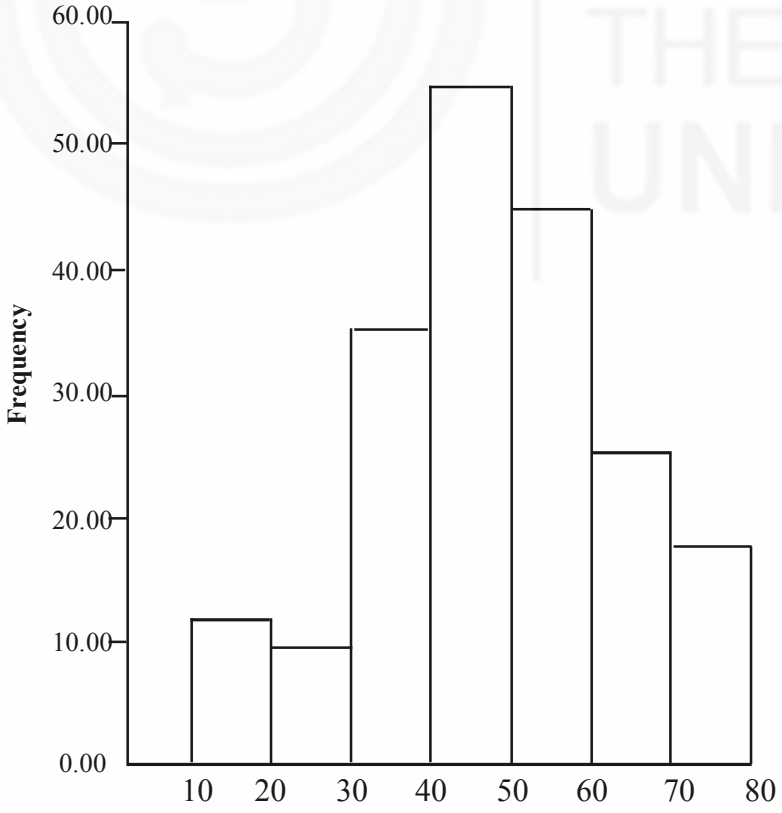


Fig. 4.1: Histogram

4.4.2 Frequency Polygon

Prepare an abscissa originating from 'O' and ending to 'X'. Again construct the ordinate starting from 'O' and ending at 'Y'. Now label the class-intervals on abscissa stating the exact limits or midpoints of the class-intervals. There is also a fashion to add one extra limit keeping zero frequency on both side of the class-interval range. The size of measurement of small squares on graph paper depends upon the number of classes to be plotted. Next step is to plot the frequencies on ordinate using the most comfortable measurement of small squares depending on the range of whole distribution. To obtain an impressive visual figure it is recommended to use the 3:4 ratio of ordinate and abscissa though there is no tough rules in this regard. To plot a frequency polygon you have to mark each frequency against its concerned class on the height of its respective ordinate. After putting all frequency marks a draw a line joining. This is the polygon. A polygon is a multi-sided figure and various considerations are to be maintained to get a smooth polygon in case of smaller N or random frequency distribution. The very common way is to compute the smoothed frequencies of the classes by having the average of frequencies of that particular class along with upper and lower classes' frequencies. For instance, the frequency 4 of class-interval 75-79 might be smoothed as 6+4+5 /3 = 5.

The frequency polygon of data given in Table 4.8 in graph

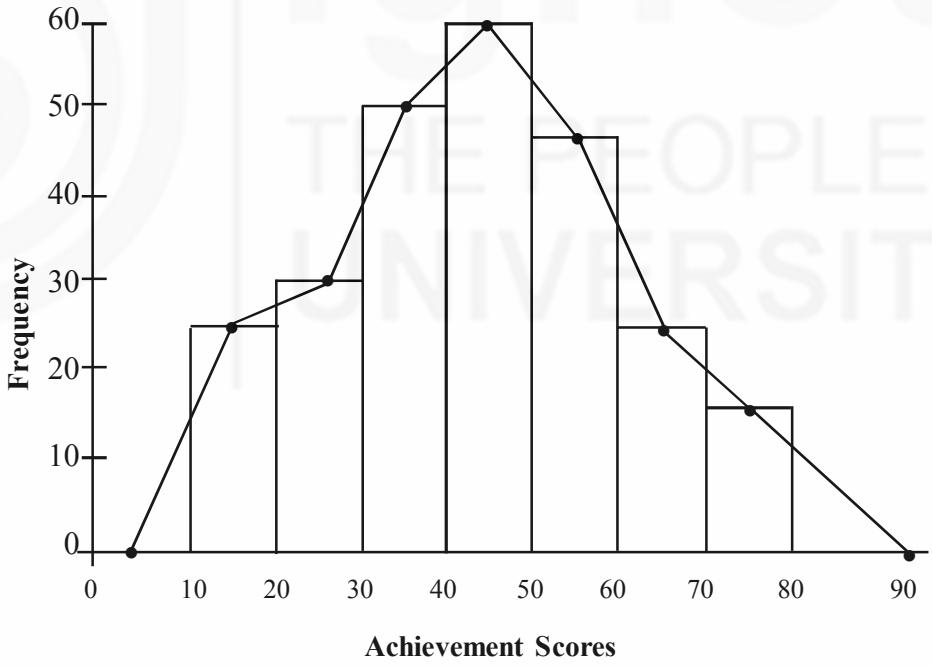


Fig.4.2: Frequency polygon

4.4.3 Frequency Curve

A frequency curve is a smooth free hand curve drawn through frequency polygon. The objective of smoothing of the frequency polygon is to eliminate as far as possible the random or erratic fluctuations present in the data.

Frequency curve is shown in Fig.4.2 based on data presented in Table 4.8

4.4.4 Cumulative Frequency Curve or Ogive

The graph of a cumulative frequency distribution is known as cumulative frequency curve or ogive. Since there are two types of cumulative frequency distribution e.g., " less than" and " more than" cumulative frequencies, we can have two types of ogives.

i) *'Less than' Ogive:* In 'less than' ogive , the less than cumulative frequencies are plotted against the upper class boundaries of the respective classes. It is an increasing curve having slopes upwards from left to right.

ii) *'More than' Ogive:* In more than ogive , the more than cumulative frequencies are plotted against the lower class boundaries of the respective classes. It is decreasing curve and slopes downwards from left to right.

Example of 'Less than' and 'more than' cumulative frequencies based on data reported in table

Class Interval	Frequency	Less than c.f.	More than c.f.
10-20	12	12	200
20- 30	10	22	188
30- 40	35	57	178
40- 50	55	112	143
50- 60	45	157	88
60- 70	25	182	43
70- 80	18	200	18

The ogives for the cumulative frequency distributions given in above table are drawn in Fig. 4.3

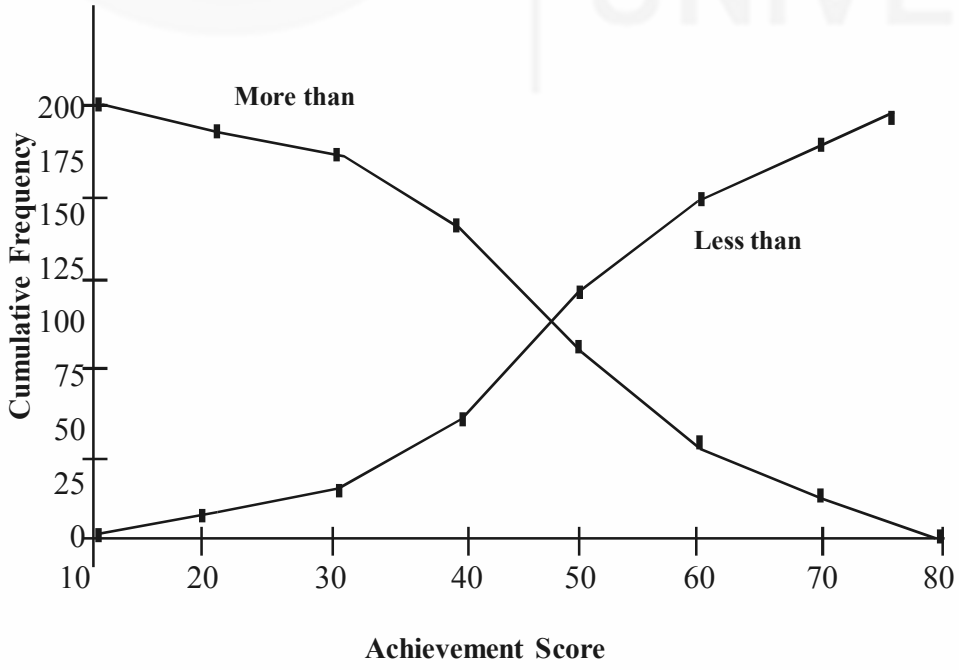


Fig. 4.3: 'Less than' and 'more than' type ogives

4.4.5 Misuse of Graphical Presentations

It is possible to mislead the observer or reader in a pictorial data presentation by manipulating the vertical (ordinate or Y-axis) and horizontal (abscissa or X-axis) lines of a graph. Elimination of zero frequency on ordinate difference among bars or ups and downs in a curve line can be highlighted in a desired way distorting the real findings of the study. Hence, utmost care should be taken while presenting the findings graphically.

4.5 DIAGRAMMATIC PRESENTATIONS OF DATA

A diagram is a visual form for the presentation of statistical data. They present the data in simple , readily comprehensible form. Diagrammatic presentation is used only for presentation of the data in visual form, whereas graphic presentation of the data can be used for further analysis. There are different forms of diagram e.g., Bar diagram, Sub-divided bar diagram, Multiple bar diagram, Pie diagram and Pictogram.

4.5.1 Bar Diagram

This is known as dimensional diagram also. Bar diagram is most useful for categorical data. A bar is defined as a thick line. Bar diagram is drawn from the frequency distribution table representing the variable on the horizontal axis and the frequency on the vertical axis. The height of each bar will be corresponding to the frequency or value of the variable. However , width of the rectangles is immaterial but proper and uniform spacing should be between different bars. It is different from the histogram when both the height and width of tha bar are important and even bars are placed adjacent to one another without any gap.

Example: In a study on causes of strikes in mills. Hypothetical data are given below.

Causes of strikes :	Economic	Personal	Political	Rivalry	Others
Occurrence of strikes:	45	13	25	7	10

Let us take the above example to demonstrate the construction of a bar diagram.

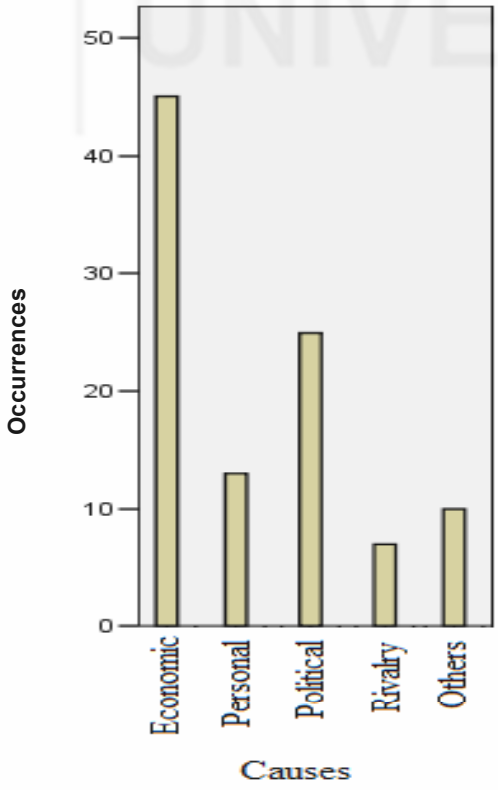


Fig. 4.4: Bar diagram

4.5.2 Sub- divided Bar Diagram

Study of sub classification of a phenomenon can be done by using sub-divided bar digram. Corresponding to each sub-category of the data, the bar is divided and shaded. There will be as many shades as there will sub portion in a group of data. The portion of the bar occupied by each sub-class reflect its proportion in the total .

Table 4.9: Hypothetical data on sales of mobile set (in thousand) in four metropolitan city.

Metropolitan City	Year		
	2006	2007	2008
Chennai	11	15	24
Delhi	15	22	30
Kolkatta	10	12	18
Mumbai	09	17	22

A sub-divided bar diagram for the hypothetical data given in above Table 4.9 is drawn in Fig. 4.5

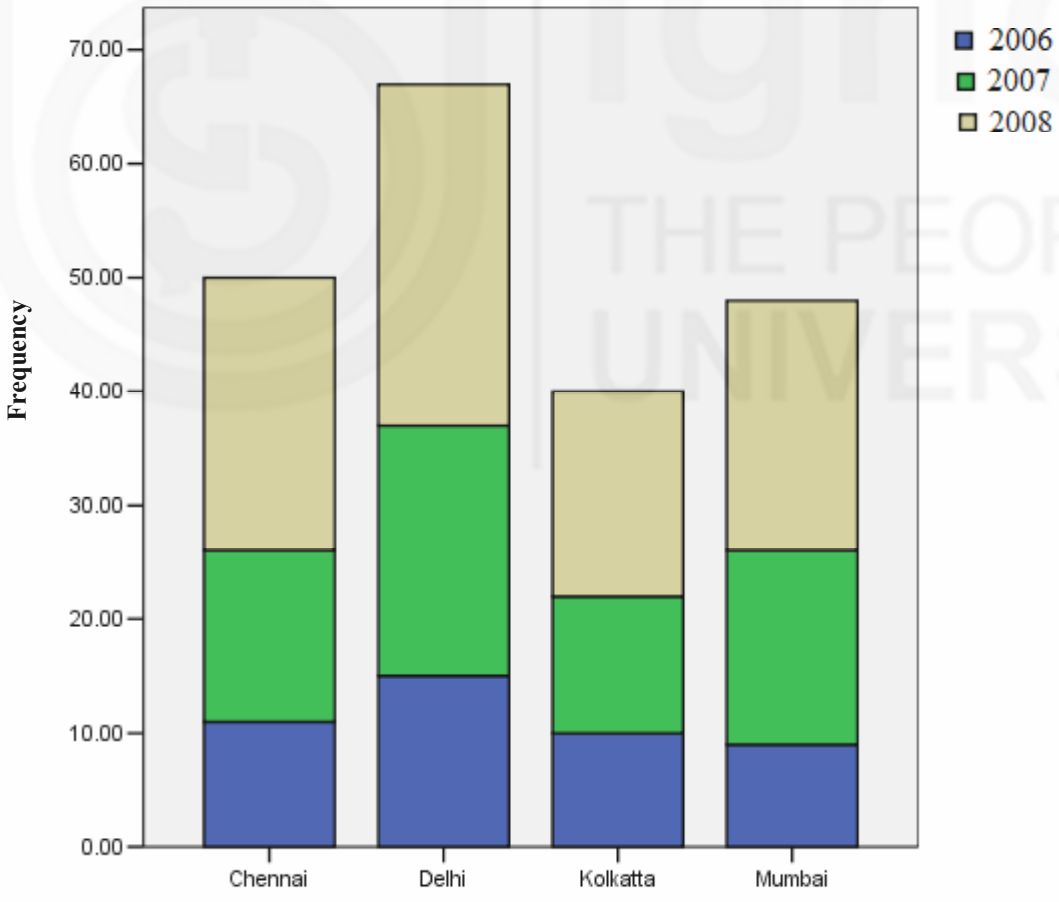


Fig. 4.5: Subdivided Bar diagram

4.5.3 Multiple Bar Diagram

This diagram is used when comparison are to be shown between two or more sets of interrelated phenomena or variables. A set of bars for person, place or

related phenomena are drawn side by side without any gap. To distinguish between the different bars in a set, different colours, shades are used.

Table 4.10: A group of three students were assessed on three different psychological parameters like, anxiety, adjustment, and stress.

Students	Anxiety	Adjustment	Stress
X	20	15	30
Y	12	25	16
Z	18	13	25

Multiple bar diagram for the hypothetical data given in table 4.10 is drawn in Fig. 4.6.

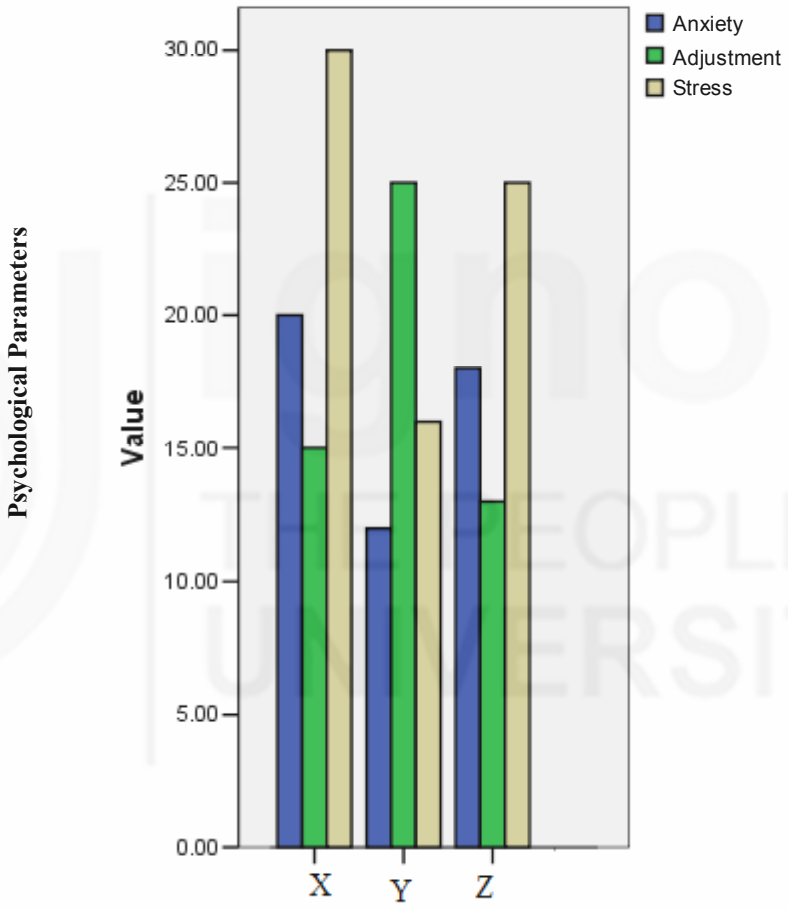


Fig. 4.6: Multiple Bar diagram

4.5.4 Pie Diagram

It is also known as angular diagram. A pie chart or diagram is a circle divided into component sectors corresponding to the frequencies of the variables in the distribution. Each sector will be proportional to the frequency of the variable in the group. A circle represent 3600. So 360 angle is divided in proportion to percentages. The degrees represented by the various component parts of given magnitude can be obtained by using this formula.

$$\text{Degree of any component part} = \frac{\text{Component Value}}{\text{Total Value}} \times 360°$$

After the calculation of the angles for each component, segments are drawn in the circle in succession corresponding to the angles at the center for each segment. Different segments are shaded with different colour, shades or numbers.

Table 4.11: 1000 software engineers pass out from a institute X and they were placed in four different company in 2009.

Company	Placement
A	400
B	200
C	300
D	100

Pie Diagram Representing Placement in four different company.

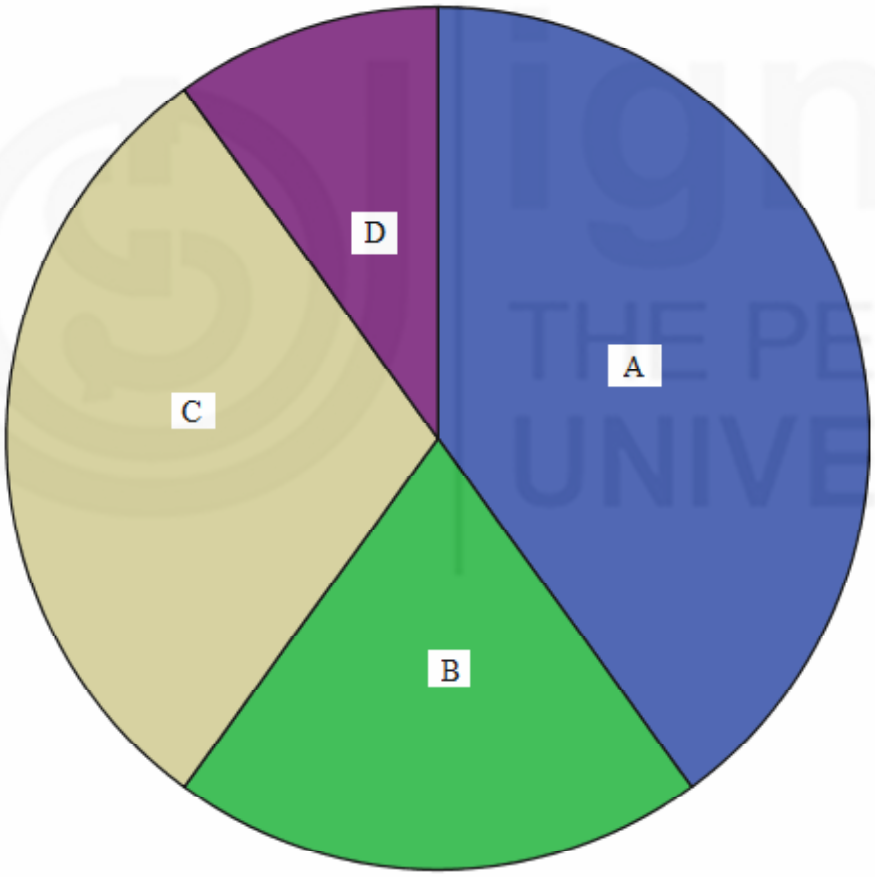


Fig.4.7: Pie diagram representing placement in four companies

4.5.5 Pictograms

It is known as cartographs also. In pictogram we used appropriate picture to represent the data. The number of picture or the size of the picture being proportional to the values of the different magnitudes to be presented. For showing population of human beings, human figures are used. We may represent 1 Lakh people by one human figure. Pictograms present only approximate values.

Self Assessment Questions

1) Explain the following terms:

   i) Frequency polygon

   .....

   .....

   .....

   ii) Bar diagram

   .....

   .....

   .....

   iii) Subdivided bar diagram

   .....

   .....

   .....

   iv) Multiple bar diagram

   .....

   .....

   .....

   v) Pie diagram

   .....

   .....

   .....

4.6 LET US SUM UP

Data collected from primary sources are always in rudimentary form. These unsystematic raw data would fail to reveal any meaningful information to us. To draw conclusions these data must be arranged or organise in a standard way. This can be done with the help of classification. There are various types of frequency distributions e.g., relative frequency distribution, cumulative frequency distribution, cumulative relative frequency distribution.

After classifying the raw data, its good presentation is also equally important. A good presentation enables us to highlight important features of the data and make fit for comparison and further statistical analysis. This can be achieved through statistical table, histogram, frequency polygon, cumulative frequency curve. Bar diagram, sub-divided bar diagram, multiple bar diagram, and pie diagram are also used for diagrammatic presentation of statistical data.

4.7 UNIT END QUESTIONS

1) What do you mean by classification? Discuss its various methods with suitable examples.

2) Following are the marks obtained by 30 students of psychology in their annual examination. Classify them in a frequency table.

30, 35, 36, 35, 19, 25, 63, 50, 32, 58, 55, 28, 43, 19, 40, 51,

56, 15, 14, 31, 56, 62, 22, 46, 52, 17, 54, 37, 16, 50.

3) Construct a "less than" cumulative and " more than" cumulative frequency distribution from the following data.

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	7	9	12	8	13	5

4) State different parts of of a statistical table.

5) Distinguish between classification and tabulation.

6) Prepare histogram and frequency polygon from the following table

Class Interval :	0-10	10-20	20-30	30-40	40-50
Frequency	5	9	16	14	6

7) Differentiate between the following pairs of terms

i) Histogram and bar diagram

ii) Frequency polygon and cumulative frequency curve

iii) Sub-divided bar diagram and multiple bar diagram.

4.8 GLOSSARY

Abscissa (X-axis) : The horizontal axis of a graph.

Array : A rough grouping of data.

Bar diagram : It is thick vertical lines corresponding to values of variables.

Body of the Table : This is the real table and contains numerical information or data in different cells

Caption : It is part of table, which labels data presented in the column of table.

Classification : A systematic grouping of data.

Continuous : When data are in regular in a classification.

Cumulative frequency distribution : A classification, which shows the cumulative frequency below, the upper real limit of the corresponding class interval.

Data : Any sort of information that can be analysed.

Discrete : When data are counted in a classification.

Exclusive classification : The classification system in which the upper limit of the class becomes the lower limit of next class.

Histogram : It is a set of adjacent rectangles presented vertically with areas proportional to the frequencies.

Frequency distribution	:	Arrangement of data values according to their magnitude.
Frequency Polygon	:	It is a broken line graph to represent frequency distribution.
Inclusive classification	:	When the lower limit of a class differs the upper limit of its successive class.
Ogive	:	It is the graph of cumulative frequency
Open-end distributions	:	Classification having no lower or upper endpoints.
Ordinate (Y-axis)	:	The vertical axis of a graph.
Pictogram	:	In pictogram data are presented in the form of pictures.
Pie diagram	:	It is a circle sub-divided into components to present proportion of different constituent parts of a total
Primary data	:	The information gathered direct from the variable.
Qualitative classification	:	When data are classified on the basis of attributes.
Quantitative classification	:	When data are classified on the basis of number or frequency
Relative frequency distribution	:	It is a frequency distribution where the frequency of each value is expressed as a fraction or percentage of the total number of observations.
Secondary data	:	Information gathered through already maintained records about a variable.
Stub	:	It is a part of table. It stands for brief and self explanatory headings of rows.
Tabulation	:	It is a systematic presentation of classified data in rows and columns with appropriate headings and sub headings.

4.9 SUGGESTED READINGS

Asthana, H. S. and Bhushan, B. (2007). *Statistics for Social Sciences* (with SPSS Application). Prentice Hall of India, New Delhi.

Yale, G. U., and M.G. Kendall (1991). *An Introduction to the Theory of Statistics*. Universal Books, Delhi.

Garret, H. E. (2005). *Statistics in Psychology and Education*. Jain publishing, India.

Nagar, A. L., and Das, R. K. (1983). *Basic Statistics*. Oxford University Press, Delhi.

Elhance, D. N., and Elhance, V. (1988). *Fundamentals of Statistics*. Kitab Mahal, Allahabad.

Sani, F., and Todman, J. (2006). *Experimental Design and Statistics for Psychology*. A first course book. Blackwell Publishing.

UNIT 1 CONCEPT OF CENTRAL TENDENCY

Structure

1.0 INTRODUCTION

In the day to day situations you must have heard that the average height of the Indian boys is 5 feet10 inches. The average longevity of Indians has now increased the average number of working women have gone up in the last ten years. Ever wondered what this 'average' is all about. This is nothing but the measure of central tendency. Tests and experiments and survey studies in psychology provide us data mostly in the shape of numerical scores. In their original form these data have little meaning for the investigator or reader. The obtained data are complicated and scattered and no inference can be arrived at, unless they are organised and arranged in some systematic way.

The classifications and tabulation makes the data easy and clear to understand. In the previous units you have learned about how you can classify and organise the data as well you had learned about presentation of data in the forms of graph. But you want to describe a data. A useful way to describe a data as a whole is to find a single number that represents what is average or typical of that set of data. This can be obtained by way of measures of central tendency, because it is generally located toward the middle or center of a distribution, where most of the data tend to be concentrated.

In this unit we will learn about measures of central tendency. There are many measures of central tendency, we will focus on only the three most commonly used in psychology, i.e. arithmetic mean, median and mode. We will learn about their properties and their limitations.

1.1 OBJECTIVES

After reading this unit, your will be able to understand:

- Describe the meaning of measures of central tendency;
- State the functions of measures of central tendency;
- Name the different types of measures of central tendency; and
- Explain the properties and limitations of mean, median, and mode.

1.2 MEANING OF MEASURES OF CENTRAL TENDENCY

The term central tendency was coined because observation (numerical value) in most data show a distinct tendency of the group to cluster around a value of an observation located some where in the middle of all observations. It is necessary to identify or calculate this typical central value to describe the characteristics of the entire data set. This descriptive value is the measure of central tendency and methods of compiling this central value are called *measures of central tendency*.

According to English & English (1958) "Measure of a central tendency is a statistic calculated from a set of distinct and independent observations and measurements of a certain items or entity and intend to typify those observations.

According to Chaplin (1975) "Central tendency refers to the representative value of the distribution of scores".

If we take the achievement scores of the students of a class and arrange them in a frequency distribution, we may sometimes find that there are very few students who either score very high or very low. The marks of most of the students lie somewhere between the highest and the lowest scores of the whole class. This tendency of the data to converge around a distribution, named as *central tendency*.

1.3 FUNCTIONS OF MEASURES OF CENTRAL TENDENCY

Measure of central tendency *provides us a single summary figure* that best describes the central location of an entire distribution of observation.

It is helpful in reducing the large data into a single value. For example , it is difficult to know an individual family's need for water during summer in cities. But the knowledge of the average quantity of water needed for the entire population of a city, will help the water works departments in planning for water resources.

In psychology we draw the representative sample from the population and information are gathered regarding different attributes. *The mean of the sample provide us the idea about the mean of the population*.

These measures are helpful in making decisions. For example the education department may be intended to know the average number of boys and girls

enrolled for primary classes. The sales manager wants to know the average number of calls made per day by salesman in the fields. Measures of central tendency are valuable in estimating such averages and planning in different fields.

Since these measures provide one single value that represent data *they facilitate comparison*.

Such comparisons can be made between two groups, two conditions at the same time or two conditions over a period of time. For example, researcher may be interested in whether video games can improve mental and physical functioning in the elderly.

In one study, a psychologist administered an IQ test to 11 volunteers in their sixties and seventies, then gave them 30 minutes to play on a video game twice a week for two months and then tested them again. The average IQ was 101.8 before the two months and 108.2 afterwards (Drew & Waters 1985). The average suggests that video game improve mental functioning.

Similarly for example an investigator wants to study whether the linguistic ability of 8 years old girls is better than the linguistic ability of the 8 years old boys. The investigators administered a test for the measurement of linguistic ability. Suppose the mean scores for girls was found to be 42.9 and mean scores for boys was found to be 38.7. Then on the basis of this mean score we can conclude that the 8 year old girls are better in linguistic ability in comparison to 8 year old boys.

Self Assessment Questions

1) What do you understand by the term measures of central tendency.

……..
……..
……..
……..
……..

2) What are the functions of measures of central tendency.

……..
……..
……..
……..
……..

3) What are the characteristics of a good measure of central tendency.

……..
……..
……..
……..
……..

Definitions of Mean, Mode and Median

In statistics, we use a term called statistical distribution. This actually tells us how a group of data is distributed in a population. For instance if you want to know amongst the population of India, how many are male children, how many female children, how many adult males and how many adult females, all these we can represent in the data in a statistical distribution in terms of actual numbers or in terms of percentage. Which ever way we do, as we look at the graph, we will know how the males and females are distributed across the population. In this statistical distribution, one can describe the properties of the distribution in terms of mean, median, mode, and range. Thus in statistics, a distribution is the set of all possible values for terms that represent defined events.

In statistical distributions, there are actually two types, viz., (i) the discrete random variable distribution and (ii) the continuous random variable distribution.

The discrete random variable distribution means that every term in the distribution has a precise, isolated numerical value. An example of a distribution with a discrete random variable is the set of results for a test taken by a class in school. The continuous random variable distribution has generally values within an interval or span. To give an example, scores of test taken by 5 students are 45,55,59,50,40. These are independent scores and called discrete scores. If the same scores are given as 45-47, 48-50, 51-53,54-56, 57-59, then this is called continuous random distribution. In the latter continuous distribution, a term can acquire any value within an unbroken interval or span. Such a distribution is also called a probability density function, which is used in weather forecasts.

Mean

As mentioned in the introduction mean is the average of all the scores in a discrete or continuous distribution. Method of calculation of this mean is different for discrete distribution and the continuous distribution. This average so calculated is called the Mean or mathematical average. It is easier to calculate mean from discrete data by adding up all the scores of X1 to Xn and divide by X1 + X2+… / Xn. On the other hand to calculate mean from continuous distribution a formula has to be used and generally there are more than one method to calculate the mean from continuous distribution.

Median

The median is the midpoint of a series of data. If it is a discrete data and having even number of scores (5,7,9, etc.) then the middle item leaving equal number of scores below and above the middle item is considered to be the median. If the discrete data is of even number, then the middle two items have to be added and divided by two to get the median. (The calculation of the median will be taken up in another unit as here we discuss the clear concept about the measures of central tendency). From a continuous distribution also median can be calculated but as mentioned in the case of calculation of mean here too certain formula has to be applied. Median thus is the 50^{th} % igtem in a series whether it is discrete or continuous.

Mode

When the data is arranged in a frequency, that is, for example, all the test scores obtained by 15 students in English, certain scores like 55 marks out of 100 may appear 4 times, that is 4 students would have scored 55. (For example, 45,**55,**43,44,**55**,45,65,63,67,**55**,,41,42,**55**,67.) The remaining have scored different, marks in the test. At one glance we could see that the largest number of times a score appears is 55 that is by 4 students while all other scores appear only once). Thus the mode here is 55 which has appeared the largest number of times or we can say 55 has the largest frequency of 4 students getting that mark in English. Thus one may state that the mode of a distribution with a discrete random variable is the value of the term that occurs the most often. It should also be kept in mind that there could be 4 students getting marks of 65. Thus 55 and 65 both have a frequency of 4 students each and both can be considered as the Mode. Thus it is possible that for a distribution with a discrete random variable can have more than one mode, especially if there are not many terms. A distribution with two modes is called bimodal. A distribution with three modes is called trimodal. The mode of a distribution with a continuous random variable is also calculated with the help of a formula, which will be taken up in another unit.

Self Assessment Questions

1) What are the types of measures of central tendency?

……..

……..

……..

……..

2) Define mean and put forward the properties of the mean. What are its limitations?

……..

……..

……..

……..

3) Define median and discuss its properties and limitations.

……..

……..

……..

……..

4) What is mode ? How will you identify mode. Discuss its properties and limitations.

……..

……..

……..

……..

Range

Range is defined as the difference between the lowest and the highest scores or values in a series. The series could be a discrete random variable series or continuous random variable series. In either case the range is difference between the lowest value and the highest value.

1.4 CHARACTERISTICS OF A GOOD MEASURES OF CENTRAL TENDENCY

A central tendency to be considered a good measure must be:

1) Rightly defined
2) Simple to calculate
3) Easy to understand
4) Based on all the observations
5) Should be least affected by fluctuation in sampling

Let us take up one by one and see what these mean.

1) It should be rigidly defined. This means that the definition of the measure should be so clear that everyone interprets the measure in the same manner.

2) The definitions of measure of Central tendency should be so clear that it should lead to one interpretation by all persons.

3) A measure to be considered good should be possible to calculate in a simple manner. Too many complex and high calculations will not make the measure a good one.

4) Whatever the calculation the resulting measure of central tendency, irrespective of it being mean, mode or median must be possible to easily understand what it conveys.

5) In order to be a good measure, the central tendency should be based on all observations. That is it should take into consideration all the scores. For instance if X1,X2,X3,X4,X5 have scored 15, 18, 20, 23, 24 respectively out of 25 marks in English test, all these marks should be taken as it is in the case of Mean. If any mark is left out say the extremes, that is 15 and 24, then the measure of central tendency that results would show 20 which is in fact not the correct measure.

Sampling is a term used when we take a sample from a population for our study. For instance, amongst class 240 class 3 students, if we take only 24 that is $1/10^{th}$ of the total class 3 students, this 24 is the sample. If we take these 24 randomly then this sample will be called random sample which is considered to represent the population from each of these students has been drawn. This sampling if not correctly selected or is defective in some way, then it may be subjected to fluctuations. For instance in some cases only the best students may be selected or only the worst and so on and some students will be totally left out as they are not good performers and so on. These fluctuations should not affect the measure of central tendency if it is a good measure. In other words, if we select the two samples randomly from the same universe or population the value of average for both of them should be near to each other. The differences in the averages of two samples drawn from the same population are technically known as 'fluctuation of sampling'.

1.5 TYPES OF MEASURES OF CENTRAL TENDENCY

There are many measures of central tendency. We will consider only the three most commonly used in psychology: The Mean, Median and Mode.

1.5.1 The Mean

This, is the most commonly used measure of central tendency. There are different types of means, that is, Arithmetic mean, Geometric mean and Harmonic mean, but we will focus only on arithmetic mean in this unit.

The arithmetic mean is the sum of all the scores in a distribution divided by the total number of scores.

M is the symbol for the Mean. M is used in research articles in psychology and are recommended by the style guidelines of the American Psychological Association (2001).

It should be noted that the Greek letter (μ) pronounced as *mue* is used to denote the mean of the population and X pronounced as "X-bar" or M is used to denote the mean of a sample.

Let us now look at the properties of these measures of central tendencies.

1.5.1.1 Properties of the Mean

The mean is responsive to the exact position of each score. In the next unit when we will learn how to compute the mean you will see that increasing or decreasing the value of any score changes the value of the mean.

The mean is sensitive to the presence (or absence) of extreme scores. For, example. The means of the scores 6,7,9,8 is 7.5

6+7+9+8 /4 = 7. 5

But the mean of scores 6,7,9,22 is

6+7+9+22 / 4 = 11

It means that one extreme value in a series of values, can make drastic changes in the mean values.

When a measure of central tendency should reflect the total of the scores, the Mean is the best choice because the Mean is the only measure based on the total scores. For example, if a teacher wants to see the effect of training on pupils performance then the teacher can get the mean scores of each student, before and after the training and compare the mean scores obtained before and after the training. The difference will give an idea of the effect of training on the students. If the difference is great then one may be able to state that the training had a good effect on the students.

The mean will best suit this kind of problems because the teacher is interested in knowing the usefulness of the training for the students. Similarly insurance company expresses life expectancy as a Mean, because on the basis of this, companies can come to know the total income from policy holders and total pay off to survivors.

When we have to do the further statistical computations, the mean is the most useful statistic to use.

As you will see in the next unit, Mean is based on arithmetic and algebraic manipulations, so it fits in with other statistical formulas and procedures. When we have to do further calculations, mean is often incorporated implicitly or explicitly.

1.5.1.2 Limitations of the Mean

One of the most important limitation of the mean it is too sensitive to the extreme scores. As mentioned earlier, in calculation of the Mean if any one score is extreme and all other scores are near each other, it would give a wrong idea about the average. Let us take an example, again of marks obtained by students of class 10 in a mathematics examination. Let us say the 10 students had scored the following: 55,65,45,35,45,25,50,40,45,100. This extreme score of 100 will affect adversely the Mean (for instance the M = 50.5. If we remove just that one score of 100, the mean will become 40.5. Thus the score of 100 has made enormous difference to the Mean, that is the measure of central tendency.

As has been pointed out, in the calculation of the Mean, every value is given equal importance. But if one extreme value is present in the series, then the value of mean as seen above becomes misleading.

Similarly let us say that a teacher wants to make comparisons between progress in performance of students of two sections in mathematics. Let us suppose that the average scores obtained by group A in first, second and third term is 60, 20 and 70; let us say the average scores obtained by group B is 45,50,55 in the three terms respectively, then the mean in both the cases is 50. (150 / 3 = 50). However, the lowest and highest scores are 20 and 70 in Group A and 45 and 55 in group B. That is the range is higher (50) in Group A as compared to B (range = 10). This means group B is more homogeneous than Group A in that their performance in the three terms vary far lesser in group B as compared to that in Group A. Group B's performance in the three terms appear more consistent than Group A. Hence it is important use the Mean, one of the measures of central tendency with caution, especially when there are extreme scores that may affect the Mean.

1.5.2 The Median

The median is another measure of central tendency.

According to Minium, King & Bear (2001) "Median is the value that divides the distribution into two halves."

According to Garrett (1981) 'When scores in a continuous series are grouped into frequency distribution, the median by definition is the 50% point in the distribution.'

If we arrange the items of series in ascending or descending orders of magnitude, the value of the central item in the series is the median. We may say that median is that value of the scores below which 50% of cases lie and also above which 50% of cases lie.

You should keep in mind that the central item itself is not the median but the value of central item, that is the *50th percent item* is known as the median. For example, if you arrange the marks of seven students in ascending order, for example, (3,5,7,9,11,13,15), then the marks obtained by the fourth student (Mr. X) will be known as the median of the scores of the groups of students under consideration. The 4th person has scored 9 and thus the median is = 9. That is there are 3 students below the student X and 3 students above the student X making the student the mid point.

1.5.2.1 Properties of the Median

Median is less sensitive to extreme values in the distribution. For example, the median of the value 6,7,9,12,16 is 9 and the median of the value 6,7,9,12,46 is also 9.

In some situations median is better than the mean as a representative value for a group of scores. For example, in the above example of scores, the mean of the first scores is 10 (6+7+9+12+16 / 5) and the means of the second scores is 16 (6+7+9+12+46 / 5) but the median is 9 in both the cases, which is much more representative of most of the scores.

An extreme score like this (46) is called an outlier. Outlier may be much higher or much lower than the other scores in the distribution. Where outliers are present, it would be better to use the Median as the central tendency measure.

1.5.2.2 Limitations of the Median

The median tells about how many scores lie below or above it but it does not tell about how far away the scores may be. A value that is slightly below the median or considerably below the mean— both count the same in determining the median. For example, in the scores 10, 25, 30, 49, 50, the Median is 30 but the difference between 25 and 30 is 5 and difference between 30 and 49 is 19. This means the median at times can be misleading when the data has a very wide range of scores with minimum at one extreme and maximum at another extreme.

Further more, another limitation of the Median is that the value of the Median is not based on each and every item of the distribution so it does not represent the complete data. Also we leave out most of the scores in the median taking only the mid point value. Thus it is not the representative of the sample.

1.5.3 The Mode

The literal, meaning of the word, ‘mode’ is frequent and "fashionable”. The word ‘Mode’ originates from the French word La Mode. According to Minium et al (1997) “Mode is the score that occurs with the greater frequency.”

According to Guildford (1965) “The mode is strictly defined as the point on the scale of measurement with maximum frequency in a distribution.”

On the basis of the above definition it can be said that mode is that value in a series which is most frequent. For example in a factory maximum number of labourers (out of 100, 80 labourers) earn Rs. 100 per day and those earning more than 100 or less than 100, is less than 80. Thus the mode wage of the factory is Rupees 100. To give another example, if inn the Maths test the

scores are as given below for 10 students, 35, 45, 42, 54, 42, 35, 42, 46, 42, 36, here 42 marks has been obtained by 4 students while all other marks has been obtained by lesser than this number of students. Thus the mode of marks is 42.

1.5.3.1 Properties of the Mode

The mode is easy to obtain. It can easily be identified many times by observation only.

The mode is the only measure of Central tendency which can be used for nominal level variables. For example, there are more Hindus in India than people of any other religion. Here, Hindu religion is referred to as the mode. No other measure of central tendency is appropriate for distribution of religion in India, as one can use mode to describe the most common scores in any distribution.

1.5.3.2 Limitations of the Mode

- Mode is not stable from sample to sample.
- It is affected more by sampling fluctuation.
- There may be more than two modes for a particular set of scores. For example, if the scores are 4, 9, 5, 6, 5, 4, 8, 7, 3, 10 here 4 and 5 both are mode as both these occur two times whereas all other numbers occur only once.

1.6 LET US SUM UP

A measure of central tendency provides us a single value, which represents the characteristic of the group. The mean, median and mode are most commonly used as measures of central tendency.

Mean can be computed for grouped and ungrouped data. This is the only measure of central tendency based on all the scores in the series.

Median is the score or value of that central item, which divides the series into equal parts. Median is the midpoint of the class intervals.

Mode is the most frequently occurring value.

1.7 UNIT END QUESTIONS

1) Given below are statements. Indicate in each case whether statement is true or false.

   i) Mean is most stable measure of central tendency T/F

   ii) The median is less affected than the mean by extreme values of observation is a distribution. T/F

   iii) If the number of observations is even, the median is in the middle of the distribution. T/F

   iv) Mode is not influenced by extreme values T/F

2) Fill in the blanks in the series

   i) The value of …… is affected by extreme data values.

   ii) The sign ……. is used for population mean and sign …… is used for sample mean.

   iii) The most frequently occurred value is the series is know as …….

   iv) While calculating ……. every individual item in the data is taken into consideration.

3) What do you understand by 'Central tendency'? Describes the functions of measures of central tendency.

4) Under what condition median is most suitable than other measures of central tendency?

5) What are the characteristics of a good measure of central tendency?

Ans. 1 i)T, ii)T, iii)F, iv)T

2) i) mean, ii) mue and M iii) mode iv) mean

1.8 GLOSSARY

Measures of Central Tendency : Measure that describes the center of the distribution. The mean, median and mode are three measures of central tendency.

Arithmetic Mean **:** A measure of Central tendency calculated by dividing the sum of observation by the number of observation in the data set.

Median **:** The value of the middle item in the data set arranged in ascending or descending order. It divide the data into two equal parts.

Mode **:** The most occurring value or the value that has the maximum frequency.

1.9 SUGGESTED READINGS

Chaplin, J.P.(1975). *Dictionary of Psychology,* Alaurel original

Drew, B., & Waters, J.(1985).Video games :utilization of a novel strategy to improve perceptual –motor skills in non-institutionalized elderly . Proceeding and Abstract of the Eastern Psychological Association 5,56.

English, H.B. & English, A.C (1958). *A Comprehensive Dictionary of Psychological and Psycho Analytical Terms , New York :*Longmans Green

Garrett , H.E. (1981)*Statistics in Psychology and Education*, Bombay, Vakils, Feffer and Simons Ltd.

Guilford, J.P(1965).*Fundamental Statistics for Students of Psychology and Education* (4th ed),New York :McGarw-Hill

UNIT 2 MEAN, MEDIAN AND MODE

Structure

2.0 INTRODUCTION

When we conduct research in psychology we use statistical methods to analyse the data. To decide which statistical method to use for which purpose is an important decision to make in research while analysing the data. For example we obtained the large amount of data. We organise the data in a tabular form. Now we are interested in obtaining one single value which represents the total group of data. There are various methods which can be used for this purpose, but we are not sure which method will be most suitable for our data and for our purpose, until and unless we have conceptual knowledge regarding the measures which suits to our purpose. Once we have the knowledge then, we can select suitable measure, can find out the values and can interpret the results. In the last unit, we discussed properties of different measures of the central tendency, in this unit, we are going to learn procedure or methods of calculating these measures of central tendency.

2.1 OBJECTIVES

After reading this unit, you will be able to:

- Calculate mean for ungrouped data;
- Find out the mean for grouped data by long and short method;
- Work out the median for grouped and ungrouped data;

- Compute median when frequencies are missing;
- Find out the mode for grouped and ungrouped data; and
- Select the most appropriate method of central tendency for a given set of data.

2.2 SYMBOLS USED IN CALCULATIONS OF MEASURES OF CENTAL TENDENCY

Let use get familiar with symbols which we will use in Calculating measures of Central tendency. Some of the symbols that you should know are given below:

∑= Sum of (Add are the score)
N= The total number of observation in the
X= Raw scores.of X1….Xn
M = Mean of the sample
∑x = the sum of X

2.3 THE ARITHMETIC MEAN

2.3.1 Calculation of Mean from Ungrouped Data

Example 2.1 An academic achievement test was administered on 10 students and they obtained following (hypothetical) scores.

8,7,4,4,7, 5, 6, 9, 2 ,8 (X1, X2,….X10)

{The formula for calculating mean for ungrouped data is} $M = \sum X / N$

$\sum X$ Sum of X1 + X2+ X3 +…X10

$M = \sum X / N$

N= Total number of students.

Calculation: The mean will be

8+7+4+4+7+5+6+9+2+8 = 60

$M = \sum X / N = 60 / 10 = 6$

Mean = 6

Steps involved in computing means for ungrouped data are given below:

1) Add up all the scores of all the students.
2) Divide this sum by the number of students whose scores have been added.

2.3.2 Calculation of Mean from Grouped Data by Long Method

When we have large data it would take a long time adding up all the scores of all the students. Let us say there are 1000 students whose scores are given, and to physically or manually add these scores is rather difficult. So we group them into smaller divisions and then try to find the Mean for this. The manner in which we group the data is as given below.

Take the lowest score and the highest score

Decide how many categories of the data you want.

Let us say you want to have 5 groups in all and the data (marks) ranges from 30 to 90 for a total of 1000 students .

Now work out the difference between the lowest and the highest score which is = 60.

You need 5 categories, that means we need 60 / 5 = 12 class intervals in each category.

Now we group the data into a frequency distribution, as follows,

Categories of marks	No. of students(f)	Mid point (X)	fX
30-41	250	36	9000
42-53	200	48	9600
54-65	300	60	18000
66-77	100	72	7200
78-89	150	84	12600
Total	1000 students.		∑fX = 56400

In the above the marks of 1000 students have been grouped into 5 categories . In each category the class interval (the difference between low and high scores is 12). In this you can also have 30-40, 40-50, 50-60, 60-70, 70-80 and 80-90. You have to see how many students fall in each category, that is the frequency for each category.

Calculation of Mean by long method:

Mean = ∑fX / N

Mid point for each class interval = X

N=The total number of observation =1000

∑fX =Sum of the midpoints weighted by their frequencies.= 56400

Let us now calculate the Mean

Mean = ∑fX / N 56400 / 1000 = 56.4.

Thus the Mean marks obtained by this group of students = 56.4.

One more example is given below to make the matters clear and how to calculate the Mean

Example 2
Calculation of Mean by long method.

Class interval	Mid Point (X)	F	f(X)
195-199	197	1	197
190-194	192	2	384
185-189	187	4	748
180-184	182	5	910
175-179	177	8	1416

170-174	172	10	1720
165-169	167	6	1002
160-164	162	4	648
155-159	157	4	628
150-154	152	2	304
145-149	147	3	441
140-144	142	1	142
	N = 50		∑fX=8540

M = ∑fX / N
N = 50
∑fX = 8540 / 50
M = 170.80

Steps

- Find out the midpoint of each class interval (X)
- Add all the number of scores (N)
- Multiply X values by the respective frequency the opposite to it (fX)
- Add all the fX (∑fX)
- Finally divide ∑fX by N
- Thus in the long method we are multiplying the frequencies with the mid points. When the data is large the calculationof mena becomes cumbersome and hence we employ shorter method called the calculation of mean by short method.

2.3.3 Calculation of the Mean for Grouped Data by Short Method

In the table given above, the Mean was calculated by long method. But sometimes we have to handle large numbers or midpoints whose values are in points, then further calculations becomes tedious, and hence the short method has been devised for calculating the Mean.

Remember that the short method does not apply to the calculation of the median or mode.

The most important point to remember in calculating the means by short method is that we assume a mean at the outset. There is no set rule for assuming a mean. The best plan is to take the midpoint of an interval somewhere near the center of the distribution or midpoint of that interval which contains the largest frequency.

The following formula is used to calculate means by short method or assumed means method.

AM = Assumed Mean
x' = (X – AM) the difference between the assumed mean and the actual scores of each category.

∑fx' = the difference obtained weighted with the frequency of that category or interval.

i = class interval

M= AM + ∑fx' × i.

N = the no. of cases that is the total number of subjects.

The use of this formula can be easily understood through the following illustration:

Table: Calculation of Mean by short method with Assumed Mean

Class interval	Mid Point (X)	f	x'	fx'
195-199	197	1	5	5
190-194	192	2	4	8
185-189	187	4	3	12
180-184	182	5	2	10
175-179	177	8	1	8
170-174	172	10	0	0
165-169	167	6	-1	-6
160-164	162	4	-2	-8
155-159	157	4	-3	-12
150-154	152	2	-4	-8
145-149	147	3	-5	-15
140-144	142	1	-6	-6
	N = 50	50		-55

In the above table the class interval 170-174 has the highest frequency of 10 and so we can assume the mean to be 172.

From 172 we go up by one point each , then we get 1,2,3,4,5.

As we go down by one point each from 172 we get -1, -2, -3, -4, -5, -6

Thus we have f the frequency and x' the difference in the scores by 1 point difference towards up and down the AM.

Now to calculate f x' = we have 5, 8, 12, 10 , 8, 0all these are positive as they are taken up from the AM. Sum of these fx' = 43

The other set we have of fx' is the minus scores as we go down the AM and these are -6,-8,-12,-8,-15,-6, Sum of these fx' = -55

The ∑fx' = -55 +(43) = -12.

Now the correction has to be inserted as we took above and below mean 0,1,2,3, etc., which inn the real sense would have been the actual differenced. We had taken the difference without the class interval of 5 and so we have to

add the correction here so that we know how much distance from the assumed Mean is the actual mean.

For this correction we take the ∑fx' and multiply by class interval of 5 and divide by the total number 50

∑fx' x i = -12 x 5 = -60 / 50 = -1.2.

Actual Mean = AM + correction

= 172 -1.20 = 170.80

Steps

- Find out mid point of each class interval (X)
- Assume one value as mean In table the largest f is on intervals 170-174, Which also happens to be almost in the centre of the distribution, 172 in taken as AM.
- 3.Find out the difference between mid point and assumed mean and divide it by class interval x) eg. 177-172 = 5 /5 = 15
- Multiply each x by f respective frequencies (fx').
- Find the algebraic sum of the plus and minus fx' divide this sum by N. some time ∑fx' will be positive and sometimes negative.
- Multiply this value by class interval. This gives the correction to be applied to the Assumed Mean
- The assumed Mean + correction = the Actual Mean.

Now that we have learnt the long and short methods of calculation of mean, let us do a sum here.

Scores	freq.	Mid point	x'	fx'
10-19	10	15	2	20
20-29	8	25	1	8
30-39	6	35	0	0
40-49	4	45	-1	-4
50-59	12	55	-2	-24
60-69	10	65	-3	-30
Total	50			∑ fx'= -30

Correction = ∑ fx' /N × i = -30 / 50 × 10 = -6

Assumed Mean = 35. Correction = 6

Actual Mean = 35-6 = 29

2.4 THE MEDIAN

2.4.1 Computation of Median for Ungrouped Data

For ungrouped data two situations are their in the calculation of the *median*.

a) When N is *odd*, and

b) When N is *even*.

Where N is odd median can be calculated by formula

Mdn= $(n+1)/2^{th}$ item

Example: Suppose we have the following scores,

17,14,15,25,44,32,30

First we will arrange the above scores in ascending order

14,15,17,25,44,32,30

Here N=7(N+1)/2th =4 the 4th item i.e. 25 is the median. The scored 25 has lies in the middle of the series three scores lies above and three score lies below 25.

When the N is *even* numbers of scores,

When N is even the medians can be calculated by following formula

The value of (N/2) the item the value of (N/2) + 1]the item

Mdn = 2

Example: 14,15,17,25,35,47.50,,54,

The scored of the (N/2)th i.e. 4th is 25

The score of the (N/2) +1]th 5th is 35

The median is

25+35 /2 = 60 / 2 = 30

Mdn = = 30

2.4.2 Calculation of the Median for Grouped Data

The formula for calculating the median when the data are grouped in class interval is

Mdn = l + {((N/2)- F) / fm } × i

Where :

l= exact lower limit of the class intervals within which the Mdn lies.

N/2= One half the total number of scores

F= Sum of the scores on all intervals *below l*

fm =Frequency *within* the interval upon which the median falls.

i=length of the interval

The use of the formula can be illustrated by following example.

Example: Calculation of Median from Grouped data

Table:

Class interval	F	Cumulative frequency
195-199	2	50
190-194	3	48
185-189	4	45
180-184	4	41
175-179	5	37
170-175	10	32
165-169	6	22
160-165	5	16

155-159	4	11
150-154	4	7
145-149	2	3
140-144	1	1

Let us apply the formula to derive median

Mdn = *l*+ (N/2)- F) / fm x i

Where

l=169.5, N/2 50/2=25 F = cum. Freq = 22 fm = 10 i=10

169 + 25-22 / 10 x 10

169 + 3/10 x 10 = 3

The Median = 169 = 3 = 172.

Steps

To locate the median we take 50% i.e. N/2 of our scored and count into the distribution until the 50% point is reached. In the above example there should 25 scores above and 25 scores below the median. Start adding the frequencies (+) from below we discover that 25 lies in the class interval 170-174. If we add the frequencies from above to below then again 25 lies on the class intervals 170-174.

Find the lower limit of the calls interval on which Mdn falls. The lower limit of class interval 170-174 is 169.5

To find out F begin adding the frequencies from the below and count off the scores in order up to interval which contain median. The sum of these score is F.

Compute N/2-F. Divide this quantity by the frequency on the interval which contain the median (fm) and multiply it by the size of the class interval (i).

Add the amount obtained by the above calculations to the exact lower limit (l) of the intervals which contains the Mdn.

Calculation of the Mdn when the frequency distribution contains gaps

Example: Calculation of the median when these are gaps in the distribution.

Class Interval	F
20-21	2
18-19	1
16-17	0
14-15	0
12-13	2
10-11	0

8-9	0
6-7	2
4-5	1
2-3	1
0-1	1
N	10

(N/2)=5

F = 5

((N/2)- F) = 5-5 = 0

Mdn = l + N/2-F / fm x i

fm = 2

9.5+(0/2) =9.5

Steps

Find N/2 Since N=10 and N/2=5.

Count up the frequency from below we find that 5 cases lie upto the class interval 6-7. By adding the frequencies from above, we find that 5 cases lie upto class interval 12-13. The median should then fall mid-way between the two classes 8-9 and 10-11. IT should be the common scores represented by both these classes. This will be 9.5 the upper limit of 8-9 and lower limit of 10-11. Computing from the two ends of the series now give the consist results.

2.5 THE MODE

2.5.1 Computation of Mode for Ungrouped Data

Mode can easily be computed by looking at the data. In ungrouped data mode is that single score which occurs most frequently.

Example

If we have to find out the value of the mode from the following scores.

25,25,29,29,29,30,32,36

Here the scores 29 is repeated maximum number of times therefore 29 is the "crude' mode.

2.5.2 Calculation of Mode for Grouped Data

When data is available in the form of frequency distribution then, we differentiate between *crude* mode and "*true*" mode.

Crude mode is the point of greatest concentration in the distribution. For example in the frequency distribution given in table. The class interval 170-174 contain the largest frequency and 172 is the midpoint therefore 172 is the 'crude; mode.

The '*true*' mode can be obtain by the following formula

Mode-= 3 Mdn-2Mean

Md = Median

M= Mean

Therefore if the mean is 170.80 and median is 171.00 the mode will be

Mode 171 x 3-170.80X 2 =174.40

2.6 WHEN TO USE THE VARIOUS MEASURES OF CENTRAL TENDECY

Now you have the knowledge of how to compute mean, median and mode. But sometimes we got puzzled what measures of central tendency is most appropriate for particular problem. Certain general rules for decision is as follows.

2.6.1 When to Use the Mean

1) Mean has the greatest stability. Therefore when the measure of central tendency having the greatest stability is required .
2) When other statistics (e.g. SD or Coefficient of Correlation) are to be commutate later on.
3) When the scores symmetrically fall around a central point i.e. when the distribution is normal distribution.

2.6.2 When to Use the Median

When the exact midpoint of the distribution is required.

When there are extreme scores in the distribution.

When it is desired to be found the position of an individuals score in terms of its distance from the mid point.

2.6.3 When to Use the Mode

1) When a quick and approximate measure of central tendency is required.
2) When we have to know the most often recurring value. For example when We have to describe the most common style of the dresses worn by girls.

2.7 LET US SUM UP

In this unit we learn about the three important measures of central tendency, that is the mean, mode and the median, their concepts and definitions. We then took up the mean, and learnt about the calculation of the mean from both grouped and ungrouped data. This was followed by defining the median and learning to calculated the median from ungrouped and grouped data. We also learnt how to calculate median when there are gaps. We then took upin the next section the mode and learnt about its definition and then learnt how to calculate the mode. From these we also came to know when to use these three measures of central tendency, when to use the mean, median and the mode and what would be more appropriate etc. Then we worked out all the formulas and gave them together in this unit so that it is easy to remember.

2.8 UNIT END QUESTIONS

1) An intelligence test was administered on 8 students, they obtained following scores on the test, find out the mean intelligence score.

   80,100,120,105,90,110,115,112

2) Compute the medians for the following data
   a) 12,16,20,5,19,36,15
   b) 72,80,84,60,69,54

3) Compute the mode for the following data
   a) 18,12,14,16,18,13
   b) 4,4,7,12,9,12,4,7

4) Find the mean, median and mode for the following scores.
   a) 22,21,24,18,19,23,12,20
   b) 9,8,13,10,11,10,12,10,14

5) Compute the mean, median and mode for the following frequency distribution:

a)

Class Interval (Scores)	f
70-71	2
68-69	2
66-67	3
64-65	4
62-63	6
60-61	7
58-59	5
56-57	1
54-55	2
52-53	3
50-51	1

b)

Class Interval (Scores)	f
100-109	5
90-99	9
80-89	14
70-79	19
60-69	21
50-59	30
40-49	25
30-39	15
20-29	10
10-19	8
0-9	6

6) In the situation describe below, what measures of central tendency would you like to compute?

a) The average intelligence of a class.

b) The most popular dress of teenagers

c) Determine the midpoint of the scores of a group in and examination

7) "Every measure of central tendency has its own particular characteristics. It is difficult to say which measure is best". Explain with example

Answer

1) 104

2) a) 16, (b), 70.5

3) a) 18 (b) 4

4) a) Mean-19.87 Mdn 20.5 Mode 18.61

   b) Mean 10.8 Mdn 10 Mode- 10.0

5) a) Mean 60.76 Mdn=60.79 Mode= 60.85

   b) Mean= 55.43 Mdn= 55.17, Mode- 54.65

6) a) Median-14.5

   b) Mode

   c) Median

2.9 LIST OF FORMULA

Mean for ungrouped data = $\sum X / N$

Mean for grouped data by long method = $\sum f X / N$

Mean for grouped data by short method = AM + {fX/N × i

Median for ungrouped data

When N is odd Mdn = the value of $(n+1)/2)^{th\ item}$

N is the number of items in the odd array.

When N is even

The value of (N/2) the item the value of (N/2) + 1]the item

Median for groped data =

Mdn = l + ((N/2)- F) / fm × i

Mode = 3 Mdn-2 Mean

2.10 SUGGESTED READINGS

Pagano, R. (2004). *Understanding Statistics in the Behavioural Sciences (7th edition)*. Pacific grove, ca: brooks/cole publishing co.

Guilford, J.P...(1956). *Fundamental Statistics in Psychology and Education*. Mcgraw-hill book company. NY

Garrett, E.H. (1969). *Statistics in Psychology and Education* Greenwood Press, NY.

UNIT 3 CONCEPT OF DISPERSION

Structure

3.0 INTRODUCTION

In this unit we will be dealing with the concept of dispersion. Dispersion actually refers to the variations that exist within and amongst the scores obtained by a group. We have seen how in average there is a convergence of scores towards a mid point in a normal distribution. In dispersion we try and see that how each score in the group varies from the Mean or the Average score. The larger the dispersion, less is the homogeneity of the group concerned and if the dispersion is less that means the groups are homogeneous. Dispersion is an important statistic which helps to know how far the sample population varies from the universe population. It tells us about the standard error of the mean and also gives us indication of the mean and standard deviations. In this unit we will understand the concept of dispersion, mean and standard deviation.

3.1 OBJECTIVES

After reading this unit, you will be able to:

- Understand the concept of dispersion;
- Explain the significance of measuring dispersion;
- Identify the differences between measures of central tendency and measures of dispersion; and
- Describe the main properties and limitation of the Range, Quartile deviation, Average deviation and Standard deviation.

3.2 CONCEPT OF DISPERSION

Dispersion in statistics means deviation of scores in a group or series, from their mean scores. It actually refers to the spread of scores in the group in relation to the mean. In other words in a group of 10 subjects who have scored differently on a mathematics test, each individual varies from the other in terms of the marks that he or she has scored. These variations can be measured and that is called the measure of dispersion. This measure refers to many methods of measuring dispersion or variations. It measures the dispersion of different values for the average value or average score. It means also the scatter of the values in a group, as for instance how the marks in mathematics test varies amongst the 10 students.

3.2.1 Functions of Dispersion

- It is used for calculating other statistics such as F-value, degree of correlation, regression etc.
- It is also used for comparing the variability in the data obtained as in the case of Socio economic status, income, education etc.
- To find out if the average or the mean/median/mode worked out is reliable If gthe variation is small then we could state that the average calculated is reliable, but if variation is too large, then the average may be erroneous.
- It is also used in time series to overcome the various fluctuations due to time factor, seasonal factors etc.
- Dispersion gives us an idea if the variability is adversely affecting the data and thus helps in controlling the variability.

We had seen in the earlier units that the measures of central tendencies (i.e. means) indicate the general magnitude of the data and locate the center of a distribution of measures. They do not establish the degree of variability or the spread out or scatter of the individual items and their deviation from (or the difference with) the means.

According to Nciswanger, there may be distributions which may have common mean, median and mode as well identical frequencies, however they may differ considerably in their scatter or in their values in regard to the average measures or measure of central tendencies.

According to Simpson and Kafka Average alone cannot give an idea about the data or its characteristic features, because it is not representative of a population unless how the individual items are scattered around that average.

For this we need further description of the particular series and if we are to gauge how representative the average is.

In order to understand the scatter, or variability known as dispersion, let us take the following three sets of data X, Y, Z.

Students	Group X	Group Y	Group Z
1	50	45	30
2	50	50	45
3	50	55	75
Total	150	150	150
Mean	50	50	50

As is observed in the above table, the three groups X, Y, and Z have the same mean of M=50. In fact the median of group X and Y are also equal. Yet if we state that the students from the three groups are of equal capabilities, or are of equal proficiency, it may be erroneous. This is so because as we look at the data we find that students in group X have equal marks of 50 each, and while one has obtained 50 equal to the mean, the other two vary by 5 points. In the third group Z, though the mean is 50, each individual subject varies a great deal from each other, as for example the first subject has only got 30 marks, which is 20 points below the mean, the second subject has obtained 45 which is 5 points less than the mean of 50, whereas the third subject has obtained 75 which is 25 points above the mean. Thus the variation is very high amongst the three subjects in the Z group. It is thus clear that the measures of the central tendency alone is not sufficient to describe the data.

The term dispersion is also known as the average of the second degree, because here we consider the arithmetic mean of the deviations from the mean of the values of the individual items.

3.2.2 Measures of Dispersion

In measuring dispersion, it is imperative to know the amount of variation (absolute measure) and the degree of variation (relative measure). In the former case we consider the range, mean deviation, standard deviation etc. In the latter case we consider the coefficient of range, the coefficient mean deviation, the coefficient of variation etc.

Measures of dispersion are descriptive statistics that describe how similar a set of scores are to each other. The greater the similarity of the scores to each other, lower would be the measure of dispersion. The less the similarity of the scores are to each other, higher will be the measure of dispersion. In general, the more the spread of a distribution, larger will be the measure of dispersion. To state it succinctly, the variation between the data values in a sample is called dispersion.

It is possible to have two very different data sets with the same means and medians. While measure of the central tendencies are indeed very valuable, their usefulness is rather limited. That is the measures of the middle are useful but limited in their usefulness. Hence one has to think of other measures too around the center and one such measures is the dispersion or variability about its middle. The most commonly used measures of dispersion

are the range, percentiles, and the standard deviation. The range is the difference between the largest and the smallest data values. Therefore, the more spread out the data values are, the larger the range will be. However, if a few observations are relatively far from the middle but the rest are relatively close to the middle, the range can give a distorted measure of dispersion.

In the previous two units you saw that measures of central tendency provided the central measure or the centre value that represent a set of scores as a whole. Although through these measures we can compare the two or more groups, a measure of central tendency is not sufficient for the comparison of two or more groups. They do not show how the individual scores are spread out. For example a math teacher is interested to know the performance of two groups of his students. He takes 10 weekly 40 point tests. The marks obtained by the students of groups A and B are as follows:

Test scores of Group A: 5,4,38,38,20,36,17,19,18,5

Test scores of Group B: 22,18,19,21,20,23,17,20,18,22

The mean scores of both the groups is 20, as far as mean goes there is no difference in the performance of the two groups. But there is a difference in the performance of the two groups in terms of how each individual student varies in marks from that of the other. For instance the test scores of group A are found to range from 5 to 38 and the test scores of group B range from 18 to 23.

It means that some of the students of group A are doing very well, some are doing very poorly and performance of some of the students is falling at the average level. On the other hand the performance of all the students of the second group is falling within and near about the average that is Mean = 20. It is evident from this that the measures of central tendency provide us incomplete picture of a set of data. It gives insufficient base for the comparison of two or more sets of scores.

We need in addition to a measure of central tendency an index of how the scores are scattered around the center of the distribution. In other words, we need a measure of *dispersion* or *variability*. In this unit we will study the measures of dispersion.

Whereas a measure of central tendency is a summary of scores, a measure of dispersion is summary of the spread of scores. Information about variability is often as important as that about the central tendency.

3.2.3 Meaning of Dispersion

According to Minium et al (2001), measures of variability express quantitatively the extent to which the score in a distribution scatter around or cluster together. They describe the spread of an entire set of scores, they do not specify how far a particular score diverges from the centre of the group. These measures of variability do not provide information about the shape of a distribution or the level of performance of a group.

Self Assessment Questions

1) Given below are statements, indicate whether statement is true or false.

   i) The dispersion in series indicates the reliability of the measure of central tendency. (T/F)

   ii) A major limitation of rang is that it ignore the large number of observation is a series. (T/F)

   iii) Mean Deviation is capable of further algebraic treatment. (T/F)

   iv) Standard deviation is the square of the variance of a distribution. (T/F)

   v) The difference between the largest and the smallest observation is known as Quartile deviation. (T/F)

Answer: i) F, ii) T, iii) F, iv) T, v) F

3.2.4 Absolute Dispersion and Relative Dispersion

Absolute dispersion usually refers to the standard deviation, a measure of variation from the mean, the units of standard deviation are the same as for the data.

Relative dispersion, sometimes called the coefficient of variation, is the result of dividing the standard deviation by the mean and it may be presented as a quotient or as a percentage. A low value of relative dispersion usually implies that the standard deviation is small in comparison to the magnitude of the mean. To give an example if standard deviation for mean of 30 marks is 6.0, then the coefficient of variation will be 6.0 / 30 = 0.2.(about 20%) If the mean is 60 marks and the standard deviation remains the same as 6.0, the coefficient of variation will be 6.0 / 60 = 0.1, (10%).

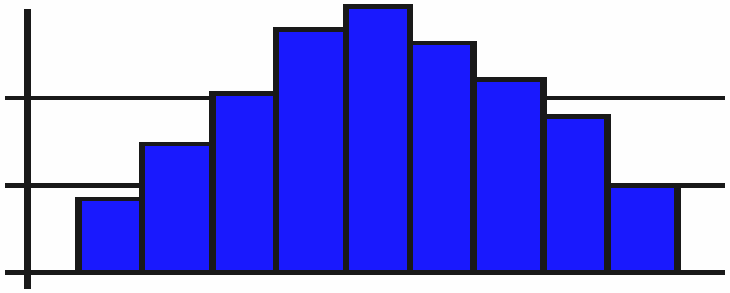
However with measurements on either side of zero and a mean being close to zero the relative dispersion could be greater than 1.

At the same time we must remember that the two distributions in quite a few cases can have the same variability as mentioned in the above example. Irrespective of the mean both are having the same standard deviation.

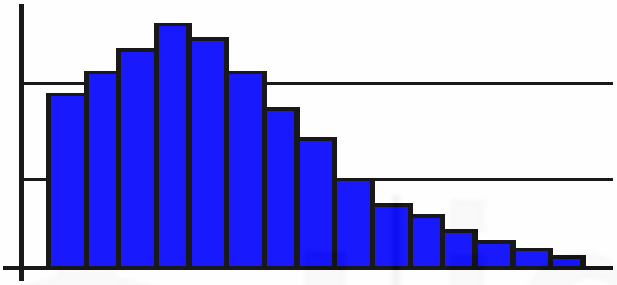
Sometimes the distributions may be skewed and not normal with mean, mode and median at different points in the continuum. These distributions are called skewed distributions.

The figures given below present the different distributions. The first one is the standard normal distribution, the second one is skewed distribution showing the distribution falling basically on the left side and tapering off towards the right.

- The standard normal distribution, with its zero skewness and zero kurtosis.



- A skewed distribution, with one tail longer than the other.



Above figures demonstrate that it is possible to have two distributions that have equal variability but unequal means or different shapes. To describe a distribution adequately, therefore, we usually must provide a measure of central tendency and a measure of variability.

Measures of variability are important in statistical inference. With the help of measures of dispersion we can know about fluctuation in random sampling. How much fluctuation will occur in random sampling? This question in fundamental to every problem in statistical inference, it is a question about variability.

Self Assessment Questions

1) Differentiate between absolute and relative dispersion.

……..

……..

……..

……..

……..

2) What is a skewed distribution?

……..

……..

……..

……..

……..

3) Why are measures of variability inportant in statistics

……..

……..

……..

……..

……..

3.3 SIGNIFICANCE OF MEASURES OF DISPERSION

The measures of variability is needed for the following purposes

- Measures of variability are used *to test to what extent an average represents the Characteristics of a data.* If the variation is small then it indicates high uniformity of values in the distribution and the average represents the characteristics of the data. On the other hand if variation is large then it indicates lower degree of uniformity and unreliable average.
- Measures of variation *helps to identify the nature and cause of variation* such information is useful to control the variation.
- Measures of variation *help in the comparison of the spread in two or more sets of data* with respect to their uniformity or consistancy.
- Measures of variation facilitate the use of other statistical techniques such as correlation, and regression analysis, and so on.

Self Assessment Questions

1) What are the purposes of measures of variability?

……..

……..

……..

……..

……..

2) Fill in the blanks

   a) For open-ended distribution ……. is used.

   b) The …… is the only measure of dispersion capable of algebraic treatment.

   c) …… is related to median in its property.

Answers: 2) (a) Quartile deviation (b) Standard Deviation, (c) Quartile Deviation

3.4 TYPES OF MEASURES OF VARIABILITY / DISPERSION

There are mainly 5 types of measures of variability and these are given below:

i) Method of limits

(1) The range (2) Inter-quatrile range (3) Percentile range

ii) Method of Averages

(1) Quartile deviation (2) Mean deviation

(3) Standard Deviation and (4) Other measures.

There are four measure of variability or dispersion

1) Range
2) Quartile Deviation
3) Average Deviation
4) Standard Deviation

We will introduce each and discuss their properties in detail.

3.4.1 The Range

This is the simplest measure of the variability. It can be defined as the difference between the highest and lowest score in the distribution.

3.4.1.1 Properties of Range

It is easier to compute than the other measures of variability and its meaning is direct. The range is ideal for preliminary work or in other circumstances where precision is not an important requirement (Minimum et. al., 2001). It is quite useful in case where the purpose is only to find out the extent of extreme variation, such as temperature rainfall etc.

3.4 1.2 Limitations of Range

1) The calculation of range is based only on two extreme values in the data set and does not consider any other values of the data set. Some times the extreme values of the two different data sets may be almost the same, but the two data set may be different in dispersion.

2) Its value is sensitive to change in sampling. The range varies more with sampling fluctuation, that is different sample of the same size from the same population may have different range.

3) Its value is influenced by sampling size. In many types of distribution including normal distribution the range is dependents on sample size. When sample size is large the value of the range is also large.

3.4 2 The Quartile Deviation

Since a large number of values in the data lie in the middle of the frequencies distribution and range depends on the extreme of a distribution, we need another measure of variability. The Quartile deviation symbolised by the letter Q is a measure depends on the relatively stable central portion of a distribution. Quartile is defined as " one half the distance between the first and third Quartile point." (Minium, 2001).

According to Garret (1966) the Quartile deviation or Q is half the scale distance between 75^{th} and 25^{th} percent is a frequency distribution.

According to Guilford (1963) the Semi inter Quartile range Q is the one half the range of the middle 50 percent of the cases.

On the basis of above definition it can be said that Quarterly deviation is half the distance between Q_1 and Q_3 .

3.4.2.1 Properties of the Quartile Deviation

The Quartile deviation is closely related to the median because median is responsive to the number of scores laying below it rather than to their exact positions and Q_1 and Q_3 are defined in a same manner. The medians and Quartile deviation have common properties. Both the median and the Quartile deviation is not effected by extreme values.

In the non-symmetrical distributions the two quartiles Q_1 and Q_3 are at equal distance from the median- Q_1 =Q_3- Median. Thus Median Quartile ± Deviation covers exactly 50 percent of observed values in the data. If the distribution is open ended than Quartile deviation is the only measures of variability that is reasonable to compute.

3.4.2.2 Limitation of Quartlie Deviation

1) The valued of Quartile deviation is based on the middle 50 percent values, it is not based on all the observations.

2) The value of Quartile deviation is affected by sampling fluctuation.

3) The value of Quartile deviation is not affected by the distribution of the individual's values within the intervals of middle 50 percent observed values.

3.4.3 The Average Deviation

Deviation Scores

Before discussing the average deviation first we should know about the meaning of deviation. Deviation score express the location of the scores by indicating how many score points it lies above or below the mean of the distribution. Deviation score may be defined as (X-Mean) i.e. when we subtract the means from each of the raw scores the resulting deviation scores states the position of the scores, relative to the mean.

The two measures of variation, Range and Quartile deviation which we discussed earlier do not show how values of the data are scattered about a centrals values. To measure the variation, as a degree to which values within a data deviate from their mean we use average deviation.

According to Garrett (1981). "The average deviation is the mean of the deviation of all of the separate scores is a series taken from their mean".

According to Guilfdord (1963), "The average deviation is the arithmetic mean of all the deviation when we disregard the algebraic signs"

3.4.3.1 Properties of the Average Deviation

The calculation of average deviation is easy therefore it is a popular measure. When we calculate average deviation equal weight are given to each observed values and thus it indicates how far each observation lie from mean.

3.4. 3.2 Limitation of the Average Deviation

The main limitation of average deviation is that while calculating average deviation we ignore the plus minus sign and consider all values as plus. Because of this mathematical properties it is not use in inferential statistics.

3.4.4 The Standard Deviation

The term standard deviation was first used in writing by Karl Pearson in 1894. A useful property of Standard Deviation is that unlike variance it is expressed in the same unit as the data.

This is most widely used method of variability. The standard deviation indicates the average of distance of all the scores around the mean.

According to Guilford (1963) "Standard deviation is the square root of the arithmetic means of squared deviation of measurements from their means". Standard deviation is also known as root-mean, square deviation. The standard deviation of population is denoted by **σ** (Greek letter sigma) and that for a sample is **S.**

3.4.4.1 Properties of the Standard Deviation

Standard deviation shows how much variation there is, from the mean. If standard deviation is low it means that the data is close to the mean, where as, the high Standard deviation shows that the data are spread out over a large range of values.

Standard deviation may serve as a measure of uncertainty. If you want to test the theory or in other word, want to decide whether measurements agree with a theoretical prediction the Standard deviation provide the information. If the difference between mean and Standard deviation is very large then the theory being tested probably needs to be revised. The mean with smaller standard deviation is more reliable than with large Standard deviation. The smaller Standard deviation shows the homogeneity of the data.

The value of standard deviation is based on every observation in a set of data. It is the only measure of dispersion capable of algebraic treatment therefore Standard deviation is used in further statistical analysis.

3.4.4.2 Limitation of the Standard Deviation

While calculating standard deviation more weight is given to extreme values and the less to those, near to means. When we calculate Standard deviation we take deviation from mean (X-M) and squared these obtained deviation therefore, large deviation, when squared are proportionally more than small deviation. For example the deviation 2 and 10 are in the ration of 1:5 but their square 4 and l00 are in the ration 1:25.

3.4.5 Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1913. The concept of variance is of great importance in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variations in their original series. Variance is defined as follows:

Variance is a measure of the dispersion of a set of data points around their mean value. It is a mathematical expectation of the average squared deviations from the mean.

Investopedia defines variance as the measure of the variability (volatility) from an average. Variance is expressed as $V = sd^2$.

The variance and the closely-related standard deviation are measures indicate how the scores are spread out in a distribution is. In other words, they are measures of variability.

The variance is computed as the average squared deviation of each number from its mean. For example, for the numbers 1, 2, and 3, the mean is 2 and the variance is:

$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = 0.667$$

.

The formula for the variance in a population is

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

where μ is the mean and N is the number of scores.

When the variance is computed in a sample the statistic used is

$$S^2 = \frac{\sum (X-M)^2}{N}$$

S = standard deviation

M = Mean of the sample

Calculating the variance is an important part of many statistical applications and analysis.

3.4.5.1 Merits and Demerits of Variance

Merits of Variance

1) It is rigidly defined and based on all observations.
2) It is amenable to further algebraic treatment.
3) It is not affected by sampling fluctuations.
4) It is less erratic.

Demerits

1) It is difficult to understand and calculate.
2) It gives greater weight to extreme values.

3.4.5.2 Co-efficient of Variation

To compare the variations (dispersion) of two different series, relative measures of standard deviation must be calculated. This is known as co-efficient of variation or the co-efficient of s. d. Its formula is

V = A statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows:

V = (Standard deviation / expected return or the mean of the group scores)

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

Thus it is defined as the ratio of standard deviation to its mean.

It is given as a percentage and is used to compare the consistency or variability of two more series. The higher the V, higher the variability , and lower the V, higher is the consistency of the data.

3.5 LET US SUM UP

Mean and other measured of central tendency is not sufficient to describe the data. To describe distribution adequately we must provide a measure of variability.

The measures of variability are summary figures that expresses quantitatively, the extent to which, scores in a distribution scatter around or cluster together.

There are four measures of variability i.e rage,quartile deviation, average deviation and standard deviation.

Range is easy to calculate and useful for preliminary work. But this is based on extreme items only, and not considers intermediate scores therefore it is not useful as descriptive measures.

Quartile deviation is related to the median in its properties. It take into consideration the number of scores lying above or below the outer Quartile point but not to their magnitude. This is useful with open ended distribution.

The average deviation is take into account the exact positions of each score in the distribution. The means deviation gives a more precise measure of the spread of scores but is mathematically inadequate. The average deviation is les effected by sampling fluctuation.

The standard deviation is the most stable measure of variability. Standard deviation shows how much the score depart from the mean. Because is express in original scores unit it is most widely used measure of variability in descriptive statistics.

3.6 UNIT END QUESTIONS

1) Explain the term dispersion, discus the importance of studying dispersion.
2) What are the different measures of dispersion ?
3) What are the different types of dispersions we have? explain
4) Discuss the properties and limitation of range.
5) Discuss the properties and limitation of quartile deviation.
6) Why is the standard deviation most widely used measured of dispersion?
7) What is meant by Coefficient of Variation? Explain with examples

3.7 GLOSSARY

Dispersion : The Spread or variability is a set of data.

Deviation : The Difference between raw score and mean.

Range : Difference between the largest and smallest value in a data.

Quartile Deviation : A measure of dispersion that can be obtained by dividing the difference between Q_3 and Q_1 by two.

Average Deviation : A measure of dispersion that gives the average difference (ignoring plus and minus sign) between each item and the mean.

Standard deviation : The square root of the variance in a series.

3.8 SUGGESTED READINGS

Minium, E.W., King, B.M.& Bear .G (2001). *Statistical Reasoning in Psychology and Education*, (third edition), Singapore, John Wiley &Sons, Inc,

Garrett, H.E. (1981), *Statistics in Psychology and Education*, (Tenth edition), Bombay, Vakils Feffer and Simons Ltd.

UNIT 4 RANGE, MD, SD AND QD

Structure

4.0 INTRODUCTION

In the last unit we have learned that average like mean, median and mode condense the series into a single figure. These measures of central tendency tell us something about the general level of magnitude of the distribution but they fail to show anything further about the distribution. It is not fully representative of a population unless we know the manner in which the individual items scatter around it. A further description of the series is necessary if we are to gauge how representative the average is. To cite an example, in a country the average income may be very high. Yet there may be great disparity in its distribution among people. As a result, a majority of the people may be living below the poverty line. When we want to make comparison between two groups, it is seen that at times the value of the means is the same in both the groups but there is a large difference between individual subjects in the groups. This difference amongst the subjects within the same group is termed as variation, that is within the groups the subjects vary a great deal even though they have the same means. Therefore to make accurate and meaningful comparisons between groups, we should use variability along with central tendency. In the last unit we have learned about the concept of variability, different measures of variability their properties and limitations. In this section we will learn how we compute the range, quartile deviation, average deviation and standard deviation. We will also discuss when to use various measures of variability.

4.1 OBJECTIVES

After reading this unit, your will be able to:

- Compute the Range;
- Compute Quartile deviation, (for ungrouped and grouped data);
- Compute Average deviation (for ungrouped and grouped data);
- Compute Standard deviation (for ungrouped and grouped data); and
- Select the most appropriate measure of dispersion for a given set of data.

4.2 COMPUTING DIFFERENT MEASURES OF VARIABILITY

There are four measures of computing variability or dispersion within the set of scores:

Range (R)

Quartile Deviation (Q)

Average Deviation (AD)

Standard Deviation (SD)

Each of the above measures of variability gives us the degree of variability or dispersion by the use of a single number and tells us how the individual scores are spread throughout the distribution. In the following paragraphs we will discuss the methods of computation of the above measures of dispersion .

4.2.1 Range

Range is the difference between the highest and the lowest score inn a group of subjects whose scores are given. For example, if there are 10 students and they have obtained marks in history, as given here:

45,42,46,50,55,54,59,60,62,64, In this group the lowest score is 42 and thed highest score is 64. The range therefore is 64 – 42 = 22.

The formula for Range is

R=H-L (Highest scores minus Lowest scores)

R=range

H= Highest scores in the distribution

L= Lowest score in the distribution

The use of the formula can be explained by the following illustration

To give another example, let us say that the following are the scores obtained by 6 students in a GK test.

Scores 25,17,14,18,20,13.

Now arrange these scores in ascending order : 13,14,17,18,20,25

By applying the formula

Highest score =25

Lowest score= 13

Range= 25-
13=12

Steps for Calculating range

- Arrange the sores in ascending order
- Find out highest and lowest score of the series.
- Find out the difference between highest and lowest scores.

4.2.2 Quartile Deviation (QD)

The inter-quartile range is a measure of dispersion and is equal to the difference between the third and first quartiles. Half of the inter-quartile range is called semi inter-quartile range or **Quartile deviation**. Symbolically it is defined as;

$Q.D = (Q_3 - Q_1)/ 2$

Where Q_1 and Q_3 are the first and third quartiles of the data. What are qualrtiles? Quartiles are an additional way of disaggregating data. Each **quartile** represents one-fourth of an entire population or the group The quartile deviation has an attractive feature that the range "median + Q.D" contains approximately 50 % of the data. The quartile deviation is also an absolute measure of dispersion. Its relative measure is called coefficient of quartile deviation or semi inter-quartile range. It is defined by the relation;

Coefficient of quartile deviation= $(Q_3 - Q_1)/(Q_3 + Q_1)$

The quartile deviation or Q is one half the scale distance between the 75^{th} and 25^{th} percentile in a frequency distribution . The 25^{th} percentile or Q1 is the *first quartile* on the score scale , the point below which lie 25% of the scores.

The 75^{th} percentile or Q3 is the *third quartile* on the score scale the point below which lie 75%of the scores . To find Q we must first compute the Q3 and Q1.

There are grouped data and ungrouped data as in all cases and thus to compute quartile deviation we have to find out first if it is a grouped data ore ungrouped data. We will first see how the Q is calculated from ungrouped data.

4.2.2.1 Calculation of Quartile Deviation for Ungrouped Data

The inter-quartile range is a measure of dispersion and is equal to the difference between the third and first quartiles. Half of the inter-quartile range is called semi inter-quartile range or **Quartile deviation**. Symbolically it is defined as;

$Q.D = (Q_3 - Q_1)/ 2$

Where Q_1 and Q_3 are the first and third quartiles of the data. The quartile deviation has an attractive feature that the range “median + Q.D” contains approximately 50 % of the data. The coefficient of quartile deviation is also an absolute measure of dispersion. Its relative measure is called of quartile deviation or semi inter-quartile range. It is defined by the relation;

Coefficient of quartile deviation= $(Q_3 - Q_1)/(Q_3 + Q_1)$

Example:

Obtained Scores = 24,25,23,26,29,30,27,35,34,36,28

First arrange the data in ascending order

23,24,25,26,27,28,29,30,34,35,36

Q1 = (N+1)/4th position

N=11 Q1=11+1/4th position = 3rd position= 25

Q3 =3(N + 1) /4th position

N=11 Q3=3(11+1)/4=9th position =34

$Q= Q_3-Q_1 / 2$

Q3=34

Q1=25

Q=34-25/2 =4.5

4.2.2.2 Calculation of Quartile Deviation for Grouped Data

From the grouped data Quartile Deviation can be computed by the formula

Q= Q3-Q1 **/ 2**

$Q1= l+i \{(N/4-cumf_l)\}/fq$

Q3= *l* + i (3N/4-cum fi)/fq

Where *l* = the exact lower limit of the interval in which the quartile falls.

i = the length of the interval

cum fi= cumulative f up to interval which contains the quartile

fq = the f on the interval containing the quartile.

The use of the above formula can be illustrated by the following example

Class intervals	Frequencies	Cumulative frequencies
195-199	1	50
190-194	2	49
185-189	4	47
180-184	5	43
175-179	8	38
170-174	10	30
165-169	6	20
160-164	4	14
155-159	4	10 (1+3+2+4)
150-154	2	6 (1+3+2)
145-149	3	4 (1+3)
140-144	1	1

First quartile deviation Q1 is calculated using the formula given below:

$Q1=l+i \{(N/4-cumf_l)\}/fq$

l= 159.5 (50/4 = 12.5 the item from down below) (falls in 160-164) the *l*= 159.5. (Lower ,limit of that class interval)

fi= 10 cumulated scores up to interval containing Q_1

fq= 4 the f on the interval on which Q_1 falls.

i = 5 (Class Interval)

Substituting in formula we have that.

Q1=159.5+ 5{(12.5-10)}/4 = 162.62

Now to calculate the third quartile that is Q3.

Q3=l+i{(3N/4-cumfl)}/fq

3/4N= 37.5 (37.5^{th} item is 175-179)

l=174.5 is the exact lower limit of interval which contains Q3

Cum fi= 30, sum of scores up to interval which contains Q_3

i=5

fq=8

Q_3=174.5+ 5(37.5-30)/8 = 179.19

Finally, substituting in formula, we have the Quartile Deviation Q as given below in the formula

Q = Q_3-Q_2 / 2

Q= { (179.19)- (162.62)} / 2 = 8.28

Thus the Quartile deviation for the above data is = 8.28

Steps involved in Calculation of Quartile Deviation

1) To locate the Q_1 we take N/4 and to locate Q_3 we takes 3 x N/4 of our scores

   In the above example N/4 = 12.5 and 3N/4 is 3x 50/4 =37.5

   The 12.5^{th} item falls on 160-164 (when you add the frequencies form 140-144 upto 160-164, you will find that the 12.5^{th} item falls on this class interval.)

   The 37.5^{th} item falls on 175-179 (Continue adding the frequencies from 140-144 upto reaching 37.5 and you will find the 37.5^{th} item falls on the calss interval 175-179. 0

2) Now find out the exact lower limit of the class interval on which Q_1 and Q_3 fall. The exact lowe limit of 160-164 is 159.5 and the exact lower limit of 175-179 is 174.5.

3) Now compute the cumulative frequency (that is, as you keep adding the frequencies from one class interval to an other it is cumulative)

4) For Q1 it is 10 and for Q3 it is 30.

5) Now find out the fq which is the frequency of the interval upon which the Q1 and Q3 fall . For it is 4 and for Q 3 it is 8.

6) Now apply the formula and calculate the Q1 and Q3.

   Q1 = *l* + i(N/4- cumulative frequency) / fq

   Substituting the numbers in the above formula we get the following:

   Q1 = 159.5 + 5 (50/4 – 10) / 4

   = 159.5 + 5 x 2.5 / 4

   = 162.62

Q3 = *l* + i(3N/4- cumulative frequency) / fq

= 174.5 + 5(37.5 – 30) / 8

= 174.5 + 4.7 = 179.2

7) Subtract the obtained Q3 from Q1 and divide by 2

179.2 - 162.2. / 2

17 / 2 = 8.5.

4.2.3 The Average Deviation (A.D)

4.2.3.1 Computation of Average Deviation from Ungrouped Data

In the case of ungrouped data the average deviation is calculated by the following formula

$$AD=\sum |x| /N$$

Above formula can be illustrated by an example given below:

Table: Scores of 5 students

Scores	Deviation from the Mean of 10.
6	-4
8	-2
10	0
12	2
14	4
Total = 50 MEAN = 50/5=10	Total = 12 (Ignore signs)

$M= \sum X/N$

M= 50/5 = 10 Now let us see from this average of 10 how much each score deviates

It is seen from the table, the deviations (x) = 0, -2, -4, +2, +4

Now we add up these deviations without bothering about the + and – signs.

$AD= \sum |x| /N$

$\sum |x| = 12$, N= 5

AD=12/5=2.4.

Steps

- Find out the mean by adding scores and divide it by the number of observations
- Find out the deviation of each score from this mean x= (X-M) .
- Add these deviations disregarding plus and minus sign ($\sum |x|$)
- Then divide it by the total number of subjects that is N

4.2.3.2 Calculation of Average Deviation from Grouped Data

The Average Deviation for grouped data can be computed by the following formula

$$AD = \sum |fx| / N$$

AD= Average deviation

$\sum |fx|$ = Add all the fx without considering the + and – sign

N = Number of observations

The above formula and calculation of AD can be illustrated by the example given below.

Class Interval	Frequency	Mid Point(X)	fX	X (M-X) M=91.67	Fx (freq. × difference)
110-114	3	112	336 (112x3)	20.33 (112-91.67= 20.33)	60.99 (20.33 × 3)
105-109	4	107	428 (107×4)	15.33	61.32
100-104	6	102	612 (102×6)	10.33	61.98
95-99	8	97	776	5.33	42.64
90-94	15	92	1380	.33	4.95
85-89	10	87	870	-4.67	-46.67
80-84	7	82	574	-9.67	-67.69
75-79	4	77	308	-14.67	-58.68
70-74	3	72	216	-19.67	-59.01
	60		5500		

M= 5,500/60=91.67

$AD=\sum |fx| / N$

AD=463.96/60=7.73

Thus the Mean deviation is = 7.73

4.2.4 The Standard Deviation

The standard Deviation is the most stable measure of variability. Therefore it is most commonly used in research studies.

4.2.4.1 Calculation of Standard Deviation for Ungrouped Data

Standard Deviation for ungrouped data can be computed by the following formula.

Formula

$$SD = \sqrt{\sum x^2} / N$$

The above formula can be explained by the following example.

Example

Mean of the given scores = $\sum X/N = 480/8 = 60$ From this 60, how do the scores in the table deviate. We find that the deviations are -8, -10,….-3, =10 etc. (the deviations are given in the table below.

Table: Scores obtained by 8 subjects

Scores	Deviation from the mean (x)	(x^2) deviation square
52	-8	64
50	-10	100
56	-4	16
68	8	64
65	5	25
62	2	4
57	-3	9
70	10	100
Total		382

To calculate standard deviation using the formula , we get the following

SD= $\sqrt{(\sum x^2)}$ /N

Here

$\sum x^2$=382 N = 8

SD= $\sqrt{}$ 382/8 = $\sqrt{}$ 47.7 = 6.91

Thus the standard deviation for this data is 6.91.

Steps

For ungrouped data

1) Add all the scores ($\sum x$) and divide this sum by the number of scores (N) and find out mean.
2) Find out the difference between scores and means (X-x) and find out deviation (x).
3) Square all the deviation to get x^2.
4) Add all the squared deviation to get$\sum x^2$.
5) Divide $\sum x^2$ by N.
6) Find out square root of the obtained values.

4.2.4.2 Computations of SD from Grouped Data by Long Method

Standard deviation of grouped data can be computed by the formula

$$SD = \sqrt{\sum fx^2}/N$$

The use of the formula can be illustrated with the help of following example.

Table: Class interval of scores

Class interval	Frequency	X	Mean	Deviation (x)	Deviation squared (x^2)	fx^2
122-129	1	128	115	13	169	169
124-126	2	125	115	10	100	200
121-124	3	122	115	7	49	147
118-120	1	119	115	4	16	16
115-117	6	116	115	1	1	6
112-114	4	113	115	2	4	16
109-111	3	110	115	5	25	75
106-108	2	107	115	8	64	128
103-106	1	104	115	11	121	121
100-102	1	101	115	14	196	196
	24					$\sum fx^2=1074$

SD= $\sqrt{}$ $(\sum fx^2)$ /N

Here

$\sum fx^2=1074$ N = 24

SD = $\sqrt{}$ 1074/24 = 6.99

Steps

For grouped data Long method

1) Add all the scores ($\sum x$) and divide this sum by the number of scores (N) and find out mean.
2) Find out the difference between scores and means (X-x) and find out deviation (x).
3) Square all the deviation to get x^2.
4) Multiply each of the squared deviation by the frequency which is opposite ie. (x)x (fx) . This multiplication gives the fx^2.
5) Add the fx_2 ($\sum x^2$) and divide it by number of observation (N).
6) Find out square root of the obtained value

4.2.4.3 Calculation of SD from Grouped Data by Short Method

Standard deviation from grouped data can also be computed by the following formula.

SD= i $\sum fx^2$ /N - $[\sum(fx) / N]^2$

The use of this formula is illustrated in the following example.

Example

Table: Class interval of scores

Class interval	Frequency	X	X'=X-A/i	fx'	fx'²
122-129	1	128	4	4	16
124-126	2	125	3	6	18
121-124	3	122	2	6	12
118-120	1	119	1	1	1
115-117	6	116	0	0	0
112-114	4	113	-1	-4	4
109-111	3	110	-2	-6	12
106-108	2	107	-3	-6	18
103-106	1	104	-4	-4	16
100-102	1	101	-5	-5	25
	24				$\sum fx^2=1074$

The use of this formula is will restated is the following example.

SD= i $\sum fx^2$ /N - $[\sum(fx) / N]^2$

Here

i = 3

$\sum fx^2$ =122

N=24

$\sum fx$= -8 $\{\sum fx / N\}^2$

(-8 /24) ² = 64 / 576

SD= 3x {122/24-(8x8/24x24) }

Sd = 3 x 2.2. = 6.69

Steps

1) Find out midpoint of each class interval.
2) Assume one value as mean.
3) Find out the difference between mid point and assumed mean and divide it by class intervals to get x'.
4) Multiply each x' by respective frequency and get fx'.
5) Multiply fx' by frequency opposite to it to get fx'^2.
6) Add all the fx'^2 ($(\sum x^2)$ divide it by number of observation.
7) Divide $\sum x'$ by number of observation and get whole square of it.

8) Find the difference between values obtained by step 6 and 7 and find out the square root.

9) Multiply intervals by obtained square root value.

4.3 WHEN TO USE DIFFERENT MEASURES OF DISPERSION

The following rules are useful to guide us:

4.3.1 Use the Range

The use of range is recommended

When

1) We want to know about highest and lowest scores.
2) When quick and easy computation of measures of variability is required .
3) When the data too scant or too scattered so computation of other measured of variability is not useful.

4.3.2 Use the Quartile Deviation

1) The distribution contain few and very extreme scores.
2) When the median is the measure of centrals tendency.
3) When our primary interest is to determine the concentration around the median.

4.3.3 Use the Average Deviation

1) When it is desired to weight all deviation from the mean according to their size.
2) When the standard deviation in unduly influenced by the presence of extreme scores.
3) Distribution of the score is not near to normal.

4.3.4 Use the Standard Deviation

When coefficient of correlation, significance of difference between means and other statistics are subsequently to be calculated.

1) Measure of central tendency is available in the form of mean.

4.4 KEY FORMULAS

1) **Range** = H- L
2) **Quartile deviation Q** = Q_3-Q_1/2

   Q_1=1+i(N/4-Cumfi)/fq

   Q_3=1+i(3N/4-Cumfi)/fq

   Average deviation from ungrouped data $\sum |x|/N$

   Average deviation from grouped data=$\sum |fx|/N$

   Standard deviation from ungrouped data = $\sqrt{\sum x^2/N}$

Standard deviation from grouped data by long method

S.D.= $\sqrt{\sum fx^2/N}$

Standard deviation from grouped data by short method

SD = $i\sqrt{\sum fx^2/N - \sum(fx)^2/N}$

4.5 LET US SUM UP

In this unit we dealt with the concept of disopersion and related materials. We gave an introduction to computing range and how the range varies from lowest to the highest value. We then dealt with quartile deviation and also worked out the example with the help of formulate. We then took up the definition of Mean or Average deviation and dealt with it in detail as to when and where it should be used. This was followed by how the average deviation should be calculated. After this we took up standard deviation and understood the way it should be calculated and the different formula that we could use for the same. We then listed out use of range, average deviation, quartile deviation and the standard deviation.

4.6 UNIT END QUESTIONS

1) Compute the range, average deviation and standard deviation from the following ungrouped data:
   a) 30,35,36,39,42,46,38,34,35
   b) 52,50,56,68,65,62,57,70
2) Calculate Q from the following scores:
   6,3,9,9,5,7,9,6,8,4,8,5,7,9,3,2,9,5,7
3) Calculate Average deviation of the following scores:

a)

Class Interval	Frequency
40-44	3
35-39	4
30-34	6
25-29	12
20-24	7
15-19	5
10-14	1
	N = 38

b)

Class Interval	Frequency
50-59	6
40-49	3
30-39	5
20-29	8
10-19	4
0-9	2
	N = 28

4) Calculate the Quartile deviation and Standard deviation for the following frequencies distribution:

Scores	Frequency
70- 71	2
68-69	2
66-67	3
64-65	4
62-63	6
60-61	7
58-59	5
56-57	1
54-55	2
52-53	3
50-51	1
N	36

Scores	Frequency
90-94	0
85-89	0
80-84	1
75-79	5
70-74	6
65-69	11
60-64	9
55-59	7
50-54	5
45-49	0
40-44	2
N	56

5) Which measure of variability would you prefer in the following situation

   i) When one wants to quickly have some idea of the variability of a set of data
   ii) When one have open ended distribution
   iii) When there are extreme values in a distribution
   iv) When stable measure of variability is required
   v) The co efficient of correlation is subsequently to be computed

6) Do you think standard deviation is the best measure of dispersion amongst all the measures of variability ? Comment .

Answers

1) i) R=16,AD=3.9,SD=4.68
   ii) R=10,AD=6.25,SD=6.91
2) Q=2
3) a) 5.78
   b) 13.21
4) a) SD=13.55,Q=9.79
   b) SD=11.33,Q=8.12
5) i) RANGE, ii) Q, iii) Q, iv) SD, v) SD

4.7 SUGGESTED READINGS

Aron A., Aron, E.N., & Coups, E..J. (2006), *Statistical for Psychology* (4th ed), New Delhi. Pearson Prentice Hall.

Asthana, H.S.& Bhushan, B (2007). *Statistics for Social Sciences*, New Delhi, Prentice Hall of India Private Limited..

Hopkins, K.D.,& Glass, G.V.(1978). *Basic Statistics for the Behavioral Science.* Englewood Cliffs, N.J.Prentice Hall.

Levin, and Fox, J.A. (2006), *Elementary Statists in Social Research*, New Delhi, Pearson education

Web link http://argull.epsb.Ca/freed/mathrs/str/and 4/central tendency.htm

UNIT 1 INTRODUCTION TO PARAMETRIC CORRELATION

Structure

1.0 INTRODUCTION

This Unit presents an idea about correlation, that is, how two variables vary with each other. For e.g. if height increases weight increases and vice versa. In this unit we will be discussing how two or more variables are related to each other. We will also see how the variables are spread in a population and learn the method of working out a scatter diagram. This unit also discusses the linear and non-liner relationship among variables and the direction of correlation in terms of positive or negative or no correlation. While correlation presents the relationship, whether this relationship is strong and if so the degree of relationship and how the same is measured are all presented in this unit.

1.1 OBJECTIVES

After reading and doing exercises in this unit you will be able to:

- Describe and explain concept of correlation;
- Plot the scatter diagram;
- Explain the difference between linear and nonlinear relationships;

- Explain the concept of direction of correlation and differentiate between positive, negative and no relationship;
- Compute and explain the strength of relationship;
- Differentiate between various measures of correlations;
- Evaluate conceptual issues in correlation and causality; and
- Create problems suitable for correlation analysis.

1.2 INTRODUCTION TO CORRELATION

On some lazy Sunday afternoon, after a long summer break, you are awaiting for your results which may be out in couple of days. As everyone does, you would also tend to speculate about your grade or marks. And you would wonder about the factors that would determine your academic performance, that is, your marks. Often it is thought that *amount of study* is associated with the *marks* obtained in the examination. Generally, we will believe that if we study more, then we will get more marks. And if we study little, then we will get lesser marks in examination. Though, it is generally true, you will also note that that some of your friends have scored well in spite of studying less. At the same time you would also realise that there are few unlucky students who, in spite of studying more, did not obtain the better marks. Still, generally, most of the students would obtain marks as per their studies. Hence, you would believe that there is a positive relationship between amount of study and marks obtained.

Actually, *Mukta* has really got interested in this question. She quickly called up some of her friends and found out some interesting information. She enquired about two things: first, number of hours they have spend in studying per day for the examination and second, the marks obtained in that examination. She collected this information from five of her friend which is given below .. Table 1.1 Data showing number of hours spent in studies and marks obtained for five individuals.

Table 1.1: Number of hours studied and marks obtained

Friends name	Hours spent in studies/per day	Marks obtained
Sujata	2	55
Jasbir	3	60
Sidharth	4	65
Naseem	5	70
Yohan	6	75

What do you observe? You will quickly realise that table 1.1 indicates that as the number of hours spent in studies increases, the marks obtained in the examination also increase. So we can say that there is a positive association between number of hours spent in studies and marks obtained. This is a kind of data that is useful in understanding the correlation. Since this is a first

example, I have kept the data very simple. In the further examples, we will have a look at comparatively more realistic (and hence more complex) data.

In this example, we have observed that 'more you study better you score'. Can you think of such examples on your own? Hmmm....you are taking time...ok...I shall provide you some such examples and then you can try. Have a look at the following statements. They are examples of positive and negative relationships.

- As the experience of employees increases, their salaries increase.
- As kids grow up, their memory improves.
- If people have more interest in a particular subject, then they are likely to score well in that subject.
- Those who are better at literature are not good at mathematics.
- The more you practice a skill, lesser the mistakes you make.
- As the depression increases, the optimism decreases.

Activity for students

Now, it's your turn. Generate five such examples. Let's start....!

Are you done with the five examples? I know you must have found this task interesting. Now we shall learn to plot *scatterplot*.

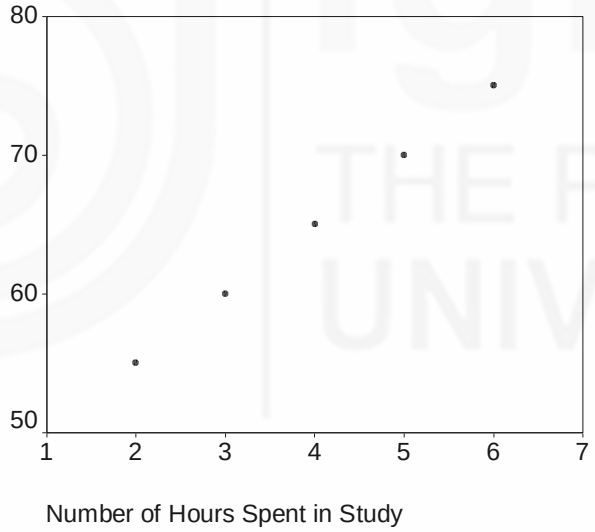


Fig. 1.1: Scatter diagram showing the relationship between number of hours spent in studies and marks obtained.

1.3 SCATTER DIAGRAM

Scatter diagram (also called as *scatterplot*, *scattergram,* or *scatter*) is one way to study the relationship between two variables. Scatter diagram is to plot pairs of values of on a graph. Let's learn to make a scatter diagram.

1.3.1 How to Make Scatter Diagram?

Step 1. Draw the x-axis and y-axis on the graph. The x-axis is horizontal and y-axis is vertical. Plot one of the variables on x-axis and another on y-axis. You can plot any variable on any axis for correlation analyses. If the two

variables share a cause-effect relationship, then plot the causal variable on x-axis and effect variable on y-axis. Please note that correlation does not necessarily imply causality. In our example, we plot 'number of hours spent in studies' on x-axis and 'marks obtained' on the y-axis.

Step 2. Now we decide the range of values. Usually the lowest score is zero. But, if the range of values begins from higher value, then you can start from a higher value than zero. In our example, marks are starting from 55, so we can start the y-axis from 50. The axis can continue till the highest value you have in the data. Conventionally, the scatterplot is square. So plot x and y values about the same length.

Step 3. Identify the pairs of values. A pair of value is obtained from a data. A pair of values is created by taking a one value on first variable and corresponding value on second variable. For example, Sujata's scores on first variable (number of hours spent in studies) is 2, and corresponding value on other variable (marks obtained) is 55, so the pair is 2 and 55. Similarly you have four other pairs.

Step 4. Now, locate these pairs in the graph. Find an intersection point of x and y in the graph for each pair. Mark it by a clear dot. Then take second pair and so on. For example, Jasbir's three hours are plotted with his 60 marks, and so is the case with others. The graph (Figure 1.1) below shows *Mukta's* finding more clearly.

Figure 1.1 above — Scatter diagram shows the relationship between number of hours spent in studies and marks obtained.

You will realise that the pairs are plotted in the graph; the scatter diagram clearly shows that as the number of hours spent in study increase, the marks are also increasing.

Let's take some more data and plot the graph for those data (note that 'data' is a plural word, 'datum' is singular. So do not say 'datas'. You will note that 'data is' is not opposite expression, but still commonly used. 'Data' actually 'are'. Pronounce it as '*deyta*').

Let's look at the data of experience and salaries for 10 employees. Following table 1.2 shows the data.

Table 1.2: Experience in months and salary

Employee	Experience (in months)	Salary (in Rupees)
1	2	5000
2	6	5500
3	12	6000
4	18	6500
5	24	7000
6	24	7000
7	30	7500
8	35	8000
9	40	8500
10	48	10000

Let's draw a scatterplot for this data. This data shows positive relationship. As the experience is increasing the salary of the employee is also increasing. The plot is shown below in figure 1.2.

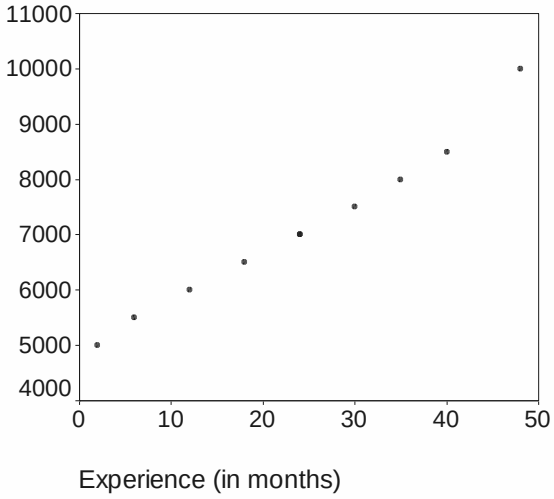


Fig. 1.2: Scatterplot shows relationship between experience and salary.

Table 1.3 shows data between extroversion and number of friends. Extroversion score and information about number of friends has been obtained for six individuals. This information is given below. We will plot a scatter for this information.

Table1.3: Extroversion and No. of friends

Extroversion scores	Number of Friends
12	**3**
10	**1**
14	**5**
18	**9**
16	**4**
20	**7**

The data in table 1.3 is plotted in figure 1.3. You will also realise in this plot, there is no one to one relationship between extroversion and number of friends. It denotes that the relationship is not perfect. Still, the general trend is that as extroversion increases the number of friends' increase. Look at the figure 1.3. Figure 1.3. Scatter showing relationship between Extroversion and Number of Friends.

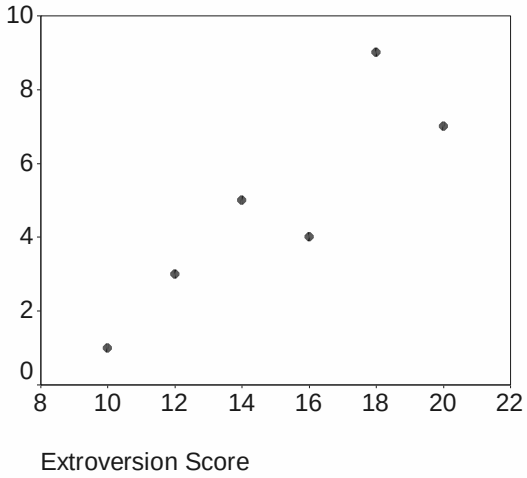


Fig. 1.3: Relationship between Extroversion and Number of Friends.

The Table 1.4 shows the data for intelligence and mistakes on reasoning task. Plot a scatter plot for this data.

Table 1.4: Intelligence and mistakes on reasoning task.

Individual	Intelligence (IQ)	Mistakes on reasoning task
1	110	6
2	113	4
3	115	5
4	106	6
5	102	8
6	98	11
7	111	3
8	120	0

Have you plotted this data? Is it looking like the plot shown in figure 1.4? What do you observe in this data? Look at the scatter below and try to understand.

Figure 1.4. Scatter diagram showing relationship between Intelligence and the number of Mistakes on reasoning Tasks.

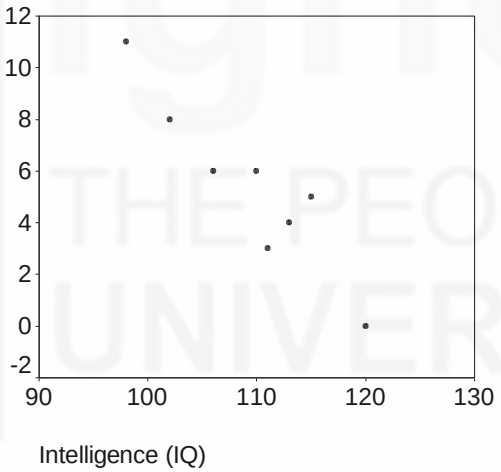


Fig. 1.4: Intelligence and no. of mistakes on reasoning task

The relationship between intelligence and mistakes is different than the one we have observed in other figures. But still, there is a relationship between them. Generally, as intelligence is increasing, the mistakes appear to be reducing.

Self Assessment Questions

Now, it's your turn to plot scatter plot for the data shown in table 1.5 and table 1.6. This exercise will give an opportunity to you to test your skills to understand and plot scatter.

Table 1.5. Data showing scores on time taken to complete 100 meters race and duration of practice for 5 swimmers.

Table 1.5: Time taken and duration of practice

Time taken (In Seconds)	Duration of Practice (In Months)
30	10
32	12
34	8
27	14
37	6

2) Table 1.6. Data showing scores on sleep deprivation (in hours) and scores on irritability measured by standardised test for seven individuals. Plot the scatterdiagram

Table 1.6: Sleep deprivation and degree of irritability

Sleep Deprivation (In hours)	Irritability Scores
12	5
16	7
19	9
27	13
30	16
25	11
22	6

Ok. Have you plotted the scatter for the examples shown in the table 1.5 and 1.6. So far, we have learned about the scatterplot. Now we shall learn about the direction of the relationship.

1.4 CORRELATION: LINEAR AND NON-LINEAR RELATIONSHIP

The relationship between two variables can take a variety of forms. They can be understood as linear and nonlinear relationships.

1.4.1 Linear Relationship

The most basic form of relationship is linear relationship. *Linear* relationship can be expressed as a relationship between two variables that can be plotted as a *straight* line on a graph. As the name suggests, the *non-linear* relationships, cannot be plotted as a *straight line*. Nonlinear relations may take various forms. For example, cubic, quadratic, polynomial, exponential, etc. We have seen variety of examples of scatterplot in the previous section. All these examples are examples of linear relationship. The linear relationship can be expressed in the following equation (eq. 1.1):

$$Y = \alpha + \beta X \quad \text{(eq. 1.1)}$$

In this Y is a variable on y-axis (often called as dependent), α (alpha) is a constant or Y intercept of straight line, β (beta) is slope of the line and X is variable on x-axis (often called as independent). We again plot scatter with the line that best fits for the data shown in table 1.3. So you can understand the linearity of the relationship. Figure 1.7 shows the scatter of the same data. In addition, it shows the line which is best fit line for the data. Figure 1.7 shows that there is a linear relationship between two variables, extroversion and number of friends. The graph also shows the straight line relationship indicating linear relation.

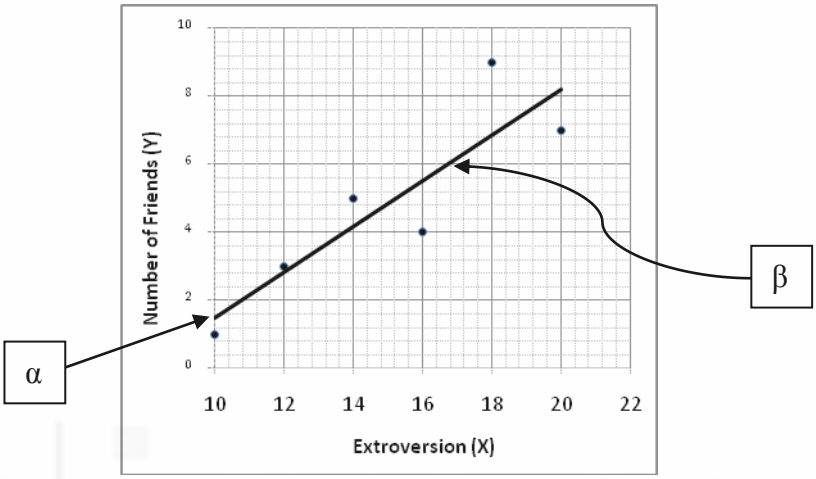


Fig. 1.7: Linearity of the relationship between extroversion and number of friends.

1.4.2 Non-Linear Relationship

There are other forms of relationships as well. They are called as curvilinear or non-linear relationships. One such example is an example of relationship between stress and performance, popularly known as Yorkes-Dodson Law. It suggests that the performance is poor when the stress is too little or too much. It improves when the stress is moderate. Figure 1.8 shows this relationship.

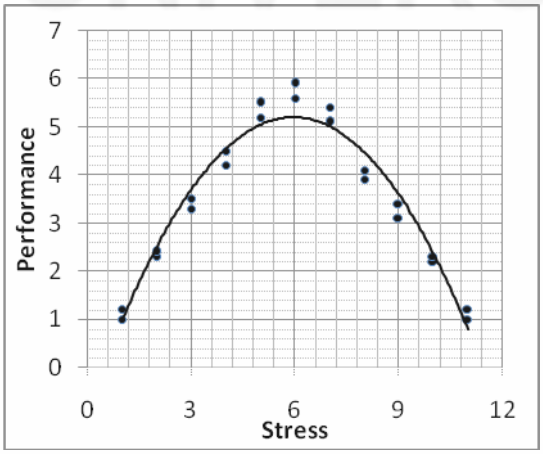


Fig. 1.8: Typical relationship between stress and performance. The performance is poor at extremes and improves with moderate stress. This is one type of curvilinear relationship.

The curvilinear relationships are of various types (cubic, quadratic, polynomial, exponential, etc.). This is not an appropriate place to discuss them. The point we need to note is that *relationships can be of various types*. This block discussed **only** *linear* relationships. Other forms of relationship are *not* discussed. The types of correlation presented in this block represent linear relationships. Pearson's product-moment correlation, Spearman's rho, etc. are linear correlations.

1.5 DIRECTION OF CORRELATION: POSITIVE AND NEGATIVE

The examples we have discussed in introduction and scatter diagram sections are called as 'correlations'. These statements are discussing relationship between 'two' variables. When the two variables are correlated, then they simply go together. They can go together in two different ways: positive and negative. So the relationship between them is either positive or negative. The scatter diagram can show us the direction of the relationship.

1.5.1 Positive Correlation

The positive correlation indicates that as the values of one variable increases the values of other variable also increase. Similarly, as the values of one variable decreases, the values of other variable also decrease. It means that both the variables move in the similar direction. For example,

As the temperature increases, sales of cold-drinks increase.

As openness to experience increases, the creativity scores increase.

The figure 1.9 shows *scatterplot* of the positive relationship.

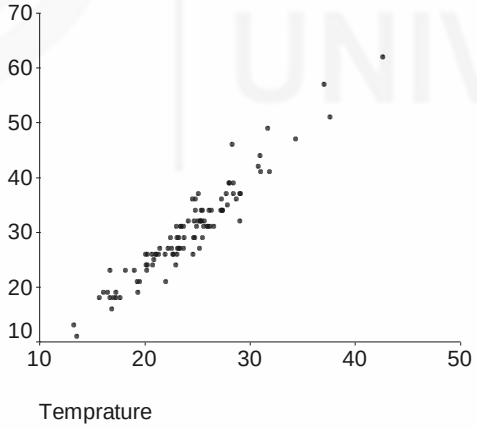


Fig. 1.9: Positive relationship. Scatter showing relationship between temperature and sales of cold-drinks for 100 data points.

You will realise that the higher scores on X axis are associated with higher score on Y axis and lower scores on X axis are generally associated with lower score on Y axis. In the 'a' example, higher scores on temperature are associated with the higher score on sales of cold-drinks. Similarly, as the temperature drops down, the sales of the cold-drinks has also dropped down.

1.5.2 Negative Correlation

The negative correlation indicates that as the values of one variable increases the values on other variable also decrease. Similarly, as the values on one variable decreases the values on other variable increase. It means that both the variables move in opposite direction. For example, As the temperature increases, the sales of woollen cloths decrease.

As social anxiety increases, assertiveness decreases.
Figure 1.10 shows *scatterplot* of the negative relationship. You will note that the higher scores on x-axis are generally paired with lower scores on y-axis and lower scores on x-axis are generally paired with higher scores on y-axis. In the first example (example a), higher scores on temperature are paired with the lower score on sales of woollen clothes. Similarly, lower temperatures are paired with higher sales of the woollen cloths. Often young students are likely to believe that negative correlation is 'bad' or 'undesirable'. Indeed, positive and negative are just directions. Neither of them is desirable or undesirable. Figure 1.10. Negative relationship. Scatter showing relationship between temperature and sales of woollen cloths.

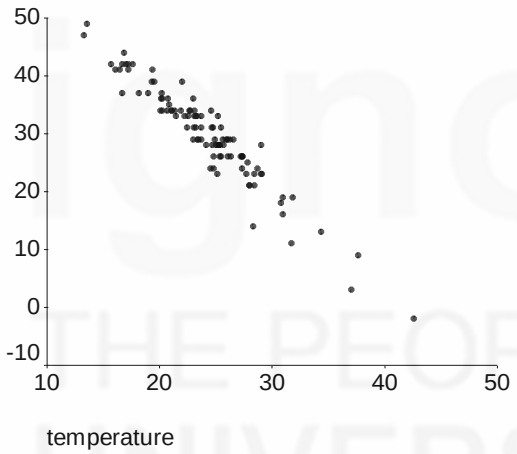


Fig. 1.10: Negative relationship between temperature and woollen sales.

1.5.3 No Relationship

Now, you have understood the positive and negative correlation. But there is a third possibility as well. That is the two variables do not share any relationship. If they do not share any relationship (that is, technically the correlation coefficient is zero), then, obviously, the direction of the correlation is neither positive nor negative. It is often called as zero correlation or no correlation. (Please note that 'zero order correlation' is a different term than 'zero correlation' which we will discuss afterwards). For example, what would be the relationship between income and height of an individual? You can easily guess that there is no relationship between them. The data of one hundred individuals is plotted in figure 1.11. It shows the scatterplot for no relationship.

Figure 1.11. Scatter between height of an individual and income in Rs. (per month) of the individual.

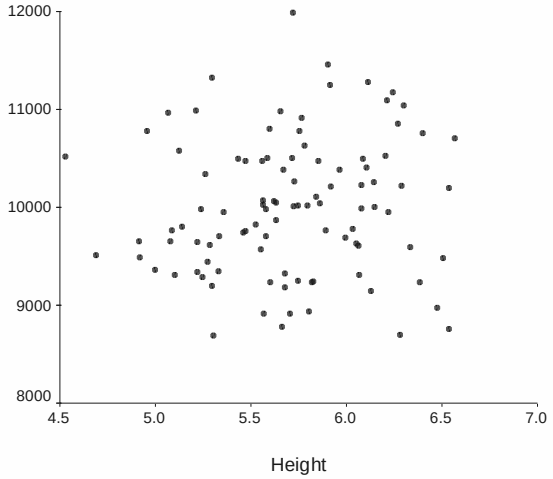


Fig. 1.11: Scatter between height of an individual and income in Rs. (per month) of the individual.

Self Assessment Questions

Check whether the following statements are true or false.

1) Positive correlation means as X increases Y increase. (True/False)
2) Negative correlation means as X decreases Y decreases. (True/False)
3) Generally, in a scatter, lower scores on X are paired with lower scores on Y for positive correlation. (True/False)
4) The scatter diagram indicates the direction of the relationship. (True/False)

1.6 CORRELATION: THE STRENGTH OF RELATIONSHIP

So far we have discussed the direction of relationship. Obviously, any reader would ask a question "how strong is the relationship"? The strength of relationship can be determined from the degree of linearity of the relationship. We need to know more about the correlation in order to understand strength of association.

The correlation between any two variables is expressed in terms of a number, usually called as correlation coefficient. The correlation coefficient is denoted by various symbols depending on the type of correlation. The most common is '*r*' (small 'r') indicating the Pearson's product-moment correlation coefficient. The representation of correlation between X and Y is r_{xy}. The range of the correlation coefficient is from –1.00 to + 1.00. It may take any value between these numbers, for example, – 0.78, – 0.54, – 0.21, + 0.02, + 0.35, + 0.98, etc. If the correlation coefficient is 1, then relationship between the two variables is perfect. This will happen if the correlation coefficient is – 1 or + 1. As the correlation coefficient moves nearer to + 1 or – 1, then the strength of relationship between the two variables increases. If the correlation coefficient moves away from the + 1 or – 1, then the strength relationship between two variables decreases (that is, it becomes weak). So correlation coefficient of + 0.84 (and similarly – 0.79, – 0.84, etc.) shows strong association between the two variables. Whereas, correlation coefficient of + 0.24 or – 0.24 will indicate weak relationship. Figure 1.12. The Range of Correlation Coefficient.

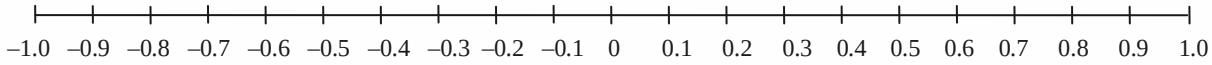


Fig. 1.12: The Range of Correlation Coefficient

The correlation coefficient also shows the direction of relationship. If the correlation coefficient has positive sign, then it shows positive correlation. If the correlation coefficient has negative sign, then it shows negative correlation. Some readers would mistakenly believe that negative correlation is weak than positive. Indeed, the sign (positive and negative) indicates the direction of relationship. The *sign* of correlation coefficient is not an indicator of *strength* of association. The strength of association of the two correlation coefficients with different signs (for example, +0.58 and – 0.58) is identical if the absolute value is identical.

Another way to understand strength of association is to understand the common variance between two correlated variables. The correlation coefficient is not percentage. So correlation of 0.30 does not mean 30% variance is common between two variables. The shared variance between two correlated variables can be calculated. I explain this point further. See, every variable has variance. We denote it as (variance of X). Similarly, Y also has its own variance (Variance of Y). In the previous block you have learned to compute them. From the complete variance of X, it shares some variance with Y. It is called as covariance. Figure 1.13 shown below explains the concept of shared variance. The circle X indicates the variance of X. Similarly, the circle Y indicates the variance of Y. The overlapping part of X and Y, indicated by shaded lines, shows the shared variance between X and Y. One can compute the shared variance.

Figure 1.13: The circles X and Y represent variances of X and Y respectively. The shaded part is the variance that is common between X and Y. It is covariance. Covariance indicates the degree to which X shares variance with Y. The shaded part is the variance common between X and Y.

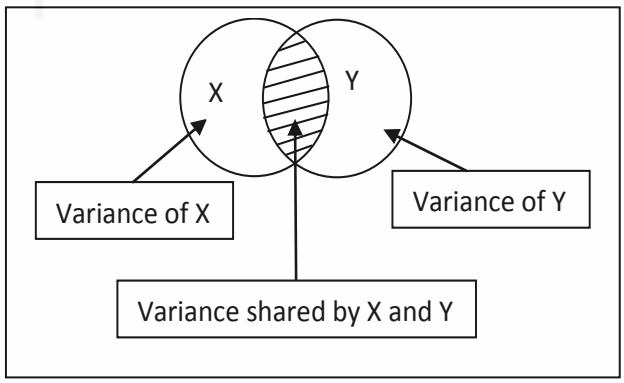


Fig. 1.13: The circles X and Y represent variances of X and Y respectively

To calculate the percentage of shared variance between X and Y (common variance), one needs to square the correlation coefficient (*r*). The formula is given below:

Percentage of common variance between X and Y= ×100 (eq. 1.2)

For instance, if the correlation between X and Y is 0.50 then the percent of variation shared by X and Y can be calculated by using equation 1.2 as follows:

Percentage of common variance between X and Y = × 100 = 0.50^2 × 100 = 0.25 × 100 = 25%

It indicates that, if the correlation between X and Y is 0.50 then 25% of the variance is shared by the two variables, X and Y. You would note that this formula is applicable to negative correlations as well. For instance, if r_{xy} = – 0.85, then shared variance is:

Percentage of common variance between X and Y = × 100 = – 0.85^2 × 100 = 0.7225 × 100 = 72.25%

Self Assessment Questions

1) What is correlation coefficient?

   .....

   .....

2) t is the range of correlation coefficient?

   .....

   .....

3) correlation coefficient a percentage?

   .....

   .....

4) to calculate common variance from correlation coefficient?

   .....

   .....

5) t is the percentage of variance shared by X and Y if the r_{xy} = 0.71?

   .....

   .....

6) t is the percentage of variance shared by X and Y if the r_{xy} = – 0.40?

   .....

   .....

1.7 MEASUREMENTS OF CORRELATION

Correlation can be calculated by various ways. The correlation coefficient is a description of association in the sample. In that sense it is a descriptive statistics. Various ways to compute correlation simply indicate the degree of association between variables *in the sample*. It can also be shown that distributional assumptions are *not* required to compute correlation as a descriptor of relationship. So it is not a parametric or nonparametric statistics.

The calculated sample correlation coefficient can be used to estimate population correlation coefficient. The sample correlation coefficient is usually denoted by symbol '*r*'. The population correlation coefficient is denoted by symbol '*ρ*'. It is Greek letter *rho*, pronounced as *row*

(Unfortunately, Spearman's correlation coefficient is also symbolised as *rho*. This may create some confusion among the readers. Therefore, I shall use symbol r_s for Spearman's *rho* as a sample statistics and ρ_s to indicate the population value of the Spearman's *rho*. Henceforth, I shall also clearly mention the meaning with which ρ is used in this block). When the population correlation coefficient is estimated from sample correlation coefficient, then the correlation coefficient becomes an inferential statistic. Inference about population correlation (ρ) is drawn from sample statistics (r). The population correlation (ρ) is always unknown. What is known is sample correlation (r). The population indices are called as parameters and the sample indices are called as statistics. So ρ is a parameter and r is a statistics. While inferring a parameter from sample, certain distributional assumptions are required. From this, you can understand that the descriptive use of the correlation coefficient does not require any distributional assumptions.

The most popular way to compute correlation is 'Pearson's Product Moment Correlation (r)'. This correlation coefficient can be computed when the data on both the variables is on at least equal interval scale or ratio scale (we will learn about them on Unit 3). Apart from Pearson's correlation there are various other ways to compute correlation. Spearman's Rank Order Correlation or Spearman's *rho* (r_s) is useful correlation coefficient when the data is in rank order. Similarly, Kendall's *tau* (τ) is a useful correlation coefficient for rank-order data. Biserial, Point Biserial, Tetrachoric, and Phi coefficient, are the correlations that are useful under special circumstances. Apart from these, multiple correlations, part correlation and partial correlation are useful ways to understand the associations (Please note that the last three require more than two variables).

1.8 CORRELATION AND CAUSALITY

As we have noted earlier, correlation does not imply causality. If the correlation between X and Y is high, then it is incorrect to conclude that X is a cause of Y or Y is a cause of X. There are various possibilities of causation when the two variables show high correlation. The first possibility is that the X is cause of Y. Second possibility is that Y is a cause of X. The third possibility is that both X and Y are caused by third variable Z. The correlation coefficient or its significance does not confirm any of these possibilities. Hence, it is always premature to conclude abut causality from correlation. Having said so, correlation can be used to infer causation. For doing so, one need to understand the complex issue of causal modelling (exogenous and endogenous variable, etc.). The regression analysis is a predictive analysis. It uses the logic underlying the correlations. But it must be remembered that statistics does not give causality. Statistics only confirms or rejects properly developed causal models. The development of causal model is a theoretical exercise which requires sufficient understanding of the content area, that is, psychology in our case.

1.9 USES OF CORRELATION

The correlations can be used to understand the degree and direction of association between two variables. In addition it can be used for various other purposes. Regression analysis is a predictive analysis. The association between the variables forms the basis for regression analysis. It can also be

used in reliability analysis. For example, the test retest reliability requires data on same set of subjects on the same test on two different occasions. Then these two sets are correlated to obtain test-retest reliability. Other type of reliabilities also uses the correlation logic. Correlation can be used in estimating validity. The concurrent validity for example, is correlation between two scales/tests measuring similar constructs. Factor analysis is almost a backbone of research in psychology. The factor analysis requires the data in terms of correlation among the variable. From correlation among the variables, the factors are calculated. It is useful in various other multivariate techniques. The correlation is an very important and useful technique in psychological research.

1.10 LET US SUM UP

We have learned the basic information about the correlation. Now you will be able to describe and explain concept of correlation. The correlation has positive or negative direction. You have also understood the strength of relationship. We have learned to plot the scatter diagram. We have understood difference between linear and nonlinear relationships. We have obtained information about various measures of correlation. We have also understood that conceptual issues in correlation, like causality. Now you should be able to create problems suitable for correlation analysis. In the next units, we shall learn the procedures to solve those problems.

1.11 UNIT END QUESTIONS

1) Define correlation and explain scatter diagram.
2) How would you draw a scatter diagram?
3) Discuss the different types of correlation.
4) What are the range of correlation?
5) How would you measure correlation?
6) What do you understand by the term strength of correlation. Explain with an example
7) What are the various uses of correlation?

1.12 SUGGESTED READINGS

Aron, A., Aron, E. N., Coups, E.J. (2007). *Statistics for Psychology*. Delhi: Pearson Education.

Minium, E. W., King, B. M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education*. Singapore: John-Wiley.

Guilford, J. P., & Fructore, B. (1978). *Fundamental Statistics for Psychology and Education.* N.Y.: McGraw-Hill.

Wilcoxon, R. R. (1996). *Statistics for Social Sciences*. San Diego: Academic Press.

Answers to SAQs.

Answers to “Answers to self assessment questions 1.1”

Scatter showing relationship between duration of practice (in months) and time taken (in seconds)

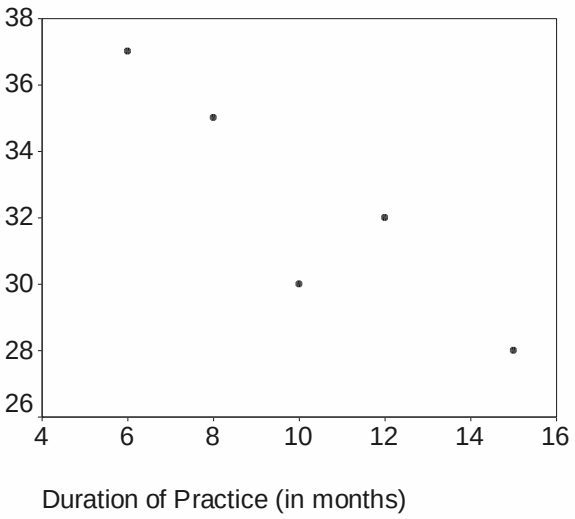


Fig. 1.5: Scatter for time taken and duration of practice

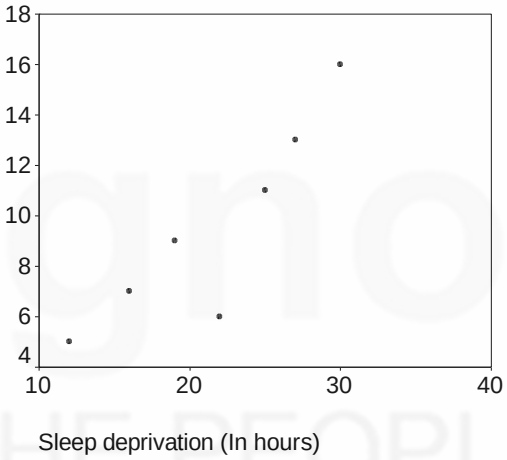


Fig. 1.6: Scatter showing relationship between sleep deprivation and irritability

Answers: 1 = True, 2 = False, 3 = True, 4 = True, 5 = True.

Test yourself 1.3:

We have already plotted the scatter for table 1 to table 6. Your job is to identify the direction of relationship for those data sets.

Answers: Table 1.1 = Positive, Table 1.2 = Positive, Table 1.3 = Positive, Table 1.4 = Negative, Table 1.5 = Negative, Table 1.6 = Positive.

1.4. Answers:

A number expressing the relationship between two variables.

The range of correlation coefficient is from –1.00 to + 1.00.

No. Correlation is not a percentage. But it can be converted into percentage of variance shared.

Common variance is calculated from correlation coefficient by using a formula: $r_{xy}^2 \times 100$.

50.41%

16.00%

UNIT 2 PRODUCT MOMENT COEFFICIENT OF CORRELATION

Structure

2.0 INTRODUCTION

We have studied theoretical aspects of correlation in the previous unit. In Unit 2, we shall learn more about Pearson's product-moment coefficient of correlation. Pearson's product-moment correlation coefficient is developed by Karl Pearson 1986. This correlation coefficient is usually calculated on continuous variables (if the data are in rank-order, frequencies, dichotomous, etc. then we can use other procedures). In this unit we will learn about the logic of correlation, computational steps, significance testing, and interpretation of Pearson's product moment correlation. We will also learn about other types of correlations which are specialised Pearson's correlation coefficients.

2.1 OBJECTIVES

On completion of this unit, you will be able to:

- Understand what are the pre requisites for calculation of correlation;
- You will get to know the formulae for mean, variance and covariance;

- Understand what is product moment coefficient of correlation or Pearson's r;
- Understand the coefficient of correlation formula and how it is to be used to calculate the correlation coefficient;
- Understand how to interpret the correlation coefficient;
- Understand the direction and strength of correlation as well as issues in interpreting correlation coefficient;
- Understand how to calculate r from raw scores and which formula to use for the same;
- Know how to test the significance of r and the related assumptions; and
- Know how to and when to use other methods of correlation when r is not appropriate to use.

2.2 BUILDING BLOCKS OF CORRELATION

Understanding product moment correlation coefficient requires understanding of mean, variance and covariance. We shall understand them once again in order to understand correlation.

2.2.1 Mean

Mean of variable X (symbolised as) is sum of scores ($\sum$ **X)** divided by number of observations (*n*). The mean is calculated in following way.

$$\sum \mathbf{X} / \mathbf{N}$$

You have learned this in the first block. We will need to use this as a basic element to compute correlation.

2.2.2 Variance

The variance of a variable X (symbolised as V) is the sum of squares of the deviations of each X score from the mean of X (σ) divided by number of observations (*n*).

$$V = \sigma^2 / N$$

You have already learned that standard deviation of variable X (σ) is square root of variance of X $\sqrt{(fx^2 / N)}$.

2.2.3 Covariance

The covariance between X and Y (or) can be stated as

$$CV=\sum xy / N$$

xy = product of the deviations of X and Y group scores from their respective means
x = deviation from the mean X
y = Deviations from the Mean Y
N= the total number of subjects

Covariance is a number that indicates the association between two variables X and Y. To compute covariance, first you need to calculate deviation of each score on X from its mean ($M1_{.}$) and deviation of each score on Y from its mean (M2). Then multiply these deviations to obtain their product. Then, sum these products. Divide this sum by number of observations (*n*). The resulting number is covariance. Let's quickly learn to compute the covariance. We shall use the data from table 1.5 for this purpose. We shall call duration of practice as X and Time taken as Y.

Table 2.1: Calculation of covariance from data in table 1.5

Duration of Practice (In months) X	**Time taken (In Seconds) Y**	**x** $_{.}$**Deviation from Mean = 10**	**y** $_{.}$**Deviation from Mean = 32**	**Xy**
10	30	0	– 2	0 × (– 2) = 0
12	32	2	0	0
8	34	– 2	2	– 4
14	27	4	– 5	– 20
6	37	– 4	5	– 20
$\sum X = 50$	$\sum Y = 160$			$_{.}\sum xy = -44$
$_{.} = 50/5 = 10$	$_{.} = 160/5 = 32$			

So the covariance between Duration of Practice (X) and Time Taken (Y) is – 8.8. (-44 / 5)

Self Assessment Questions

1) Define mean?

.....

.....

.....

.....

2) Define variance?

.....

.....

.....

.....

3) Define covariance?

.....

.....

.....

.....

4) For a following data calculate mean, variance and covariance of X and Y: X= 2,4,6,8,3; Y = 5,8,7,9,6

2.3 PEARSON'S PRODUCT MOMENT COEFFICIENT OF CORRELATION

Now we shall understand the formula for computing Pearson's product-moment correlation coefficient. We will also solve an example for the same.

2.3.1 Formula

Since you have learned to compute the covariance between X and Y, now we shall learn to compute Pearson's product-moment correlation coefficient (*r*). The Pearson's product moment correlation coefficient (*r*) can be defined as ∑xy / N. σx x σy

In this, there is covariance between X and Y, there is standard deviation of X and there is standard deviation of Y. Since, it can be shown that this is the maximum value correlation can take. So the maximum value of correlation coefficient is bound to be 1. The sign of Pearson's *r* depends on the sign of products of x and y from their deviations. If the product is negative, then *r* will be negative and if is positive then *r* will be a positive value. The denominator of this formula is always positive. This is the reason for a – 1 to + 1 range of correlation coefficient. By substituting covariance equation for covariance we can rewrite equation 2.4 as (equation 2.5)

$$r = \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}}{S_X S_Y}$$

By following a simple rule, a÷ b ÷ c = a ÷ (b × c), we can rewrite equation 2.5 as follows:

(equation 2.5.)

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{nS_X S_Y}$$

2.3.2 Numerical Example

We shall use this formula for computing Pearson's correlation. Let us look at another example for computing product moment correlation coefficient. Look at the example provided here. A researcher was interested in studying relationship between personality and creativity. She chose the Five-Factor model of personality as a personality conceptualisation. She realised that the Five-Factor theory proposes that one of its five main dimensions, the openness to experience (O), is uniquely related to creativity (McCrae, 1987). So she chose 12-item NEO-FFI Openness scale to measure openness. She chose the Torrance Test of Creative Thinking (Figural) for measurement of creativity. She understands that originality subscale of TTCT fits to the universal definition of creativity as uniqueness. Finally, she administered both the instruments on ten subjects (in the real research, we should take sample larger than this, roughly around 50 to 100 to have sufficient power to statistics). The data she had obtained are given below in table 2.2. Now we shall compute the Pearson's product moment correlation coefficient on this data.

Table 2.2: Calculation of Pearson's Correlation Coefficient

Subject	Openness (X)	Creativity (Y)					
1	10	26	-2	0	4	0	0
2	8	23	-4	-3	16	9	9
3	9	23	-3	-3	9	9	0
4	13	26	1	0	1	0	2
5	11	24	-1	2	1	4	0
6	14	30	2	4	4	16	0
7	16	27	4	1	16	1	0
8	12	27	0	1	0	1	8
9	15	29	3	3	9	9	9
10	12	25	0	-1	0	1	4
n = 10	=120	=260			= 60	= 50	∑xy = 44
	= 12	= 26			√60/10 =6	√50/10=5	
	√60/10 =√6 =2.45					√5 =2.24	
	R= ∑xy / N x σx x σy						
	r=44 / (10)(2.45)(2.24) = + 0.803						

Step 1. You need scores of various subjects on two variables. We have scores on ten subjects on two variables, openness and creativity. Then list the pairs of scores on two variables in two columns. The order will not make any difference. Remember, same individuals' two scores should be kept together. Label one variable as X and other as Y.

Step 2. Compute the mean of variable X and variable Y. It was found to be 12 and 26 respectively.

Step 3. Compute the deviation of each X score from its mean ( ) and each Y score from its own mean ( ). This is shown in the column labeled as and . As you have learned earlier, the sum of these columns has to be zero.

Step 4. Compute the square of and . This is shown in next two columns labelled as and . Then compute the sum of these squared deviations of X and Y. the sum of squared deviations for X is 60 and for Y it is 50. Divide them by *n* to obtain the standard deviations for X and Y. The was found to be 2.45. Similarly, the was found to be 2.24.

Step 5. Compute the cross-product of the deviations of X and Y. These cross-products are shown in the last column labeled as. Then obtain the sum of these cross-products. It was found to be 44. Now, we have all the elements required for computing *r*.

Step 6. Use the formula of *r* to compute correlation. The sum of the cross-product of deviations is numerator and n, , are denominators. Compute *r*. the value of *r* is 0.803 in this example.

Self Assessment Questions

1) What is the relationship between covariance and correlation

.....

.....

.....

.....

.....

2) How to calculate the r in the correlation formula?

.....

.....

.....

.....

.....

2.4 INTERPRETATION OF CORRELATION

Now we shall learn to interpret correlation coefficient. We will interpret the direction and strength of the correlation.

2.4.1 Understanding Direction

The correlation between openness to experience (X) and creativity (Y) is +0.803. The direction of this relationship is positive. This means that as the openness to experience increases, creativity also increases and vice-versa. The direction of the correlation is determined by the cross-product of . If most of these cross-products are positive then correlation is would be positive. It will happen if a pair of deviations has same signs. Which means that when a given score of X deviates from negatively, then the corresponding Y score has to deviate from negatively and vice-versa. The figure 2 shown below indicates a positive relationship.

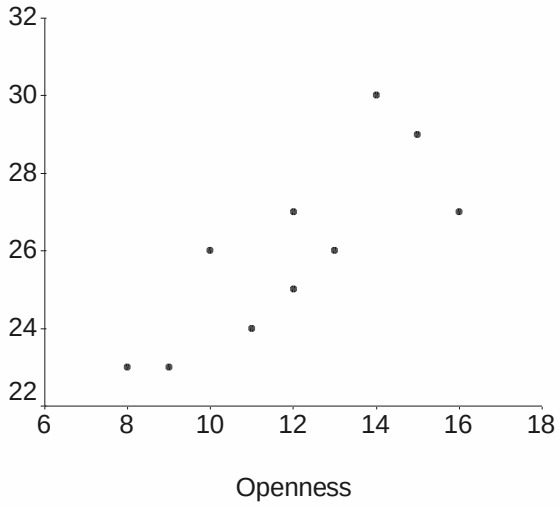


Fig. 2.1: Scatter of the relationship between openness and creativity for the data table 2.2

2.4.2 Understanding Strength

The strength of association between two variables can be understood from the value of the coefficient. The coefficient of 0.80 indicates a high correlation. It shows that the relationship between the two variables is certainly a close one. But it is still far from perfect. The low score on openness to experience decreases the chances of being creative. Similarly, high score on openness increases once chances to be creative. The common variance between openness and creativity can be calculated.

$$CV = + 0.80^2 \times 100 = 0.64 \times 100 = 64.00\%$$

The variance in creativity that is shared by openness and vice-versa is 64.00%. It is certainly a high percent variance one variable is explaining in the other. It must be kept in mind that this correlation is computed on a data of ten individuals. More representative data might change the picture. You would also realise that no assumptions were required for computing correlation.

2.4.3 Issues in Interpretation of Correlation

The interpretation of the correlation on the basis of the strength and direction looks very straightforward exercise. One should keep in mind that this interpretation of the correlation coefficient is subjected to other aspects of the data. These aspects are range of the scores, reliability of measurement, and presence of outliers. Let's look at them.

Restricted Range

While correlating X and Y, it is expected that both the variable are measured with full range. For example, suppose we want to study the correlation between hours spent in studies and marks. We are supposed to take students who have varying degree of hours of studies, that is, we need to select students who have spent very little time in studies to the once who have spent great deal of time in studies. Then we will be able to obtain true value of the correlation coefficient. But suppose we take a very restricted range then the value of the correlation is likely to reduce. Look at the following examples the figure 2.1a and 2.1b.

Figure 2.2: Scatters showing the effect of range on correlation.

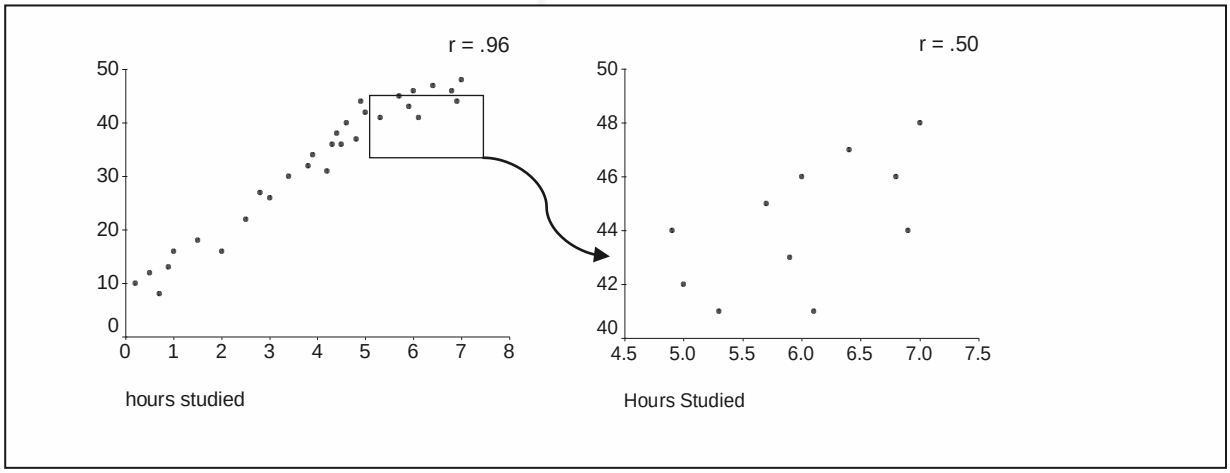


Fig. 2.2a: Scatter showing full range on both variables

Fig. 2.2b: Scatter with restricted range on hours studied

The figure 2.2a is based on a complete range. The figure 2.2b is based on the data of students who have studied for longer durations. The scatter shows that

when the range was full, the correlation coefficient was showing positive and high correlation. When the range was restricted, the correlation has reduced drastically.

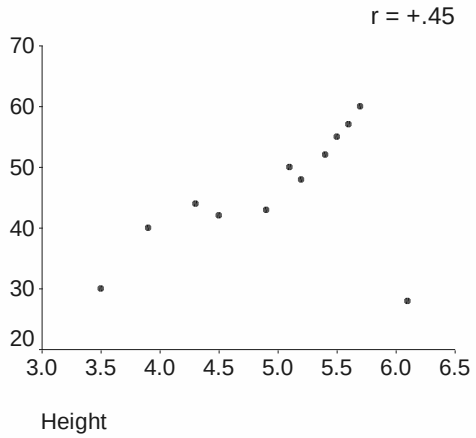
You can think of some such examples. Suppose, a sports teacher selects 10 students from a group of 100 students on the basis of selection criterion, their athletic performance. Now the actual performance of these ten selected students in the game was correlated with the selection criterion. The very low correlation between selection criterion and actual game performance. She would believe that selection criterion is not related with actual game performance. Is it true..? Why so…? Now you will realise that the range of the scores on selection criterion is extremely restricted (because these ten students were only high scorers) and hence the relationship is weak. So note that whenever you interpret correlations, the range of the variables is full. Otherwise the interpretations will not be valid. Figure 2.2. above shows the effect of range on correlation.

Unreliability of Measurement

This is an important issue in psychological research. The psychological test has to have reliability. The reliability refers to the consistency of the measurement. If the measurement is consistent, then the test has high reliability. But at times one of the variable or both the variables may have lower reliability. In this case, the correlation between two less reliable variable reduces. Generally, while interpreting the correlation, the reliability is assumed to be high. The general interpretations of correlations are not valid if the reliability is low. This reduction in the correlation can be adjusted for the reliability of the psychological test. More advanced procedures are available in the books of psychological testing and statistics.

Outliers

Outliers are extreme score on one of the variables or both the variables. The presence of outliers has extremely deterring impact on the correlation values. The strength and degree of the correlation are affected by the presence of outlier. Suppose you want to compute correlation between height and weight. They are known to correlate positively. One of the scores has low score on weight and high score on height (probably, some anorexia patient). This extreme score is called an outlier. Let us see the impact of an outlier observation on correlation. Without the outlier, the correlation is 0.95. The presence of an outlier has drastically reduced a correlation coefficient to 0.45. Figure 2.2. shows the Impact of an outlier observation on correlation. Without the outlier, the correlation is 0.95. The presence of an outlier has drastically reduced a correlation coefficient to 0.45.



Curvilinearity

We have already discussed the issue of linearity of the relationship. The Pearson's product moment correlation is appropriate if the relationship between two variables is linear. The relationships are curvilinear then other techniques need to be used. If the degree of cruviliniarity is not very high, high score on both the variable go together, low scores go together, but the pattern is not linear then the useful option is Spearman's *rho*. We shall discuss this technique in the next unit.

Self Assessment Questions

1) Which aspects of correlation coefficient help us in its interpretation?

.....

.....

.....

.....

.....

.....

2) If correlation is 0.45 then what is the direction of correlation? Interpret.

.....

.....

.....

.....

.....

.....

3) If correlation is – 0.68 then what is the direction of correlation? Interpret.

.....

.....

.....

.....

.....

.....

4) Which factors brings problems in the interpretation of correlation?

.....

.....

.....

.....

.....

5) What is an outlier?

.....

.....

.....

.....

.....

6) What is full and restricted range?

.....

.....

.....

.....

.....

2.5 USING RAW SCORE METHOD FOR CALCULATING *r*

Apart from the method we have learned in the earlier section to calculate Pearson's *r*, we can use another variation of the formula. It is called as raw score method. First we will understand how the two formulas are similar. Then we will solve a numerical example for the raw score method. We have learned following formula for calculating *r*.

2.5.1 Formulas for Raw Score

We have already learnt following formula of correlation. This is a deviation score formula.

The denominator of correlation formula can be written as $N \times \sigma x \times \sigma y$

And the numerator of the correlation formula can be written as $\sum xy$.

Thus r can be calculated by using the formula $\sum xy / N \times \sigma x \times \sigma y$

There is another formula that is raw score formula which reads as given below:

$\sum X^2 - X^2 / N$; $\sum Y^2 - Y^2 / N$; $\sum XY / \{\sum X^2 - X^2 / N\} \times \{\sum Y^2 - Y^2 / N\}$

Table:2.3: Calculation of r by using the raw scores

Subject	Openness (X)	Creativity (Y)	X^2	Y^2	XY
1	10	26	100	676	260
2	8	23	64	529	184
3	9	23	81	529	207
4	13	26	169	676	338
5	11	24	121	576	264
6	14	30	196	900	420
7	16	27	256	729	432
8	12	27	144	729	324
9	15	29	225	841	435
10	12	25	144	625	300
Summation	120	260	1500	6810	3164
	= 12	= 26			
$1500 - (120)^2/10 = 60$					
$6810 - (260)^2 / 10 = 50$					

The students may find one of the methods easier. There is nothing special about the methods. One should be able to correctly compute the value of correlation.

2.6 SIGNIFICANCE TESTING OF *r*

Statistical significance testing refers to testing a hypothesis about a population parameter by using sample statistics. When we calculate correlation as a descriptive index of sample, we need not test statistical significance of correlation. The significance need to be tested when we are interested in understanding whether the obtained value of correlation is greater than the chance finding. One may obtain a correlation of 0.22 between health and income on a sample of 30 individuals. If you take another sample you may get a different value. This simply refers to sample specific finding. Researchers are always interested in knowing whether the obtained findings are due to sample specific errors or this is correct representation of population. To do this, one need to test statistical significance of the correlation coefficients. Testing significance of correlation coefficient is complex issue. The reason for the complexity lies in the distribution of the population correlation. We will not enter into the complexities of this issue. The *t*-distribution and z-distribution are used to test statistical significance of *r*.

As you have learned, we need to write a null hypothesis (H_O) and alternative hypothesis (H_A) for this purpose. The typical null hypothesis states that population correlation coefficient between X and Y (ρ_{xy}) is zero.

$$H_O : \rho_{xy} = 0$$
$$H_A: \rho_{xy} \neq 0$$

If we reject the H_O then we accept the alternative (H_A) that the population correlation coefficient is other than zero. It implies that the finding obtained by us is not a sample specific error. Sir Ronald Fisher has developed a method of using *t*-distribution for testing this null hypothesis. The degrees of freedom (*df*) for this purpose are $n - 2$. Here *n* refers to number of observations. We can use Appendix C for testing the significance of correlation coefficient. Appendix C provides critical values of correlation coefficients for various degrees of freedom. Let us learn how to use the Appendix C. We shall continue with the example of health and income. Once we learn it, we shall use it for creativity and openness example.

The correlation between health and income is +.22 obtained on 30 individuals. We decide to do this test at 0.05 level of significance, so our $\alpha = .05$. We also decide to apply two-tailed test. The two-tailed test is used if alternative hypothesis is non-directional, i.e. it does not indicate the direction of correlation coefficient (meaning, it can be positive or negative) and one-tail test is used when alternative is directional (it states that correlation is either positive or negative). Let's write the null hypothesis and alternative hypothesis:

$$H_O : \rho_{\text{health income}} = 0$$
$$H_A : \rho_{\text{health income}} \neq 0$$

Now we will calculate the degree of freedom for this example.

$$df = n - 2 = 30 - 2 = 28$$

So the *df* for this example is 28. Now look at Appendix C. Look down the leftmost *df* column till you reach *df* 28. Then look across to find correlation coefficient from column of two-tailed test at level of significance of 0.05. You will reach the critical value of *r*:

$$r_{\text{critical}} = 0.361$$

Because the obtained correlation value of +0.22 is less than critical (tabled) value, we accept the null hypothesis that there is no significant correlation between health and income in the population. This method is used regardless of the sign of the correlation coefficient. We use the absolute value (ignore the sign) of correlation.

Now let's do the significance testing of correlation for our earlier example of openness and creativity. The obtained correlation coefficient is 0.803 with *n* of 10. The null hypothesis is there is no correlation between creativity and openness in the population. The alternative is there is a correlation between them in population.

$$H_O : \rho_{\text{openness creativity}} = 0$$
$$H_A : \rho_{\text{openness creativity}} \neq 0$$

Now we will calculate the degree of freedom for this example.

$$df = n - 2 = 10 - 2 = 8.$$

Now look at Appendix C. For the $df = 8$ and two-tailed $\alpha = 0.05$, we found out the critical value of *r*:

$r_{\text{critical}} = 0.632$

Because the obtained correlation value of +0.803 is greater than critical value of 0.632, we reject the null hypothesis that there is no correlation between openness and creativity in the population. We accept that there exists a correlation between openness and creativity in the population.

2.6.1 Assumptions for Significance Testing

One may recall that simple descriptive use of correlation coefficient does not involve any assumption about the distribution of either of the variable. However, using correlation as an inferential statistics requires assumptions about X and Y. these assumptions are as follows. Since we are using t-distribution, the assumptions would be similar to *t*.

Independence among the pairs of score.

This assumption implies that the scores of any two observations (subjects in case of most of psychological data) are not influenced by each other. Each pair of observation is independent. This is assured when different subjects provide different pairs of observation.

The population of X and the population of Y follow normal distribution and the population pair of scores of X and Y has a normal bivariate distribution.

This assumption states that the population distribution of both the variables (X and Y) is normal. This also means that the pair of scores follows bivariate normal distribution. This assumption can be tested by using statistical tests for normality.

It should be remembered that the *r* is a robust statistics. This means that some violation of assumption would not influence of the distributional properties of *t* and the probability judgments associated with the population correlation.

Self Assessment Questions

1) Explain this statement: $H_O : \rho = 0$

.....

.....

.....

.....

.....

.....

2) Explain this statement: $H_A : \rho \neq 0$

.....

.....

.....

.....

.....

.....

3) If the sample size is 34, then what is *df* ?

.....

.....

.....

.....

.....

4) Which distribution is to test the significance of *r* ?

.....

.....

.....

.....

.....

5) State the assumptions of significance testing of *r* ?

.....

.....

.....

.....

.....

2.7 OTHER TYPES OF PEARSON'S CORRELATIONS

So far we have discussed Pearson's correlation for continues measurement of X and Y. We can calculate Pearson's correlation when the variable is dichotomous (having two levels) for one or both the variables. These correlations are popularly known as Point-Biserial and Phi coefficients. The dichotomous variables are those that take two levels. For example, male-female, pass-fail, urban-rural, etc. We shall introduce ourselves to these correlations. We are not going to solve the numerical of these correlations.

2.7.1 Point-Biserial Correlation (r_{pb})

When one of the variable is dichotomous, and the other variable is continuous, then the Pearson's correlation calculated on this data is called as Point-Biserial Correlation (r_{pb}). Suppose, we want to correlate marital status with satisfaction with life. Then we take marital status at two levels: married and unmarried. The satisfaction with life can be measured by using a standardised test of 'Satisfaction with Life'. Now we have satisfaction with life as a continuously measured variable and marital status as a dichotomous variable. In this case we shall use Point-Biserial Correlation (r_{pb}).

2.7.2 Phi Coefficient (phi)

Point-Biserial Correlation (r_{pb}) was useful when one of the variable was dichotomous. If both the variables (X and Y) are dichotomous, then the Pearson's Product Moment Correlation calculated is called as Phi coefficient (phi).

Suppose, you are interested in investigating relationship between employment status and marital status. Employment status is dichotomous having two levels, employed and unemployed. Similarly, the marital status can be dichotomised by taking two levels: married and unmarried. Now we have both the variables that are dichotomous. The correlation that is useful in such instances is Phi coefficient (phi).

2.8 LET US SUM UP

In this unit we have learnt about the Pearson's correlation coefficient. Initially, we have started with basic statistics. Then we have also studied that correlation is a function of covariance. Pearson's correlation coefficient is useful to calculate correlation between two relatively continuous variables. Calculation of Pearson's correlation is possible with two methods: Deviation score method and raw scores method. The coefficient obtained can be interpreted on the basis of the strength and the direction. Range, unreliability of the measurement, outliers, and curvilinearity are the factors that need to be considered while interpreting the correlation coefficient. Using correlation coefficient as a descriptive statistics does not require assumptions. However, the use of sample correlation to estimate population parameter (ρ) requires assumptions. We can use Appendix C to test the significance of the obtained value of the correlation. It uses t-distribution for calculating the probabilities. The special cases of Pearson's correlation are known as Point Biserial coefficient and Phi coefficient. We can hope that now you can judicially be able to use this coefficient for understanding the correlation between two variables.

2.9 UNIT END QUESTIONS

Problems:

Following are scores on X and Y variables. Plot scatter diagram. Compute Pearson's Product Moment Correlation Coefficient between X and Y.

X	Y
12	1
14	2
16	3
11	4
17	5

1) A researcher was interested in understanding the relationship between hopelessness and depression. So she administered BDI and BHS to 10 patients. The data are given below. Do appropriate statistics and comment on the relationship between the two constructs. Also plot a scatter diagram. Compute covariance between them. Test the null hypothesis that population correlation coefficient is zero. State whether null hypothesis is accepted or rejected. Interpret the results.

BHS	BDI
3	3
5	2
7	5
11	6
8	4
4	8
12	13
14	10
10	7
6	4

2) A researcher in clinical psychology was curious to know the relationship between number of relaxation sessions attended by the client and the reduction in anxiety. She hypothesised that as the number of therapy sessions increase, the anxiety will reduce. She took a data of eight patients. There data on number of session of therapy attended so far and anxiety scores were obtained. The data are given below. Higher scores on anxiety scale indicate higher anxiety. Do appropriate statistics and comment on the relationship between number of therapy sessions attended and anxiety. Also plot a scatter diagram. Compute covariance between them. Test the null hypothesis that population correlation coefficient is zero. Interpret the results.

Number of sessions attended	Anxiety Scores
2	14
4	12
7	6
8	10
9	12
10	5
12	3
13	8

3) A sports psychologist has developed a ‘sport achievement scale’ to assess achievement motivation in sports. He thought of correlating the scale scores with actual sports achievements of the sportspersons to validate the scale. The data of 15 sportspersons are given below. The first column indicates the scores on newly developed ‘sports achievement scale’ and second column shows the data of their actual sports achievements. Plot the scatter. Compute covariance. Calculate correlation coefficient. State the null hypothesis. Test the null hypothesis.

Achievement Scale Scores	Actual Achievement
5	6
7	4
9	3
11	8
12	6
14	9
11	5
13	7
15	6
16	9
17	11
19	12
22	6
24	13
25	5

2.10 SUGGESTED READINGS

Aron, A., Aron, E. N., Coups, E.J. (2007). *Statistics for Psychology*. Delhi: Pearson Education.

Minium, E. W., King, B. M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education*. Singapore: John-Wiley.

Guilford, J. P., & Fructore, B. (1978). *Fundamental Statistics for Psychology and Education.* N.Y.: McGraw-Hill.

Wilcox, R. R. (1996). *Statistics for Social Sciences*. San Diego: Academic Press.

Answers:

1) Mean can be defined as the total number scores divided by the total no.of subjects

2) Variance can be defined as the square of standard deviation .

3) Covariance can be defined as the sum of the product of deviations from the mean in the groups divided by the total number of subjects. .

4) Mean X = 4.6, variance X = 5.8, Mean Y = 2.5, variance Y = 2.5, covariance(X,Y) = 3.25.

2.2. SAQ Answers:

Covariance divided by standard deviation of X and Y will be correlation.

First, calculate deviation of each X from mean of X ( ) and deviation of each Y from mean of Y ( ). Then multiply these deviation to obtain ( )( ) this product term for each subject. Then do the summation of the product term to get , which is numerator of correlation formula.

Answers: 2.3

The direction and the strength of correlation.

Positive. As X increases Y increase and as X decreases Y decrease. The strength is moderately low since the common variance is 20.25%.

Negative. As X increases Y decrease and as X decreases Y increase. The strength is moderately high since the common variance is 46.24%.

The factors that bring ambiguity to the interpretation of correlation are range, reliability of measurement, outliers, and curvilinearity.

Outlier is an extreme value on any one variable or combination of variable.

When all possible values (lowest to highest) are measured, then the range of a variable is full. When specific values (only high, only low, or only middle) are measured, then the range is said to be restricted.

Answers: for 2.4

This is null hypothesis which states that the population correlation coefficient is zero. We need to test this hypothesis.

This is alternative hypothesis which states that the population correlation coefficient is **not** zero. Alternative hypothesis is accepted if the null hypothesis is rejected.

$df = n - 2 = 34 - 2 = 32$.

t-distribution. (i) The independence among the pairs of score

2. The population of X and the population of Y follow normal distribution

3. The population pair of scores of X and Y has a normal bivariate distribution

Answers: Unit End questions.

+ 0.434

Covariance between BHS and BDI is 7.8. The correlation coefficient is 0.70.

Covariance between number of sessions attended and anxiety scores is – 9.09.

Correlation between number of sessions attended and anxiety scores is – 0.71

Covariance between sports achievement scale scores and actual achievement in sports scores is 0.50

UNIT 3 INTRODUCTION TO NON-PARAMETRIC CORRELATION

Structure

3.0 INTRODUCTION

As per the syllabus, the title of this unit is Introduction to Nonparametric Correlations. Actually, the correlation as a descriptive technique cannot be considered as parametric or non-parametric when used as a descriptor of the sample. The Pearson's correlation also requires no assumptions for the calculation of correlation coefficient. The assumptions are required only when the inference is drawn about population parameter. Even the techniques like Spearman's *rho* (r_s) and Kendall's *Tau* (τ) are, in reality, rank-based methods. In fact, it can be shown that Spearman's *rho* (r_s) is a just a special case of Pearson's correlation. But I shall use this unit to introduce you to the non-parametric methods. If you look at contents, you will realise that it is discussion of issues related to statistical inference, measurement, and ranking. Since you have never leant about the non-parametric statistics, this unit provides you with an opportunity to update yourselves about the non-parametric statistics. This unit will also help you to understand advantages and disadvantages of non-parametric methods. This information would be extremely useful when you will learn the application of other non-parametric techniques.

This unit will also help you in understanding different scales of measurement. These scales of measurement constitute an important basis for applying different correlation techniques. We have already learned in the last unit that

the Pearson's correlation is possible with interval or ratio data. In this unit we shall learn more about these scales of measurement. We shall also learn to rank order the interval or ratio scale data. These things would help us in understanding the next unit well. Indeed, some of the aspects would be more useful when you will learn other non-parametric statistics (e.g., U-test, median test, etc.). This unit is written by maintaining this broad approach.

We have seen in the last two units that correlation is a useful technique to understand association between two variables. We have also learned to compute Pearson's correlation. But if the data are not in ratio or interval scale or assumptions of Pearson's *r* are not satisfied, then there are some other alternatives available. In this unit we shall introduce ourselves to such conditions. We will use this opportunity to understand more about parameter estimation. Since you are absolutely fresh to the non-parametric statistics, I shall also provide you with introduction to the non-parametric statistics. We shall also learn to rank the data. This will help us in the unit to do the calculations. Most of this unit is a theoretical discussion, which is essential to understand next unit and other non-parametric procedures.

3.1 OBJECTIVES

Having read this unit, you will be able to:

- Describe the parameter estimation;
- Differentiate between parametric and non-parametric statistics;
- Evaluate the conditions in which non-parametric statistics is used;
- Obtain knowledge about the four scales of measurement and when they have to be used; and
- To carry out the ranking of the data.

3.2 PARAMETER ESTIMATION

As I have clearly stated in earlier sections, the descriptive use of correlation techniques does not require any assumptions regarding the distribution of variables. In that case correlation would be a simple description of relationship between two variables. The relationship is indicated by a number called as a correlation coefficient. The correlation coefficient is an indicator of the degree of relationship between two variables. In all such instances, it is a descriptive statistics, which is neither parametric nor non-parametric. The parametric and non-parametric distinctions are applicable to inferential statistics only. Inferential statistics is not concerned about the sample findings. The primary aim of inferential statistics is to infer about the population parameter. The parameter is a population value. The statistics is a sample value. From the sample value, the population value is inferred or estimated. Hence, the inferential statistics requires some assumptions about the distributional properties of parameter. Under certain instances, the distributional properties are relaxed. This creates a distinction between parametric and non-parametric statistics.

3.2.1 Parameter: Unknown to be Estimated

Parameter refers to the unknown population value. To define parameter one needs to define the population. The value or the number obtained on population is called a parameter. For example, suppose we define parameter as average height of Indian women between the age group of 18 to 25. How to find out the height of all Indian women who are in this age range? Simple…take a measurement scale to measure height and measure the height of all Indian women in this age group. To get the average height, first obtain the sum of height of all women (Total) and divide that sum by number of women (*N*). This will give you average height or mean height ($\bar{M}$) of women in this age range. Right…? Ohhh…it's not simple at all…indeed it sounds near impossible. I guess census commissioner alone has the ability to do that … (and they have lot of other important information to collect). So a simple question of measuring height parameter is also not that simple. The only reason that makes it a tough task is the sheer number of Indian women in the given age range. It is not possible to measure the height of all Indian women in that age range. In fact, height and weight are simplest things to measure. Just imagine that you have to measure personality, intelligence, marital satisfaction, quality of life, etc. sounds dreadful…! Isn't it…? That is what parameter is…! It is the value obtained on population. And as you have correctly guessed, it is not possible to obtain this value by direct measurement. So we cannot obtain the value of parameter. The parameter is always an unknown value. Usually the parameter values are written in Greek letters and sample values are written in English. For example, if mean of the population has to be written, then we write it as μ (Greek letter *mu*). The sample mean is written as (M). Similarly, population size is denoted by N and sample size is denoted by *n*.

3.2.2 Statistics

Statistics is a value obtained on the sample. Sample is a subset of a population. Sample size is denoted by *n*. The statistics helps us to guess the value of parameter. Let's continue with the earlier example of height. It is clear that we cannot study all the Indian women in the age range of 18 to 25. So we decide to take a random sample of the size *n* from the population of the size *N* (let's believe for a moment that such a random sampling is possible, though in reality even that would be a tall ask. Often psychologist doing survey research do not obtain random sample.). Suppose, *n* is 100. Then we obtain the height of all women in that sample, do summation of that, and divide the obtained summation by *n* that is 100. We will obtain an average height or mean height of the sample. We denote this as $\bar{M}$. We can use this mean height or statistic to infer about the μ or the population parameter. The inference is drawn by making some assumptions about the distribution of the parameter. One of such assumption would be the population parameter is normally distributed. Then we would be able to infer the range of the values within which the parameter would lie with some probability. Higher the sample size, lower the range of interval within which the parameter would lie for a given probability. It can be shown that the standard deviation of the sampling distribution is the standard error of the statistics.

3.2.3 Hypothesis Testing

Hypothesis testing refers to the use of the estimation logic to test a statement about the population parameter. Usually, null hypothesis (H_O) and alternative

hypothesis (H_A) are written. These hypotheses should cover all possible conditions. For our earlier example, the null hypothesis is a statement about the population mean. It can be written as:

$$H_O: \mu = 0$$

The alternative hypothesis can be written as

$$H_O: \mu \neq 0$$

You will realise that the null hypothesis states that population mean is zero. The alternative states that the population mean is not zero. Both are covering all possible conditions. We can test the null hypothesis from a sample. At some level of probability $(1 - \alpha)$ we can test this null hypothesis. If we cannot accept the null hypothesis at a given level of probability then we reject it and accept the alternative. Null hypothesis is always written about a population value. It is tested by using sample statistics.

Self Assessment Questions

Decide whether the following statements are true of false.

1) Parameter is always known to us. True/ False
2) Statistics is computed on sample. True/ False
3) Parameter is estimated from the statistics. True/ False
4) Distributional assumptions are not at all required for parameter estimation. True/ False
5) Null hypothesis is written about the sample. True/ False

3.3 PARAMETRIC AND NON-PARAMETRIC STATISTICS

The parametric and non-parametric statistics are two broad classifications of statistical methods. The parametric methods for statistical analyses make assumptions about the distribution. The most common assumption is the assumption about the normal distribution. The assumption states that the population from which the sample is obtained is normally distributed. Another assumption that is commonly made is the assumption about the variances. This assumption is required for *t*-test, ANOVA, etc. For these statistics the assumption states that the population variances are equal. Similarly, homoscedasticity assumption is made in regression.

When the real data are collected on human subjects, the underlying assumptions may not be satisfied by the data. In such cases the application of the statistical technique is problematic. The distributions may not be normal or variances may not be equal. Using parametric methods under such circumstances may yield incorrect probabilities since the statistics may not follow the given probability distribution.

The approach that is commonly used under such circumstances is known as *non-parametric statistics*. The nonparametric statistics make comparatively fewer assumptions. They are often free from the restriction of the normal

distribution. And that is the reason they are a popular option. Nowadays, nonparametric statistics has advanced a lot. Various nonparametric methods have been developed by the statisticians. Earlier limitations of nonparametric statistical methods have been considerably reduced. It is false to believe that nonparametric methods do not make any assumption. They make comparatively fewer assumptions.

3.3.1 Advantages and Disadvantages of Non-parametric Statistics

Advantages:

1) There are certain advantages of non-parametric statistical methods.
2) There is no assumption of normal distribution.
3) The assumptions are comparatively fewer in non-parametric statistics.
4) Exact computation of probability is possible.
5) Exact confidence intervals, exact experimentwise error rates for multiple comparisons are computed without relying on assumptions that the underlying populations are normal.
6) These tests do not require the data on equal interval scale or ratio scale.
7) The data on ordinal or nominal scales can be used for the purpose of statistical analysis.
8) If the data are ordinal then the parametric statistics are not possible and the only option left is of using non-parametric.
9) The nonparametric methods are relatively insensitive to outlier observation or extreme scores. Parametric statistics is comparatively very sensitive to extreme values.
10) For young students, they are comparatively easier procedures to learn.
11) Most of the statistical packages for computer (SPSS, R, etc.) have these tests.
12) It has also facilitated the computation of exact p-values.

Even though, there are advantages there are also many disadvantages and these are given below:

Disadvantages:

1) One of the serious disadvantages is the loss of data.
2) Often the data need to be converted into ordinal form from interval or ratio scales. This may lead to loss of some important information.
3) Especially, two consecutive values are assigned the ranks with a difference of one.
4) Whereas, the real values may differ from each other by a great margin which is not captured in the procedure of ranking.
5) Another point that may go against these tests is when the assumptions are followed by the data then parametric tests have higher power than non-parametric tests.

6) With this we will note that non-parametric statistics are useful options when the assumptions are not followed. They are useful when the data are not on interval or ratio scale.

Self Assessment Questions

State whether the following statements are true or false.

1) Parametric statistics usually requires an assumption about normal distribution. True/ False

2) Non-parametric statistics requires fewer assumptions than parametric statistics. True/ False

3) The assumption of normal distribution is mandatory in non-parametric statistics. True/ False

4) When the data are on the ordinal scale, parametric statistics are most useful procedures. True/ False

5) Most of the statistical packages (like SPSS, R) do not have non-parametric statistics options. True/ False

3.4 SCALES OF MEASUREMENT

Measurement is at the heart of science. Measurement can be defined as a process of assigning a number to psychological or physical attribute by following certain rule. There are four different kinds of scales of measurement. They are nominal, ordinal, interval and ratio. We will know more about them since these scales are related to the use of correlations.

3.4.1 Nominal

Nominal is simplest kind of scale. When the number is assigned for the purpose of identification, these numbers are called as *nominal* numbers. Common examples are the numbers written on cricket players jerseys. They do not indicate that one is better or worse than other. They simply serve the purpose of identification. No mathematical procedure is possible with these numbers.

3.4.2 Ordinal

When the number assigned to an attribute have a property of order, then these numbers are called as *ordinal scale*. The common example is merit list numbers. The ranking of the batsmen or bowlers by ICC is also an ordinal scale. You will note that this scale has a property of order. Which means that first is better than everyone else, the second is inferior to first but superior than others, and so on. But this does not guarantee that the distances between them are equal.

3.4.3 Equal Interval Scale

Equal interval scale is the one that has an additional property of equal distances between any two consecutive units. Look at the example of Celsius scale of measurement of heat. Answer the two questions given below:

1) Is the difference between 25^0C and 50^0C equal to the difference between 75^0C and 100^0C?

2) Is 50^0C double the temperature at 25^0C?

Most of you would get the first answer right. The answer is "Yes". The difference is the difference of 25^0C (that is 50^0C – 25^0C = 100^0C – 75^0C = 25^0C). Now, most of you would be tempted to say yes to the second question as well. But that is a wrong answer. You would wonder why? The reason lies in the Celsius scale. The lowest possible temperature is not 0^0C. The real zero of the temperature is not 0^0C but it is much below at – 273.15^0C. So 25^0C temperature is from – 273.15^0C to 25^0C. Hence the double of 25^0C would be much higher than 50^0C. It is actually 323.15^0C.

The Celsius scale is a good example of interval scale. There are two properties in addition to the property of order:

1) The zero is not a real zero. It is a reference zero. The real zero has to denote absolute absence of the quantity. The reference zero does not indicate the absolute absence of quantity.

2) The distance between any two consecutive units is equal. That's the reason the answer of the first question is yes. The difference between 4^0C and 5^0C is similar to the difference between 321^0C and 322^0C that is of 1^0C.

This is the scale employed by various psychological measures. If the score on intelligence test or quality of life scale is zero, then it does not mean that the intelligence or quality of life is absolutely absent. It simply indicates lowest possible score on that particular test or scale.

3.4.4 Ratio

The ratio scale has all the properties of the equal interval scale. But the major advantage it has is that it has a real zero. The zero on this scale indicates the absolute absence of the quantity. It has a property of order, equal distances among the consecutive units, and real zero. That's the reason it is most superior version of scales. The examples are weight, length, etc. for example, zero gram wheat simply would mean absolute absence of wheat.

Self Assessment Questions

Match the pairs: Match the scales with their properties.

	Scale		**Properties of Scale**
1	Equal Interval	A	Numbers only indicate 'greater than' 'lesser than' relationships between units
2	Ordinal	B	Real zero
3	Ratio	C	Numbers are just for identification
4	Nominal	D	Reference zero, two consecutive units have equal spacing

3.5 CONDITIONS FOR RANK-ORDER CORRELATIONS

Two important conditions warrant the use of rank-order correlations. 1. The data are presented in rank-order, 2. The distributional assumptions underlying Pearson's Correlations are not met.

The scales of measurement have a special relevance for the correlations. The Pearson's product-moment correlation can be computed on interval or ratio scale. (It can also be computed on ordinal scale. Actually, Pearson's correlation computed on ordinal scale is known as Spearman's *rho*). When the data are on the ordinal scale, then Spearman's *rho* (r_s) is a better statistics. Kendall's *tau* (τ) can also be computed on the data that are presented in rank-orders. Considering these aspects of correlations, you would understand the importance of scales of measurement in the correlations.

If Pearson's correlation is employed as an inferential statistics, then it requires an assumption of bivariate normal distribution. It also makes and assumption of linearity. If the assumptions underlying the Pearson's correlation are not met, then the alternatives need to be employed. Under such circumstances, Spearman's *rho* and Kendall's *tau* are useful options.

3.6 RANKING OF THE DATA

The ranking of the data is the most important step in carrying out the rank order correlations. Ranking data is relatively an easy procedure. It involves few simple steps. Look at the scores given below. Five scores are given. You have to rank them from lowest to highest.

Scores: 12, 18, 7, 25, 15.

One has to start from the lowest score. The lowest score is 7, so it gets the rank 1, the next lowest is 12, so it gets the rank of 2, 15 gets 3, 18 has a rank of 4 and 25 has a rank of 5. Now all five scores are ranked. Look how the ranks look like.

Scores	12	18	7	25	15
Ranks	2	4	1	5	3

Now look at slightly the difficult problem:

Scores: 14, 8, 17, 6, 20, 14, 11, 10, 10, 14.

You will realise that 6 is the smallest score. So it gets the rank of 1; 8 next smallest, so it gets the rank of 2. Ten is next smallest, but it has appeared twice. This problem is called as problem of *ties*. When a score appear two or more times, then the ranks are called as *tied*. Then we have to compute the average of the ranks. These two numbers (10) would take the ranks 3 and 4. But since they are same numbers we cannot give rank of 3 to one and 4 to another. So we need to assign the mean rank to these to numbers. So we compute the mean rank . (3+4 = 7 /2 = 3.5.) Now we will give this rank 3.5. to both the 10 scores. In this process, we have exhausted rank 3 and 4. So now we have to give rank 5 to next number, that is, 11. Next number is 14. It has

appeared thrice. We have rank 6, 7, and 8 for these three numbers. So we compute the average of the ranks .(6+7+8 = 21 / 3= 7). Which appears three times. So far, we have used the ranks 6, 7 and 8. The next number is 17, which will get a rank of 9. The last number is 20 which will get a rank of 10. Look at the ranking of these scores:

Scores	14	8	17	6	20	14	11	10	10	14
Ranks	7	2	9	1	10	7	5	3.5	3.5	7

One is likely to make an error while ranking, especially, when there are many scores and *tied* ranks. You can verify the accuracy of your ranking by a simple cross-check. This can be used despite of *tied* ranks.

Sum of the ranks = 1+ 2+ 3+ 4+ 5 = 15.

Where, *n* = number of scores being ranked.

For the first example, we had 5 scores. So the sum of the ranks is:

The ranks we have assigned are 2, 4, 1, 5, 3. The sum of the assigned ranks is:

2 + 4 + 1 + 5 + 3 = 15

This confirms that the assignment of the ranks was correct. Look at the second example. We have 10 scores to be ranked. So the sum of ranks is: 55. (7+ 2+ 9+1+10+7+5+3.5+3.5+7 = 55)

Now compute the sum of the ranks assigned. The sum of the assigned rank is also 55. This means that we have correctly assigned the ranks.

Self Assessment Questions

Assign the ranks to the following scores. Also fine out the sum of the ranks to cross check the ranking.

Scores: 3, 5, 8, 2, 6.

Scores: 6, 3, 9, 1, 6, 2, 5, 8, 5, 1.

Scores: 14, 6, 2, 8, 11, 19, 12, 26, 4, 10, 6, 11.

Scores: 110, 127, 112, 109, 105, 112, 118, 114, 119, 99, 105, 121.

3.7 RANK CORRELATIONS

Rank correlations are the procedures that are based on the ordinal or rank-order data. These procedures are especially suitable under the conditions when the assumptions of the Pearson's product-moment correlations are not satisfied by the data. If the data has monotonic relationship (consistently increasing and never decreasing or consistently decreasing and never increasing), but the relations are not linear, then the use of these procedures is better suited than Pearson's correlation. The two popular procedures among the various procedures available are Spearman's *rho* (r_s) and Kendall's *tau*. In the next unit we shall obtain more information about them.

3.8 LET US SUM UP

In this unit, we have discussed various aspects of nonparametric statistical methods. We have understood the difference between the parametric and non-

parametric methods. We also have obtained more information about the parameter, statistics and hypothesis testing. Then we have learned about the scales of measurement. They constitute important basis for the rank-order correlations. We carried out the activity of ranking. All this information is useful in understanding the next unit as well as other non-parametric methods. Now we move to the next unit.

3.9 UNIT END QUESTIONS

1) How do we make parameter estimation? Give examples.
2) Define parameter, sample, statistics and hypothesis testing.
3) What are the important features of hypothesis testing?
4) What are the differences between parametric and nonparametric statistics? When do we use parametric statistics and when do we use non-parametric statistics?
5) What are the advantages and disadvantages of non-parametric statistics. Give suitable examples
6) What are the different scales of measurement? State their significant features and characteristics. How are they different from each other?
7) What are the necessary conditions to compute rank correlations?
8) How do you rank a data. Give an example. Suppose we have a data as given here: 16, 28, 35,15, 29, 50, 38, 45, 19, 24, rank them in the order it should be done.
9) What are rank correlations? Why are they called so? Give examples of rank correlations.

Answers to SAQ

Answers to 3.2. Answers: 1: False, 2: True, 3: True, 4: False, 5: False.

Answers to 3.3. 1: True; 2: True; 3: False; 4: False; 5: False.

Answers to 3.4: 1 = D; 2 = A; 3 = B; 4 = C.

Answers: to 3.6.

Scores:	3	5	8	2	6
Ranks	2	3	5	1	4

Sum of the ranks: 15

Scores:	6	3	9	1	6	2	5	8	5
Ranks	7.5	4	10	1.5	7.5	3	5.5	9	5.5

Sum of the ranks: 55

Scores	14	6	2	8	11	19	12	26	4	10	6	11
Ranks	10	3.5	1	5	7.5	11	9	12	2	6	3.5	7.5

Sum of the ranks: 78

Scores	110	127	112	109	105	112	118	114	119	99	105	121
Ranks	5	12	6.5	4	2.5	6.5	9	8	10	1	2.5	11

Sum of the ranks: 78

3.10 SUGGESTED READINGS

Aron, A., Aron, E. N., Coups, E.J. (2007). *Statistics for Psychology*. Delhi: Pearson Education.

Minium, E. W., King, B. M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education*. Singapore: John-Wiley.

Guilford, J. P., & Fructore, B. (1978). *Fundamental Statistics for Psychology and Education.* N.Y.: McGraw-Hill.

Wilcox, R. R. (1996). *Statistics for Social Sciences*. San Diego: Academic Press.

UNIT 4 RANK CORRELATION (*Rho* AND KENDALL RANK CORRELATION

Structure

4.0 INTRODUCTION

In this unit we shall learn two important techniques of correlation. One is **'Spearman's Rank-Order Correlations Coefficient'** or **Spearman's *rho* (r_s)** and the other one is **Kendall's *tau* (τ)**. Both the correlations are rank-order correlations. These correlations are preferred correlation techniques when the data is in rank-order or does not follow the assumptions of Pearson's product-moment correlation coefficient. While learning these coefficients, we shall learn the logic, the conditions under which they can be used, the null hypothesis, the formula for computing correlation, and interpreting the results.

4.1 OBJECTIVES

On successful completion of this unit, you are expected to be able to:

- Describe rank-order correlation and evaluate conditions under which they can be applied;
- Explain the logic of Spearman's *rho*;
- Carry out the computation of Spearman's *rho*;
- Apply the significance and the interpretation of *rho*;
- Explain the logic of Kendall's *tau*;
- Carry out the computation of Kendall's *tau*; and
- Apply the significance and the interpretation of Kendall's *tau*

4.2 RANK-ORDER CORRELATIONS

We have studied Pearson's correlation in the second unit. The Pearson's correlation is calculated on continuous variables. But when the data are in the form of ranks or when the assumptions of Pearson's correlation are not followed by the data, then it is not advisable to apply Pearson's correlation coefficient. Under such circumstances, rank-order correlations constitute one of the important options. We have already discussed the scales of measurement in the last unit. The ordinal scale data is called as rank-order data. Now let us look at these two aspects, rank-order and assumption of Pearson's correlations, in more detail.

4.2.1 Rank-Order Data

When the data is in rank-order, the correlations that need to be computed are called as rank-order correlation. Unlike the continuous data, the rank-order data present the ranks of the individuals or subjects. For example, scholastic achievement can be measured by actual marks obtained by individuals in the unit test or any other examination conducted. Suppose, the school displays just the merit list without the marks of the students, then we will only know who has obtained maximum marks, and so on. But we will not be able to know the marks. The data of marks constitute a continuous variable. Whereas, the merit list (ranks) provide us with the ordinal scale data. Look at the table 4.1.

Table 4.1: Table shows the marks and the ranks of 10 students in the class. The marks have provided us with the continuous data whereas merit list has provided us with rank-order data.

Table 4.1: Marks and ranks of 10 students

Student	Marks (Interval Data)	Merit List (Ordinal Data)
A	90	1
B	86	2
C	81	3
D	76	4
E	75	5
F	64	6
G	61	7
H	59	8
I	57	9
J	52	10

If the data available is ordinal or rank-order data, then rank-order correlation needs to be computed. If marks need to be correlated with other continuous variable like 'marks in previous standard' then Pearson's correlation is an appropriate option. If marks have to be correlated with other ordinal scale variable like ranking in the previous standard, then Pearson's correlation is not

an appropriate option. There rank-order correlation is required. Here we learn an important lesson: if we want to correlate X and Y, and either X or Y or both are on ordinal scale, then rank-order procedures need to be employed.

4.2.2 Assumptions Underlying Pearson's r are not Satisfied

The significance testing of Pearson's Product-moment correlation, as we learnt in the previous unit is based on some assumptions. We have learned these assumptions in the second unit. If these assumptions are not satisfied by the data, then application of statistical significance testing to the Pearson's coefficient becomes problematic. In such situation, one can think of applying the rank-order correlation coefficient.

At this moment, one also needs to note that Pearson's correlation is known as *robust* statistics. Robust means that the distributional properties are not seriously hampered even if there is minor deviation from the assumptions of the test.

Rank-order correlations are applicable especially when the relationship is monotonic but not linear. The *monotonic* relationship is the one when values in the data are consistently increasing and never decreasing or consistently decreasing and never increasing. It implies that as X increases Y consistently increases or as X increases Y consistently decreases. One such example of monotonic relationship is shown below. Figure 4.1 shows monotonic relationship (in reality it is a power function). Since it is not a linear relationship, we cannot calculate the Pearson's product-moment correlation. In such instance, rank-order is a better option (indeed, curve-fitting is a best approach). Figure 4.1: the figure shows a monotonic relationship between X and Y..

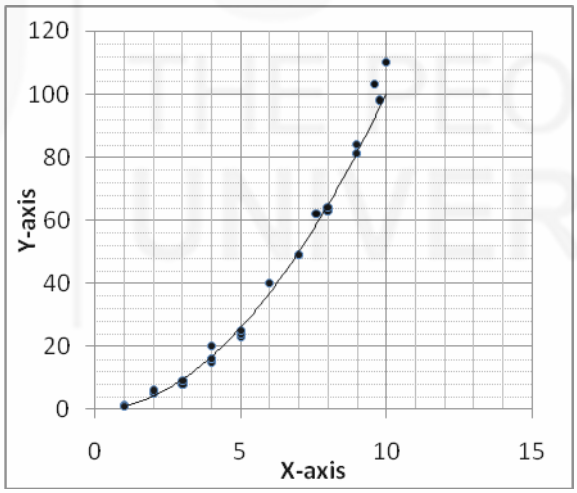


Fig. 4.1: A monotonic relationship between X and Y

4.3 SPEARMAN'S *rho* (r_s)

4.3.1 Introduction to Spearman's Rho

The Spearman's rank-order correlation or Spearman's *rho* (r_s) developed by a well-known psychologists, Charles Spearman (1904) is computed when the data is presented on two variables for *n* subjects. It can also be calculated for data of *n* subjects evaluated by two judges for inter-judge agreement. It is suitable for the rank-order data. If the data on X or Y or on both the variables are in rank-order then Spearman's *rho* is applicable. It can also be used with

continuous data when the assumptions of Pearson's assumptions are not satisfied. It is used to measure a monotonic relationship. Like Pearson's *r*, the range of Spearman's *rho* (r_s) is also from – 1.00 to + 1.00. Similar to Pearson's correlation, the interpretation of Spearman's rho is based on sign of the coefficient and the value of the coefficient. If the sign of r_s is positive the relationship is positive, if the sign of r_s is negative then the relationship is negative. If the value of r_s is close to zero then relationship is weak, and as the value of r_s approaches to ± 1.00, the strength of relationship increases. When the value of r_s is zero then there is no relationship between X and Y and if r_s is absolute value 1.00, then the relationship between X and Y is perfect. Whatever the value the r_s may take, it does not imply causation. Please refer to our discussion in the first unit for correlation and causality.

4.3.2 Null and Alternative Hypothesis

The Spearman's *rho* can be computed as a descriptive statistics. In that case we do not carry out hypothesis testing. If the r_s is computed to estimate population correlation, then null and alternative hypothesis are required.

The null hypothesis states that

$$H_O: \rho_s = 0$$

It implies that in the population represented by sample, the value of Spearman's correlation coefficient between X and Y is zero.

The alternative hypothesis states that

$$H_A: \rho_s \neq 0$$

It implies that in the population represented by sample, the value of Spearman's correlation coefficient between X and Y is not zero. This alternative hypothesis requires a two-tailed test. Depending on the theory, the other alternatives could be written. They are either 1. $H_A: \rho_s < 0$ or 2. $H_A: \rho_s > 0$. The first H_A denotes that the population value of Spearman's *rho* is smaller than zero. The second H_A denotes that the population value of Spearman's *rho* is greater than zero. Remember, only one of them has to be tested and **not** both. One-tailed test is required for this hypothesis.

4.3.3 Computation of Spearman's rho

The data on X and Y variables on are required to compute Spearman's *rho*. If the data are on continuous variables then it need to be converted into a rank-order. The computational formula is as follows:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2 - 1)} \qquad \text{(eq. 4.1)}$$

Where,

r_s = Spearman's rank-order correlation

D = difference between the pair of ranks of X and Y

n = the number of pairs of ranks

Let's start with an example. Suppose we are given the scores of 10 students in two subjects, mathematics and science. We want to find the correlation

between rank in mathematics and rank in science. The data are provided in table 4.2. The data are provided on marks obtained by students. This is continuous variable. We need rank-order data to carry out the Spearman's *rho*. So first, we need to transfer this data into ranks. The steps for computation of r_s are given below:

- **Step 1:** List the names/serial number of subjects (students, in this case) in column 1.
- **Step 2:** Write the scores of each subject on X variable (mathematics) in the column labeled as X (column 2), write the scores of each subject on Y variable (Science) in the column labeled as Y (column 3).
- **Step 4:** Rank the scores of X variable in ascending order. Give rank 1 to the lowest score, 2 to the next lowest score, and so on. This column is labeled as R_X (Column 4).
- **Step 5:** Rank the scores of Y variable in ascending order. Give rank 1 to the lowest score, 2 to the next lowest score, and so on. This column is labeled as R_Y (Column 5). Do cross-check your ranking by calculating the sum of ranks.
- **Step 6:** Now find out D, where $D = R_X - R_Y$ (Column 6).
- **Step7:** Square each value of D and enter it in the next column labeled as D^2 (Column7). Obtain the sum of the D^2. This is written at the end of the column D^2. This $\sum D^2$ is 18 for this example.
- **Step8:** Use the equation to compute the correlation between rank in mathematics and rank in science.

Table 4.2: Marks obtained in science and mathematics and computation of Spearman's rho.

Students	Mathematics (X)	Science (Y)	R_X	R_Y	$D = R_X - R_Y$	D^2
A	20	21	6	6	0	0
B	13	13	2	1	1	1
C	30	36	10	10	0	0
D	14	19	3	4	–1	1
E	16	20	4	5	–1	1
F	12	15	1	3	–2	4
G	25	23	7	7	0	0
H	29	27	9	8	1	1
I	17	14	5	2	3	9
J	26	29	8	9	–1	1
$n = 10$						$\sum D^2=18$

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

r = 6 × 18 / 10(100-1) = .891

We have computed the Spearman's *rho* for this example. The value of *rho* is 0.891. This value is positive value. It shows that the correlation between the ranks of mathematics and the ranks of science is positive. It indicates that the relationship between them is positively monotonic. The value of the correlation coefficient is very close to 1.00 which indicates that the strength of association between the two set of ranks is very high. In this example, I have deliberately avoided the use of tied ranks. I shall introduce you to the problem of tied ranks in the next section. You can do one interesting exercise.

Do Pearson's correlation on Rx and Ry. To your surprise, you will realise that the answer is identical to the one we have obtained. That is the relationship between Pearson's *r* and Spearman's *rho*. The Spearman's *rho* can be considered as a special case of Pearson's *r*.

4.3.4 Spearman's rho with Tied Ranks

The ranks are called as *tied ranks* when two or more subjects have same score on a variable. We usually get an answer that is larger than the actual value of Spearman's *rho* if we employ the formula in the equation 4.1 for the data with the tied ranks. Hence, the formula in equation 4.1 should not be used for data with tied ranks. We have to do a correction in the formula. The recommended procedure of correction for tied ranks is computationally tedious. So we shall use a computationally more efficient procedure. The easier procedure of correction actually uses Pearson's formula on the ranks. The formula and the steps are as follows:

$$r = r_s = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]\left[\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right]}} \quad \text{(eq. 4.2)}$$

Where,

r_s = Spearman's *rho*

X = ranks of variable X

Y = rank on variable Y

n = number of pairs

Look at the example we have solved for Pearson's correlation. It is an example of relationship between openness and creativity. We shall solve the same example with Spearman's *rho*.

Steps for r_s with tied ranks.

If the data are not in the ranks, then convert it into rank-order. In this example, we have assigned ranks to X and Y (column 2 and 3) in column 4, and 5.

Appropriately rank the ties (Cross-check the ranking by using sum of ranks check.). This is the basic information for the Spearman's *rho*.

Compute the square of rank of X and rank of Y for all the observations. It is in the column 6 and 7.

Multiply the rank of X by rank of Y for each observation. It is provided in column 8.

Obtain sum of all the columns.

Now all the basic data for the computation is available.

Enter this data into the formula shown in the equation 4.2 and calculate r_s.

Table 4.3: Spearman's *rho* for tied ranks

Subject	Openness (X)	Creativity (Y)	Rank X	Rank Y	(Rank X)^2	(Rank Y)^2	(Rank X) (Rank Y) XY
1	10	26	3	5.5	9	30.25	16.5
2	8	23	1	1.5	1	2.25	1.5
3	9	23	2	1.5	4	2.25	3
4	13	26	7	5.5	49	30.25	38.5
5	11	24	4	3	16	9	12
6	14	30	8	10	64	100	80
7	16	27	10	7.5	100	56.25	75
8	12	27	5.5	7.5	30.25	56.25	41.25
9	15	29	9	9	81	81	81
10	12	25	5.5	4	30.25	16	22
Sum			55	55	384.5	383.5	370.75
				Ans =	Rho = 0.84		

$$r = r_s = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]\left[\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right]}}$$

$$[370.75 - \{(55) \times (55) / 10\}] / \sqrt{[\ 384.5 - \{(55)^2/10\}]\ [383.5 - \{(55)^2/10\}]}$$

$$=68.55/81.44=0.84$$

The Spearman's *rho* for this example is 0.837. Since this is a positive value, the relationship between them is also positive. This value is rather near to 1.00. So the strength of association between the ranks of openness and ranks of creativity are very high.

4.3.5 Significance Testing of Spearman's rho

Now we will learn to test the significance of Spearman's *rho*. The null hypothesis tested is

$$H_O: \rho_s = 0$$

It states that in the population represented by sample, the value Spearman's *rho* between X and Y is zero.

The alternative hypothesis is

$$H_A: \rho_s \neq 0$$

It states that in the population represented by sample, the value Spearman's *rho* between X and Y is not zero. This alternative hypothesis requires a two-tailed test.

We need to refer to Appendix D for significance testing. The Appendix D provides critical values of Spearman's *rho* from $n = 4$ onwards. The appendix provides critical values for one-tailed as well as two-tailed tests. Let's use this table for the purpose of hypothesis testing for the first example of correlation between ranks in mathematics and ranks in science (table 4.2).

The obtained Spearman's *rho* is 0.891 on the sample of 10 individuals. For $n = 10$, and two-tailed level of significance of 0.05, the critical value of $r_s = 0.648$. The critical value of $r_s = 0.794$ at the two-tailed significance level of 0.01. The obtained value of 0.891 is larger than the critical value at 0.01. So the obtained correlation is significant at 0.01 level (two-tailed). We reject the null hypothesis and accept the alternative hypothesis. It indicates that the value of the Spearman's *rho* is not zero in the population represented by the sample.

For the second example (table 4.3), the obtained r_s value is 0.84 on the sample of 10 individuals. For $n = 10$, the critical value is 0.794 at the two-tailed significance level of 0.01. The obtained value of 0.84 is larger than the critical value at 0.01. So the obtained correlation is significant at 0.01 level (two-tailed). Hence, we reject the null hypothesis and accept the alternative hypothesis.

When the sample size is greater than ten, then the *t*-distribution can be used for computing the significance with $df = n - 2$. Following equation is used for this purpose.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \qquad \text{(eq. 4.3)}$$

For the example shown in table 4.2, the *t*-value is computed using equation 4.3.

At the $df = 10 - 2 = 8$, the critical *t*-value at 0.01 (two-tailed) is 3.355. The obtained *t*-value is larger than the critical *t*-value. Hence, we reject the null hypothesis and accept the alternative hypothesis.

Self Assessment Questions

1) Calculate Spearman Rho for the following data

Students	Marks in History	Marks in English
1.	45	52
2.	50	42
3.	55	45
4	35	46
5	62	73

Answer: Rho = 0.20

2) What do you understand by the terms null and alternative hypothesis? Suppose you want to find out if there is a relationship between intelligence and socio economic status, what would be your null hypothesis and what will be your alternative hypothesis?

4.4 KENDALL'S *tau* (τ)

4.4.1 Introduction

Kendall's *tau* is one of the measures of correlation. This correlation procedure was developed by Kendall (1938). Kendall's *tau* is based on an analysis of two sets of ranks, X and Y. It is as an alternative to Spearman's *rho* (r_s). Kendall's *tau* is symbolised as τ, which is a lowercase Greek letter *tau*. The parameter (population value) is symbolised as τ and the statistics computed on the sample is symbolised as r_s, Like Spearman's *rho*, the range of *tau* is from – 1.00 to + 1.00. Though there are some similarities in the properties of *tau* and r_s, the logic employed by *tau* is entirely different than that of *rho*. The interpretation is based on the sign and the value. Higher value indicates stronger relationship. Positive value indicates positive relationship and negative value indicates negative relationship.

4.4.2 Null and Alternative Hypothesis

The Kendall's *tau* can be computed as a descriptive statistics. In that case we do not carry out hypothesis testing. If r is computed to estimate population correlation τ, then null and alternative hypothesis are required.

The null hypothesis states that

$$H_O: \tau = 0$$

It stated that in the population represented by sample, the value Kendall's *tau* between X and Y is zero.

The alternative hypothesis states that

$$H_A: \tau \neq 0$$

It states that in the population represented by sample, the value Kendall's *tau* between X and Y is not zero. This alternative hypothesis requires a two-tailed test.

Depending on the theory, the other alternatives could be written. They are either

1. $H_A: \tau < 0$ or
2. $H_A: \tau > 0$.

The first H_A denotes that the population value of Kendall's *tau* is smaller than zero. The second H_A denotes that the population value of Kendall's *tau* is greater than zero. Remember, only one of them has to be tested and **not** both. One-tailed test is required for these hypotheses.

4.4.3 Logic of Computation of τ

The *tau* is based on concordance and discordance among two sets of ranks. For example, table 4.4 shows ranks of four subjects on variables X and Y as R_X and R_Y. In order to obtain concordant and discordant pairs, we need to order one of the variables according to the ranks, from lowest to highest (we have ordered X in this fashion). Take a pair of ranks for two subjects A (1,1)

and B (2,3) on X and Y. Now, if sign or the direction of $R_X - R_X$ for subject A and B is similar to the sign or direction of $R_Y - R_Y$ for subject A and B, then the pair of ranks is said to concordant (i.e., in agreement). In case of subject A and B, the $R_X - R_X$ is (1 – 2 = – 1) and $R_Y - R_Y$ is also (1 – 3 = – 2). The sign or direction of A and B pair is in agreement. So pair A and B is called as concordant pair. Look at second example of B and C pair. The $R_X - R_X$ is (2 – 3 = – 1) and $R_Y - R_Y$ is also (3 – 2 = +1). The sign or the direction of B and C pair is not in agreement. This pair is called as discordant pair.

Table 4.4: Four subjects and their ranks

Subjects	Rank X (Rx)	Rank Y (Ry)
A	1	1
B	4	2
C	3	4
D	2	3

How many such pair we need to evaluate? They will be $n\,(n-1)/2 = (4 \times 3)/2 = 6$, so six pairs. Here is an illustration: AB, AC, AD, BC, BD, and CD. Once we know the concordant and discordant pairs, then we can calculate by using following equation.

$$\tilde{\tau} = \frac{n_C - n_D}{\left[\frac{n(n-1)}{2}\right]}$$

Where,

τ = value of τ obtained on sample

n_C = number of concordant pairs

n_D = number of discordant pairs

n = number of subjects

I illustrate a method to obtain the number of concordant (n_C) and discordant (n_D) pairs for this small data in table 4.5. We shall also learn a computationally easy method later.

- Step 1. Ranks of X are placed in second row in the ascending order.
- Step 2. Accordingly ranks of Y are arranged in third row.
- Step 3. Then the ranks of Y are entered in the diagonal (see table 4.5).
- Step 4. Start with first element in the diagonal which is 1 (row 4).
- Step 5. Now move across the row. Compare it (1) with each column element of Y. If it is smaller then enter C in the intersection. If it is larger, then enter D in the intersection. For example, 1 is smaller than 3 (column 3) so C is entered. In the next row (row 5), 3 is in the diagonal which is greater than 2 (column 4) of Y, so D is entered in the intersection.
- Step 6. Then ∑C and ∑D are computed for each row.
- Step 7. The n_C is obtained from ∑∑C (i.e., 5) and n_D is obtained from ∑∑D (i.e., 1). These values are entered in the equation 4.4 to obtain .

Table 4.5: Table showing computation of concordant and discordant pairs

Subjects	A	B	C	D	∑C	∑D
Rank of X	1	2	3	4		
Rank of Y	1	3	2	4		
	1	C	C	C	3	0
		3	D	C	1	1
			2	C	1	0
				4	0	0
					∑∑C = 5	∑∑D = 1

4.4.4 Computational Alternative for τ

You will realise that the earlier procedure is tedious. I suggest an easier alternative. Suppose, we want to correlate time taken for solving a test and the performance on a test. Unfortunately, we do not have the refined data of time and performance. We know the order in which students have returned their answer booklets as they have finished. This gives us ordinal data on time taken to solve the test. We also know the ranks of the students on that test. The data are given below for 10 subjects.

Table 4.6. Data of 10 subjects on X (rank for time taken to complete test) and Y (ranks of performance)

Table 4.6: Time taken to complete test and performance on test

Subjects being ranked	Time taken to complete test(X)	Performance on test (Y)
A	1	1
B	2	3
C	3	5
D	4	2
E	5	4
F	6	6
G	7	8
H	8	7
I	9	10
J	10	9

First we arrange the ranks of the students in ascending order (in increasing order; begin from 1 for lowest score) according to one variable, X in this case. Then we arrange the ranks of Y as per the ranks of X. I have drawn the lines to connect the comparable ranking of X with Y. Please note that lines are not drawn if the subject gets the same rank on both the variables. Now we calculate number of inversions. Number of inversions is number of intersection of the lines. We have five intersections of the lines.

So the following equation can be used to compute (equation 4.5.)

$$\tilde{\tau} = 1 - \frac{2(n_s)}{\frac{n(n-1)}{2}}$$

Where

τ = sample value of τ

n_s = number of inversions.

n = number of subjects

$$\tau = 1 - [2\ (5)]\ /\ [10\ (10\text{-}1)\ /\ 2]$$
$$= 1\text{-}\ \ 10 / 45\ \ =\ 0.22$$
$$= .1 - 22\ \ = .78$$

The value of Kendall's *tau* for this data is 0.78. The value is positive and near 1.00. So the relationship between X and Y is positive. This means as the rank on time taken increases, the rank on subject also increases. Interpretation of *tau* is also clear. If the *tau* is 0.78, then it can be interpreted as follows: if the pair of subjects is sampled at random, then the probability that their order on two variables (X and Y) is similar, that is 0.78 higher than the probability that it would be in reverse order. The calculation of *tau* need to be modified for tied ranks. Those modifications are not discussed here.

4.4.5 Significance Testing of τ

The statistical significance testing of Kendall's *tau* is carried out by using either Appendix E and referring to the critical value provided in the Appendix E. The other way is to use the z transformation. The z can be calculated by using following equation

$$\tilde{\tau} = 1 - \frac{2(n_s)}{\frac{n(n-1)}{2}} = 1 - \frac{2(5)}{\frac{10(10-1)}{2}} = 1 - \frac{10}{45} = 1 - 0.222 = 0.778$$

You will realise that the denominator is the standard error of *tau*. Once the Z is calculated, you can refer to Appendix A in any statistics book for finding out the probability.

For our example in table 4.4, the value of = 0.664 for n = 4. The Appendix E provides the critical value of 1.00 at two-tailed significance level of 0.05. The obtained value is smaller than the critical value. So it is insignificant. Hence, we retain the null hypothesis which states H_O: $\tau = 0$. So we accept this hypothesis. It implies that the underlying population represented by the sample has no relationship between X and Y.

For example in table 4.6, the obtained value of *tau* is 0.778 with the n = 10. From the Appendix E, for the n = 10, the critical value of *tau* is 0.644 at two-tailed 0.01 level of significance. The value obtained is 0.778 which is higher than the critical value of 0.664. So the obtained value of *tau* is significant at 0.01 level. Hence, we reject the null hypothesis H_O: $\tau = 0$ and accept the alternative hypothesis H_A: $\tau \neq 0$. It implies that the value of *tau* in the

population represented by sample is other than zero. We can also apply z equation for this data.

The z table (normal distribution table) in the Appendix A in any statics book has a value of z = 1.96 at 0.05 level and 2.58 at 0.01 level. The obtained value of z = 3.313 is far greater than these values. So we reject the null hypothesis.

Kendall's *tau* is said to be a better alternative to Spearman's *rho* under the conditions of tied ranks. The *tau* is also supposed to do better than Pearson's *r* under the conditions of extreme non-normality. This holds true only under the conditions of very extreme cases. Otherwise, Pearson's *r* is still a coefficient of choice.

4.5 LET US SUM UP

We studied the conditions under which the Spearman's *rho* and Kendall's *tau* are used. Then we have studied the logic and computation of Spearman's rho. We have learned to interpret rho. Similarly, we have studied the logic and computation of Kendall's *tau*. We have learned to interpret *tau*. This would enable you to use these techniques for the data presented to you.

4.6 UNIT END QUESTIONS

Problems:

A researcher was interested in correlating ranks in languages and ranks in mathematics. She collected the data in 12 students. Their ranks are given below. Do Spearman's rho and Kendall's tau on that data. Write null and alternative hypothesis, calculate statistics, test the significance, and interpret.

Students

	A	B	C	D	E	F	G	H	I	J	K	L
Rank in Language	3	2	5	6	7	9	10	8	1	4	12	11
Rank in Science	11	8	6	9	12	1	2	5	10	7	3	4

Spearman's rho = – 0.72; $p < 0.05$

Kendall's tau = – 0.515; $p < 0.05$

4.7 SUGGESTED READINGS

Aron, A., Aron, E. N., Coups, E.J. (2007). *Statistics for Psychology*. Delhi: Pearson Education.

Minium, E. W., King, B. M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education*. Singapore: John-Wiley.

Guilford, J. P., & Fructore, B. (1978). *Fundamental Statistics for Psychology and Education*. N.Y.: McGraw-Hill.

Wilcox, R. R. (1996). *Statistics for Social Sciences*. San Diego: Academic Press.

UNIT 1 SIGNIFICANCE OF THE DIFFERENCE OF FREQUENCY: CHI-SQUARE

Structure

1.0 INTRODUCTION

In psychology some times the researcher is interested in such questions as to whether one type of soft drink preferred by the consumer is in any way influenced by the age of the consumer , and whether ? Similarly among the four colours, red, green, indigo, yellow, is there any significant difference in the subjects in regard to the preference of colours, etc. To approach questions like these, we find the number of persons who preferred particular colour. Here we have the data in the form of frequencies, the number of cases falling into each of the four categories of colours. Here the researcher compares the observed frequencies characterizing the choices of the 4 categories by a large number of individuals. Some prefer red, some white, some green and some yellow etc. Hypothetically speaking one may say that all colours are equally preferred and are equal choice for the individuals. But in reality there may be more number of persons choosing red or blue than yellow and similarly quite a few may choose white etc. Thus what is expected as choice which is equal for all colours may be seen differently when actual choice is made by the persons. Now we note the number of persons who had chosen red, blue, yellow and white respectively and note it in the table. Hypothetically we said that there will be the same number of persons choosing each of the 4 colours. This also we note in the table. Now we compare and see for each colour what is the difference between the actual number of choices made as compared to the hypothetical equal number. The differences thereof are noted. To find out

if these differences are really of statistical significance amongst the persons in regard to choice of colours, we use a statistical technique called Chi-square. In this unit we will discuss about the concept of chi-square and its application.

1.1 OBJECTIVES

After reading this unit, you will be able to:

- Define non-parametric statistics;
- Apply the non-parametric statistics at appropriate context;
- Give the Meaning of chi-square;
- Give the Assumptions for the Application of chi-square;
- Explain chi-square distribution;
- Analyse the Properties of chi-square distribution; and
- Apply chi-square test.

1.2 PARAMETRIC AND NON-PARAMETRIC STATISTICS TEST

A variety of statistical tests are available for analyzing data. Broadly statistical tests can be classified into two categories one is parametric statistical test and other is non-parametric statistical test. Which statistical method we should select for analysing our data depends on

- The scale of measurement of data
- Dependence / independence of the measurement
- Number of populations being studied
- Shape of the population distribution.

There are four scales of measurement, nominal, ordinal, interval and ratio.

In *nominal scale* numbering or classification is always made according to similarity or differences observed with respect to some characteristic. For example if we take eye colour as a variable then we will be able to classify individuals in categories like blue eyed, brown eyed, black eyed and assign number within each category to identify the individuals belonging to a group.

In *ordinal scale*, the number reflects their rank order within their group with respect to some quality, property or performance. For example on the basis of marks obtained we give the ranks 1, 2, 3 and so on and students are ordered on the basis of marks.

The defect in such scales lies in the fact that the units along the scale are unequal in size. The difference in the marks between the first and second rank holder may not be the same, as the difference in marks in second and third position holders.

The *interval scale* does not merely identify or classify individuals in relation to some attribute by a number (on the basis of sameness or difference) or rank (in order of their merit position) but advances much ahead by pointing out the relative quantitative as well as qualitative differences. The major strength of interval scales is that they have equal units of measurement. They however do not posses a true zero.

The measures of *ratio scale* are not only expressed in equal limits but are also taken from a true zero.

Independence of the observation means the inclusion or exclusion of any case in the sample should not unduly affect the results of the study.

Similarly there may be different size of sample of the study. In psychology if the number of subjects in a group is 30 or more than thirty it is known as large sample.

The distribution of the population from which the samples have been drawn may be normal or skewed.

If the level of measurement on the collected data is in the form of an interval scale or ratio scale, or if the sample is very large on which the data has been collected, then it may be assumed that the population is normally distributed. If the observations are independent and the population have the same variance as the sample which has been drawn from it, then we use the Parametric Statistics.

However is many situations, the size of the sample is quite small, the assumption like normality of the distribution of scores in the population are doubtful, and the measurement of the data is available in the form of classification, or in the form of ranks then we use the non-parametric tests.

There are number of non-parametric tests. Such as, chi-square test, Wilcoxon-Mann-Whitney U test, Rank difference methods, (rho and tau) coefficient of concordance (w), Median test, Kruskal-Wallis H test, Friedman test.

In this unit we will discuss the chi-square test.

1.2.1 Chi-Square Test: Definitions

This is one of the most important non-parametric statistics. As we see further that chi-square is used for several purposes.

The term 'chi' (is the Greek letter, it is pronounced ki). The chi-square test was originally developed by Kart-Peason in 1900 and is sometimes called the Peason Chi-square. According to Garecte (1981) "The difference between observed and expected frequencies are squared and divided by the expected number in each case and the sum of these quotients is chi-square".

According to Guilford (1973) "By definition a x^2 is the sum of ratio (any number can be summed), each ratio is that between a squared discrepancy of difference and an expected frequency".

On the basis of above definitions it can be said that the discrepancy between observed and expected frequencies is expressed in term of a statistics named chi-square (x^2).

1.3 ASSUMPTIONS FOR THE APPLICATION OF X^2 TEST

Before using chi-square as a test statistic to test a hypothesis, the following assumptions are necessary:

- Data should be in the form of frequencies, although the chi-square test is conducted in terms of frequencies or data that can be readily transformed into frequency, it is best viewed conceptually as a test about proportions.
- The chi-square test is applied only to discrete data. However, any continuous data can be reduced to the categories is such a way that they can be treated as discrete data and then the application of chi-square is justified.
- The x^2 is applied when their is quantitative data.
- The observation should be independent. For example let us say that we wanted to find out the color preference of 100 females. We select 100 subjects randomly, the preference of one individual is not predictable from that of any other. If there is a relationship between two variables or the subjects are matched, then chi-square cannot be used to test. Correlated data like matched pair are not subjected to chi-square treatment.
- The sample drawn from the population about which inference is to be made should be random.
- If there are only two cells, the expected frequencies in each cell should be 5 or more. Because for observation less than 5, the value of x^2 shall be over estimated, resulting in the rejection of the null hypothesis.
- For more than two cells, no more than 20% of the expected values may be lesser than 5 and no expected value may be lesser than 1. If more than 20 percent of the cells have expected frequencies less than 5 then x^2 should not be applied.
- A sample with a sufficiently large size is assumed. If sample size is small then x^2 will yield an inaccurate inference. In general, larger the sample size, the less affected by chance is the observed distribution, and thus the more reliable the test of the hypothesis.
- The data should be expressed in original units, rather than in percentage or ratio form. Such precaution helps in comparison of attributes of interest.

1.3.1 x^2 Distribution

The term non-parametric does not mean that the population distribution under study has no parameters. All populations have certain parameters which define their distribution. The sampling distribution of x^2 is called x^2 distribution. Other hypothesis testing procedures, the calculated value of x^2 test statistic is compared with its critical (or table) value to know whether the null hypothesis is true. The decision of accepting a null hypothesis is based on how 'close' the sample results are to the expected results. If the null hypothesis is true, the observed frequencies will vary from their corresponding expected frequencies according to the influence of random sampling variation. The calculated value of x^2 will be smaller when agreement between observed frequencies and expected frequencies are smaller. The shape of a particular chi-square distribution depends on the number of degrees of freedom.

Let us now see what is degrees of freedom?

Let us say there are 5 scores, as for example, 25,35,45,55,65

The mean here is 45. Let us say that from this mean of 45 we try to see the difference in each of the other scores. While all the other 4 scores are taken for the difference, this 45 is not available for other calculations, as from this

Mena=45 , all other calculations are made. From this Mean we note the differences in the remaining scores, and thus amongst the 5 scores one score is not movable or in other words the group of 5 subjects has lost one degree of freedom. Suppose we go for higher level calculations, then accordingly there will be more number of items that cannot be moved and thus we may have greater degrees of freedom. In a 3 × 3 table, we will be using one cell from the row and one cell from the column for the purpose of chi-square calculations. Thus the degrees of freedom will be (3 – 1) (3 – 1) = 4. That is (r- 1) (k-1) (r = row and k = column)

It may also be noted that chi-square assumes positive values only. As such a chi-square curve starting from the original (zero point) lies entirely to the right of the vertical axis.

If we draw a curve with degrees of freedom on the X-axis and chi-square values on the Y axis, the he shape of a chi-square distribution curve will be skewed for very small degree of freedom. As the degrees of freedom increase, the shape also changes. Eventually, for larger degrees of freedom, the curve looks similar to the curve of a normal distribution. A point worth nothing is that the total area under a chi-square distribution curve is 1.0 as is the case in all other continuous distribution curves.

Self Assessment Questions

1) Given below are statements. Indicate in each case whether the statement is true or false?

   i) The utility of chi-square test depends largely on the quality of data used in the test. (T/F)

   ii) The number of degree of freedom in a chi-square test depends on both the number of row and the number of column in contingency table. (T/F)

   iii) The shape of the chi-square distribution depends on the degree of freedom. (T/F)

   iv) The value of chi-square depends on the size of the difference between observed frequency and expected frequency. (T/F)

   v) Chi-square is a parametric test. (T/F)

2) Fill in the blanks :

   i) A chi-square value can never be .....

   ii) When frequencies in any cell in less than 5 we use .....

   iii) If the obtained value of chi-square is more than the critical values given in table we ..... null hypothesis.

1.4 PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The following properties of chi-square test statistic are to be kept in mind when we analyse its sampling distribution.

- Chi-square is non-negative in value; it positively valued, because all discrepancies are squared, both positive and negative discrepancies make a positive contribution to the value of x^2.
- Chi-square will be zero only in the unusual event that each observed frequency exactly equals the corresponding expected frequency.
- It is not symmetrical; it is skewed to the right as mentioned earlier. The reason for this is because all the values are positive and hence so instead of having a symmetrical distribution, we have a distribution all on the positive side.
- Since density function of x^2 does not contain any parameter of population, Chi-square test statistic is referred to as a non-parametric test. Thus chi-square distribution does not depend upon the form of the parent population.
- There are many chi-square distributions. As with the t-distribution, there is a different chi-square distribution for the value of each degree of freedom.

It we know the degrees of freedom and the area in the right tail of a chi-square distribution, we can find the value of chi-square from the Table of chi-square. We give below two examples to show how the value of chi-square can be obtained from this table.

Example 1: Find the value of chi-square for 10 degree of freedom and an area of .05 in the right tail of the chi-square distribution curve.

Solution: In order to find the required value of chi-square, we first locate 10 in the column for degree of freedom (df.) and .05 in the top row in table of chi-square. The required chi-square value is given by the entry at the intersection of the row for 10 and the column for .05. This value is 18.307.

Example 2: Find the value of chi-square for 20 degree of freedom and an area of .10 in the left tail of the chi-square distribution curve.

Solution: This example is different from the above one. Here, the area is in the left tail of the chi-square distribution curve is given. In such a case, we have to first find the area in the right tail. This is obtained as follows:

Area in the right tail= 1 - area in the left tail

Area in the right tail = 1 - .10 = .90

Now we can find, in the same manner as used earlier, the value of x^2. We locate 20 in the column for df and .90 on top row in the table of chi-square. The required value of x^2 is 12.443.

1.5 APPLICATION OF CHI-SQUARE TEST

Chi-square is used for categorical data that is for data comprising unordered quantitative categories such as colours, political affiliation etc. The important application of x^2 test are as follows:

- Test of goodness of fit
- Test of independence
- Test for homogeneity

1.5.1 Test of Goodness of Fit

On several occasion a decision - maker needs to understand whether an actual sample distribution matches with a known theoretical distribution such as binomial, poisson, normal and so on. Similarly some time researcher compares the observed (sample) frequencies characterising the several categories of the distribution with those frequencies expected according to his or her hypothesis.

The x^2 test for goodness *of fit is a statistical test of how well given data support an assumption about the distribution of a population. That is, whether it supports a random variable of interest. In other words, the test determines how well an assumed distribution fits the given data.*

To apply this test, a particular theoretical distribution is first hypothesised for given population and then the test is carried out to determine whether or not the sample data could have come from the population of interest with the hypothesised theoretical distribution. The observed frequencies come from the observation of sample and the expected frequencies come from the hypothesised theoretical distribution. The goodness of fit test focusses on the difference between the observed frequencies and expected frequencies.

This can be explained with the help of following examples.

In a particular study, a researcher wants to study that among four brands of glucose biscuits, is there a difference in the proportion of consumers who prefer the taste of each. We formulate the null hypothesis that there is no differential preference among consumers for the four brand of glucose biscuits. Here we test the hypothesis that there are equal probability of individuals preferring each brand i.e. 1/4 = .25. Or if we take 100 persons to test this hypothesis, 25 each will prefer theoretically the 4 brands of biscuits.

The alternative hypothesis is that a preference exists for the first brand of the biscuits and the reminder are equally less preferred, or that the first two are preferred more than the second two and so on.

The use of chi-square tells us whether the relative frequencies observed in the several categories of our sample frequency distribution are in accordance with the set of frequencies hypothesised.

To test the null hypothesis in the above study, we might allow subjects to taste each of the four brand of glucose biscuits and then find their preference. We control possible extraneous influence, such as knowledge of brands name and order of presentation. Suppose we had as mentioned earlier randomly selected 100 subjects and that our observed frequencies of preference are as given in the table below:

	Brand A	**Brand B**	**Brand C**	**Brand D**
Observed Frequencies	20	18	30	32

We calculate the expected frequency for each category by multiplying the proportion hypothesised to characterise that category in the population by the sample size. According to null hypothesis, the expected proportionate preference for each biscuit is 1/4 and the frequencies of preference for each is therefore 25 (1/4) (100/4) = 25.

In any experiment, we anticipate that the observed frequency of choice will vary from the expected frequencies. We do not use these differences directly. One reason is that some differences are positive and some are negative. Thus they would cancel each other out. To get around this, we square each difference. But how much variation to expect ? Some measures of discrepancy is required to test the null hypothesis, this can be tested by chi-square. We calculate the chi-square (the method of computation of chi-square will be explained in the next unit). If the obtained chi-square value is more than the critical values at .05 or .01 level than we reject the null hypothesis and if obtained value of chi-square is less than the given values then we retain the null hypothesis.

The obtained value of chi-square depends on the size of the discrepancy between observed frequency and expected frequencies. Larger the differences between observed frequencies and expected frequencies the larger will be the chi-square.

The size of the discrepancy relative to the magnitude of the expected frequency contribute in the determination of the value of chi-square. For example, if we toss a number of coins and find out how many times the 'heads' or 'tails' have appeared. To take a numerical example, when we toss coins in 12 times and 1000 times. Then according to null hypothesis our expected frequency will be 6, and 500 respectively. If we obtain 11 heads in 12 tosses the discrepancy is 5. However, if we obtain 505 heads in 1000 tosses, the discrepancy is also 5.

The value of chi-square depends on the number of discrepancies involved in its calculation. For example if we use three brands of biscuits not the four, there would be less discrepancy to contribute to the total of chi-square. This will influence the degrees of freedom, that is, when the number of discrepancies are 4 the degrees of freedom will be 3 and when number of discrepancies are 3 the degree of freedom will be 2. When degrees of freedom is 4 and obtained chi-square is more than 9.48 then we reject the null hypothesis at .05 level. On the other hand when the degrees of freedom is 4 and obtained chi-square is 7.81 then we retain the null hypothesis at .05 level.

1.5.2 Test of Independence

In the above discussion so far, we have considered the application of chi-square only to 'one' variable case. We have considered testing whether the categories in an observed frequency distribution differ significantly from one another. The chi-square test has a much broader use in social science research; to test whether one observed frequency distribution significantly differ from another observed frequency. In other words, it can be said that chi-square is used for the analysis of bivariate frequency.

For example, social researcher is interested in surveying the attitudes of high school students concerning the importance of getting a college degree. She questions a sample of 60 senior high school students about whether they

believe that college education is becoming more important, less important or staying the same and whether male students respond differently from female students ?

As we know that nominal and ordinal variables are generally presented in the form of a cross-tabulation. Specifically cross tabulation are used to compare the distribution of one variable often called dependent variable, across categories of some other variable the independent variable. In a cross tabulation the focus is on the difference between group — such as between males and females — in terms of the dependent variable — for example, opinion about the changing value of a college education. In the above example we want to study whether there exists gender differences in belief regarding the importance of a college degree - are statistically significant.

To study this question, we classify the data (observed frequencies) in a bivariable distribution. For each variable, the categories are mutually exclusive. The data is classified in the following table:

	More Important	**Less Important**	**About the Same**
Male	25	6	8
Female	10	4	7

Bivariate frequency distribution of two types are known as contingency tables. From such a table we may inquire what cell frequencies would be expected if the two are independent of each other in the population. The chi-square test may be used to compare the observed cell frequencies with those expected under the null hypothesis of independence. If the (fo-fe) discrepancy are small, chi-square will be small suggesting that the two variable could be not different. Conversely, a large chi-square points toward a contingency relationship.

As in the case of the t ratio there is a sampling distribution for chi-square that can be used to estimate the probability of obtaining a significant chi-square value by chance alone rather than by actual population difference. The test is used to study the significance of difference between mean. On the other hand, chi-square is used to make comparison between frequencies rather than mean scores. As a result the null hypothesis for the chi-square test states that the populations do not differ with respect to frequencies of occurrence of a given characteristic. In general, the null hypothesis of independence for a contingency table is equivalent to hypothesising that in the population the relative frequencies for any row (across the categories of the column variable) are the same for all rows, or that in the population the relative frequencies for any column (across the categories of the two variables) are the same for all columns.

If the null hypothesis of independence is true at the population level, we should expect that random sampling will produce obtained value of chi-square that are in accord with the tabulated distribution of that statistic. If the hypothesis is false in any way, the obtained value of chi-square will tend to be larger than when the alternate hypothesis, stating that there will be a significant difference between the variables, is true.

In the case of a 2 × 2 contingency table, we can test for a difference between two frequencies or proportions from independent samples (just as with the 1 × 2 table, where we can test a hypothesis about a single proportion.

1.5.3 Test of Homogeneity

The test of homogeneity is useful in a case when we are interested in verifying, whether several populations are homogeneous with respect to some characteristic of interest. For example a movie producer is bringing out a new movie. In order to develop an advertising strategy he wants to determine whether the movie will appeal most to a particular age group or whether it will appeal equally to all age groups. We formulate the null hypothesis that the opinion of all age groups is same about the new movie. Hence, the test of homogeneity is useful in testing a null hypothesis that several populations are homogeneous with respect to a character. This test is different from the test of independence on account of the following reasons:

i) Instead of knowing whether two attributes are independent or not, we may like to know whether different samples come from the same population.

ii) Instead of taking only one sample for this test two or more independent samples are drawn from each population.

iii) To apply this test first a random sample is drawn from each population, and then in each sample the proportion falling in each category is determined. The sample data so obtained is arranged in contingency table. The procedure for testing of hypothesis is same as for test of independence.

Self Assessment Questions

1) What are the properties of Chi-square distribution? Give examples

.....

.....

.....

.....

.....

2) What are the applications of chi-square test ?

.....

.....

.....

.....

.....

3) Describe the Goodness of fit test? Why do we say that chi-square is a goodness of fit test?

.....

.....

.....

.....

4) What is meant by test of independence? Give suitable examples

.....

.....

.....

.....

.....

.....

5) Why is chi-square test vcalled the test of homogeneity? Explain with examples.

.....

.....

.....

.....

.....

.....

1.6 PRECAUTIONS ABOUT USING THE CHI-SQUARE TEST

In order to use a chi-square test properly, one has to be extremely careful and keep in mind certain precautions:

i) A sample size should be large enough. If the expected frequencies are too small, the value of chi-square gets over estimated. To overcome this problem we must ensure that the observed frequency in any cell of the contingency table should not be less than 5.

ii) When the calculated value of chi-square turns out to be more than the critical or theoretical value at a predetermined level of significance. We reject the null hypothesis. In contrast, when the chi-square value is less than the critical theoretical value, the null hypothesis is not rejected. However, when the chi-square value turns out to be zero, we have to be extremely careful to confirm that there is no difference, between the observed and expected frequencies. Such a situation sometimes arises on account of faulty method used in the collection of data.

1.7 LET US SUM UP

- The chi-square test is the most commonly used non-parametric test. Non parametric techniques are used when a researcher works with ordinal or nominal data or with a small number of cases representing a highly asymmetrical underlying distribution.
- The chi-square test is used as a test of significance when we have data that are given in frequencies or categories.
- The chi-square test is used for two broad purposes. Firstly it is used as a test of goodness of fit and secondly as a test of independence.

- As a test of goodness of fit chi-square tells us whether the frequencies observed among the categories of a variable are tested to determine whether they differ from a set of expected hypothetical frequencies.

- As a test of independence chi-square is usually applied for testing the relationship between two variables in two way. First by testing the null hypothesis of independence and second by computing the value of contingency coefficient, a measurement of relationship existing between the two variable.

- When we have less than 5 observed frequencies in any cell then we use the Yate's correction.

1.8 UNIT END QUESTIONS

1) Describe the properties of chi-square test.
2) What do you understand by the goodness of fit test ? Describe with examples.
3) What is the chi-square test for independence ? Explain with examples.
4) What precautions would you keep in mind while using chi-square.
5) Write short notes on: Yates correction and expected frequencies in contingency table

1.9 ANSWERS TO SELF ASSESSMENT QUESTIONS

1) i) F (ii) T (iii) T (iv) T (v) F

2) i) Negative (ii) Yates Correction (iii) Reject

1.10 GLOSSARY

Chi-square Distribution : A distribution with degree of freedom as the only parameter. It is skewed to the right for small degree of freedom.

Chi-square Test : A statistical techniques used to test significance in the analysis of frequency distribution.

Contingency Table : A table having rows and column wherein each row corresponds to a level of one variable and each column to a level of another variable.

Expected Frequencies : The frequencies for different categories which are expected to occur on the assumption that the given hypothesis is true.

Observed Frequencies : The frequencies actually obtained from the performance of an experiment.

1.11 SUGGESTED READINGS

Garrett, H.E. (1971), *Statistics in Psychology & Education*, Bombay, Vakils, Seffer & Simoss Ltd.

Guilford, J.P. (1973), *Fundamental Statistics in Psychology & Education*, Newyork, McGraw Hill.

UNIT 2 CONCEPT AND CALCULATION OF CHI-SQUARE

Structure

2.0 INTRODUCTION

As we have discussed in last unit that chi-square is a commonly used non-parametric statistical test. This is used when we have data in the form of frequencies, ordinal or nominal levels of measurement. Chi-square test is used for different purposes. In the last unit you were given the conceptual knowledge of chi-square test and in this unit we will discuss the method of computation of chi-square.

2.1 OBJECTIVES

After reading this unit, you will be able to:

- Identify Chi-square as a test of goodness of fit;
- Apply Testing of equal probability hypothesis;
- Use Testing of normal probability hypothesis;
- Calculate Chi-square as a test of indepedence in contingency tables; and
- Apply Chi-square test when table entries are small.

2.2 APPLICATION OF CHI-SQUARE TEST

The Chi-square test is used for comparing experimentally obtained results with those to be expected. The important application of Chi-square is as given below:

2.2.1 Chi-square as a Test of Goodness of Fit

Chi-Square goodness of fit test is used to find out how the observed value of a given phenomena is significantly different from the expected value. In Chi-Square goodness of fit test, the term goodness of fit is used in order to compare the observed sample distribution with the expected probability distribution. Chi-Square goodness of fit test determines how well theoretical distribution (such as normal, binomial, or Poisson) fits the empirical distribution. In Chi-Square goodness of fit test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval.

In regard to the procedure, set up the hypothesis for Chi-Square goodness of fit test, that is set upo both null and alternative hypothesis.

a) **Null hypothesis:** In Chi-Square goodness of fit test, the null hypothesis assumes that there is no significant difference between the observed and the expected value.

b) **Alternative hypothesis:** In Chi-Square goodness of fit test, the alternative hypothesis assumes that there is a significant difference between the observed and the expected value.

Calculate chi-square using the formula and find out for the given degrees of freedom if chi-square value is significant at .05 or .01 levels. If so reject the null hypothesis. If not significant accept the null hypothesis.

2.2.2 Testing Hypothesis of Equal Probability

The Chi-square test is a useful method of comparing experimentally obtained results with those to be expected theoretically on some hypothesis. The formula for calculating $\chi 2$ is

$$\chi 2 = \sum[(fo-fe)^2]/fe$$

Where;

fo = observed frequency of a phenomenon or even which the experimenter is studying

fe = expected frequency of the same phenomenon based on “no difference” or “null” hypotheses.

The use of the above formula can be illustrated by the following example.

Example

An attitude scale designed to measure attitude toward co-education was administered on 240 students. They have to give their response in terms of favorable, neutral and unfavorable. Of the members in the group 70 marked favorable, 50 neutral and 120 disagreed. Do these results indicate significant difference in attitude ?

The observed data is (fo) given in the first row of table 2.1. In the second row is the distribution of answer to be expected on the basis of null hypothesis (fe), if each answer is selected equally.

Table 2.1: Responses from subjects in regard to the attitudes

Calculations	Favourable	Neutral	Unfavourable	Total
Fo	70	50	120	240
Fe	80	80	80	240
fo-fe	-10	-30	40	
$(fo-fe)^2$	100	900	1600	
$(fo-fe)^2 / fe$	100 /80	900/80	1600/80	
X^2	1.25	11.25	20	$\sum(fo- fe)^2/fe = 32.50$

The formula of $\chi 2$ is

$\chi 2 = \sum[(fo-fe)^2/fe]$

$\chi 2 = 1.25+11.25+20=32.5$

d.f.=(r-1)(k-1)=2

Entering table of $\chi 2$ (which can be obtained form any book of statistic), we find in row d.f.=2 a value 5.59 and 9.19 given under the heading .05 and .01 levels of significance respectively. Our obtained value is 32.5, which is far above the given value in the table. Thus our results will be marked significant at .01 level. We discard the null hypothesis which stated that there will be no difference in the attitude. But from the results it may be stated quite confidently that there is difference in the attitudes of people towards coeducation. It may also be mentioned that these results clearly indicate that the results are not due to any chance factor.

Example

A personnel manager is trying to determine whether absenteeism is greater on one day of the week than on another. His record for the past year shows the following scores. Test whether the absence is uniformly distributed over the week.

Table 2.2: Record of absenteeism for one year

Calculation	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Fo	23	18	24	17	18	100
Fe	20	20	20	20	20	100
fo-fe	3	-2	4	-3	-2	
$(fo-fe)^2$	9	4	16	9	4	
$(fo-fe)^2/fe$	9/20=0.45	4/20=.20	16/20=.80	9/20=.45	4/20=.20	

The formula of $\chi 2$ is

$\chi 2 = \sum[(fo-fe)^2/fe] = 0.45+0.20+0.80+0.45+0.20 = 2.10$

$\chi 2 = 2.10$

d.f.=(r-1)(c-1)

d.f.=(5-1)(2-1)

Critical value of $\chi 2$ at .05 level=9.488 (refer to statistical table given at the end of the statistics book)

Critical value of $\chi 2$ at .01 level=13.277 (refer to statistical table given at the end of the statistics book)

The computed value of $\chi 2$, i.e. 2.10 is less than the critical values at .05 and .01significance levels, we conclude that $\chi 2$ is not significant and we retain the null hypothesis. We can say that the deviation of observed absenteeism from expectation might be a matter of chance.

On the other hand if the computed value of $\chi 2$ is more than 9.48 or 13.28, then the null hypothesis is rejected and it may be concluded that there is significant difference in the absenteeism that happens on different days of the week.

However in our example we have found the chi-square value being lower than what is given in the table and so we retain the null hypothesis stating that the absenteeism does not vary interms of the days and that it is purely a chance factor.

Steps for Chi-square Testing

1) First set a null of hypotheses.
2) Collect the data and find out observed frequency.
3) Find out the expected frequencies by adding all the observed frequencies divided by number of Categories (In Ist example 240/3=80, in II example 100/5=20).
4) Find out the difference between observed frequencies and expected frequencies.(fo-fe)
5) Find out the square of the difference between observed and expected Frequency. $(fo-fe)^2$
6) Divide it by expected frequency. $(fo-fe)^2/fe$. You will get a quotient
7) Find out the sum of these quotients.
8) Determine the degree of freedom and find out the critical value of $\chi 2$ from table.
9) Compare the calculated and table value of $\chi 2$ and use the following decision rule.

Accept null hypothesis if $\chi 2$ is less than critical value given in table.

Reject the null hypothesis if calculated value of $\chi 2$ is more than what is given in the table under .05 or .01 significance levels.

Self Assessment Questions

1) What are the uses of chi-square test?

.....

.....

.....

.....

.....

2) What do you understand by the term "Test of Goodness of Fit test".

.....

.....

.....

.....

.....

3) Describe with examples testing hypothesis of equal probability.

.....

.....

.....

.....

.....

.....

4) Enumerate the steps in chi-square calculation.

.....

.....

.....

.....

.....

.....

Testing the Divergence of Observed Results from those Expected on the Hypothesis of a Normal Distribution

Some times the expected frequencies are determined on the basis of the normal distribution of observed frequencies in the entire population. The method as to how this is done is explained in the following examples.

Example

180 school teachers were classified into six categories on the basis of their performance. Do these teachers differ significantly in their performance as compared to the one expected if performance is supposed to be distributed normally in our population of school teachers.

Categories	1	2	3	4	5	6	Total
No. of school teachers	15	25	45	50	35	10	180

Solution

Here expected frequencies are to be determined on the basis of the normal distribution hypothesis. For this purpose we have to divide the base line of the curve into six equal segments of 1σ each. To mention here, the baseline in any normal distribution ranges from from -3σ to $+3\sigma$. The table below presents how many cases fall between the different areas.

Table 2.3: Frequencies on the basis of normal distribution

S.No	Category	% of cases falling between the area	No. of cases out of 180
1.	+ 3.oo σ to 2.00 σ	2.28	3.6
2.	+ 2.00 σ to +1.00 σ	13.59	25.2
3.	+ 1.00 σ to M	34.13	61.2
4.	M to - 1.00 σ	34.13	61.2
5.	-1.00 σ to - 2.00 σ	13.59	25.2
6.	- 2.00 σ to -1.00 σ	2.58	3.6

We will now test the null hypothesis by computing the values of $\chi 2$ given in table.

Table 2.4: Calculation of chi-square step by step

	1	2	3	4	5	6	Total
Fo	15	25	45	50	35	10	180
Fe	3.6	25.2	61.2	61.2	25.2	3.6	180
fo-fe	11.4	-.2	-16.2	-11.2	9.8	6.4	
$(fo-fe)^2$	129.96	.04	262.44	125.44	96.04	40.96	
$(fo-fe)^2/ fe$	36.6	.001	4.29	2.05	3.81	11.38	

The formula of $\chi 2$ is

$\chi 2 = \sum[(fo-fe)^2/fe]$

$X^2 = 36.1+.001+4.29+2.05+3.81+11.38=55.63$

d.f.=(r-1)(c-1)

d.f.=(6-1)(2-1)=5

The critical value of $\chi 2$ is 11.0700 at .05 level and 15.086 at .01 level. Obtained values of $\chi 2$ is much greater than the critical values. Hence it can be said that $\chi 2$ is significant and we reject the null hypothesis.

Example

200 salesmen were classified into five categories on the basis of their performance ranging from excellent performance to very poor performance. Does this evaluation differ significantly from the one expected from the normal distribution?

The observed frequencies are given in the following table.

Table 2.5: Frequencies of salesman based on performance

Very good	Good	Average	Poor	Very poor
30	45	65	40	20

Solution

We have to find out the expected frequencies with the help of normal probability Curve. For this we have to divide the base line of the curve into five equal segments by 6/5=1.2

Table 2.6: Computation of expected frequencies on the basis of normal distribution

Category.	Area of normal curve	% cases falling between area	No. of cases out of 200
1	+ 3 to 1.8 σ	3.45%	6.90 or 7.00
2	+ 1.8 σ to +.60 σ	23.84%	47.68 or 48
3	+ 6 σ to -.60 σ	45.14%	90.28 or 90
4	+.6 σ to - 1.8 σ	23.84%	47.68 or 48
5	-1.8 σ to - 3.σ	3.45%	6.90 or 7

We will now test the null hypothesis by computing the values of $\chi 2$ given in table.

Table 2.7: Calculation of chi-square step by step

	Very good	Good	Average	Poor	Very poor	Total
fo	30	45	65	40	20	200
Fe	7	48	90	48	7	200
fo-fe	23	-3	-25	-8	13	
$(fo-fe)^2$	529	9	625	64	169	
$(fo-fe)^2/ fe$	75.57	.19	6.94	1.33	24.14	

$\chi 2=\sum[(fo-fe)^2\}/fe$

$\chi 2 = 75.57+.19+6.94+1.33+24.14=108.17$

d.f.=(r-1)(c-1)

d.f.=4

The critical value of $\chi 2$ at 4 df is 9.49 at .05 level and 13.29 at .01 level therefore obtained values of X^2 is much greater than the critical values. Hence it can be said that $\chi 2$ is significant and we reject the null hypothesis.

Steps

- Formulate the null hypothesis
- Find out the observed frequencies.

- To find out the expected frequencies first divide the base line of the normal curve (which is divided into 6 equal σ units) equal to the number of given categories of observed frequencies. For example in example 1 there are six categories therefore we have to divided the base line of the curve in 6 units. In example 2 there are 5 categories and we have to divide the base line of the curve in 5 units. Thus in the first case these can be obtained by 6/6= 1σ unit in example. In example 2 there are five categories and 5/6= 1.2 σ unit distance.
- Now we have to decide the areas of normal curve as shown in table.
- To Decide the number of cases or percentage of cases, refer table of normal probability. In example 2 in the first category the area of normal curve is + 3 σ to + 1.8 σ. In table of normal distribution 0.4986 i.e. 49.86% cases lies between means and 3 σ and 0.4641 i.e. 46.41% cases lies in between mean and 1.80 σ. Therefore 49.86-44.41=3.45% cases are there in the first category. In the second category the area of normal curve is + 1.80 σ to + .6 σ. There are .4641 i.e. 46.41% cases lies between mean and 1.8 σ and .2257 i.e. 22.57% cases lies in between mean and .6 σ . Therefore percentage of cases between + 1.80 σ to .6 σ will be (46.41-22.57) 23.84.
  - In the third category the areas of normal curve is + 6 σ to -6 σ here .2257 means 22.57% cases lie on +6 σ and 22.57 lies on 6 σ therefore (22.84+22.57) 45.14% cases lies in the third category. Because this is based on normal distributions which is symmetrically divided in two halves, therefore number of cases in fourth and fifth categories will be the same as in the second and first category.
- No. of cases out of total number of cases can be found by dividing the obtained areas of normal curve by 100 and multiply by total number of cases. For example in example 2, 3.45% of the cases falling in area I of first category is 3.45. Therefore the number of cases out of 200 will be (3.45 × 200)/100=6.90. In this way find out few.
- After calculating the fe place the values in the table and find out the χ^2 with the help of given formula.

Self Assessment Questions

1) How will you test the divergence of observed results from the expected on the hypothesis of a normal distribution?

.....

.....

.....

.....

.....

2) Enumerate the steps needed to find out the expected frequencies with the help of normaly probability curve.

.....

.....

.....

.....

2.2.3 Chi-square as a Test of Independence

In addition to testing the agreement between observed frequencies and those expected from some hypothesis, that is, equal probability and normal probability, chi-square may also be applied to test the relationship between variables. In this we test whether two variables are dependent or independent to each other. The following table is known as Contingency table. Contingency table is a double entry or two way table in which the possession by a group of varying degree of the characteristics is represented. The same person for example may be classified as on the basis of sex that is boys and girls or can be categoriesed in terms of the colour of their eyes i.e. brown or black etc. These attributes are noted in terms of the gender factor.

To cite another example, father and sons may be classified with respect to temperament or achievement and the relationship of the attributes in the groups may be studied.

In the following table there is a double entry or it is called a two way table in which 100 boys and 100 girls who express their view on the statement "There should be same syllabus in all the universities of the country," in terms of agree, neutral, disagree. The results are shown in the following table.

Table 2.7: Frequencies of respondents in terms of views expressed

Gender	Agree	Neutral	Disagree	Total
Boys	20	30	50	100
Girls	50	20	30	100
Total	70	50	80	200

Do you think that the opinion is influenced by the gender of the students.?

For this we present the data in contingency table form as above. The expected frequencies after being computed are written within brackets just below the respective observed frequencies. The contingency table and the process of computation of expected frequencies is given in the table 2.8.

Table 2.8: Contingency table with fo and fe for the data given in table 2.7

Gender	Agree	Neutral	Disagree	Total
Boys	20 (35)	30 (25)	50 (40)	100
Girls	50 (35)	20 (25)	30 (40)	100
Total	70	50	80	200

The method to find out the expected frequency for each cell is given below:

$70 \times 100/200 = 35$ $50 \times 100/200 = 25$ $80 \times 10/200=40$

$70 \times 100/200 = 35$ $50 \times 100/200 = 25$ $80 \times 100/200=40$

The value of χ^2 may be computed with the help of usual formula,

$\chi 2=\sum[(fo-fe)^2\}/fe$

Table 2.9: Computation of χ^2 from contingency table 2.5

fo	Fe	fo-fe	$(fo-fe)^2$	$(fo-fe)^2/fe$
20	35	-15	225	6.43
30	25	5	25	1
50	40	10	100	2.5
20	35	-15	225	6.43
30	40	-5	25	1
20	40	-10	100	2.5
				χ^2 =19.86

The value of χ^2 may be computed with the help of normal formula.

For 2 d.f. critical value at .05 level is 5.99 and at .01 level 9.21. Our obtained value of χ^2 is 19.86 . It is far higher than the table value. Therefore we reject the null hypothesis and conclude that gender influences the opinion.

Steps

1) Formulate the null hypothesis
2) Find out the expected values by the method shown in table.
3) Find out the difference between observed and expected values for each cell.
4) Square each difference and divide this in each cell by the expected frequency.
5) Add up these and the sum of these values gives χ^2.

2.2.4 2 × 2 Fold Contingency Table

When the contingency table is 2 x 2 fold χ^2 may be calculated without first computing the four expected frequencies. Example below illustrates the method.

Example

The mother of 180 adolescents (some of them were graduates and others non-graduates) were asked whether they agree or disagree on a certain aspect of adolescent behaviour. Is the attitude of their mother related to their educational qualification? Data are arranged in a four fold table below.

	Agree	**Disagree**	
Graduate mother	30 (A)	50 (B)	80 (A+B)
Non graduate mother	70 ©	30 (D)	100 (C+D)
Total	100 (A+C)	80 (B+D)	180 (A+B+C+D)

In four fold table Chi-square is calculated by the following formula:

$\chi^2 = N(AD-BC)^2/(A+B(C+D)(A+C)(B+D)$

Substitute for A,B,C,D, in the formula , we have

$= 180\ [(30\times30)-(50\times70)\}2/80\times100\times100x80$

$=180(2600)^2/64000000$

$=19.01$

$x^2=19.01$

d.f. = (r-1)(c-1)

= (2-1)(2-1)=4

On 1 df the critical value at .05 level is 3.84 and on .01 is 6.63.

The obtained values is much higher therefore χ^2 is significant at .01 level, we reject the null hypothesis and conclude that attitude is related with education.

Self Assessment Questions

Given below are statements. Indicate in each case whether the statement is true or false.

1) A 4×3 Contingency table has four Columns and three rows. (T/F)
2) The Yates' Correction is used when observe Frequency in any cell is less than (T/F)
3) When the calculated value of χ^2 is less than the critical value at a specified level of significance the null hypothesis is rejected. (T/F)
4) On account of simple calculation involved , χ^2 test is very frequently used by the statistician. (T/F)
5) Indicate how you will use chi-square as a test of independence.

.....

.....

.....

.....

.....

6) Enumerate the steps in calculation of chi-square in 2×2 contingency table.

.....

.....

.....

.....

.....

2.3 THE CHI-SQUARE TEST WHEN TABLE ENTRIES ARE SMALL (YATES' CORRECTION)

When χ^2 is computed from a table in which any experimental frequency is less than 5 then χ^2 is not stable. Moreover in 2 × 2 contingency table (df=1) there are more chances of error. When any observed frequency is less than 5. In these situations we make a correction for continuity known as Yates' correction.

To apply the Yates correction we subtract .05 from the absolute value of the difference between observed and expected frequency.

What is the reason for applying Yates correction and what is its effect upon χ^2 can be illustrated from the following example.

Example

A student is asked to respond on 10 objective type questions requiring response in 'yes' or 'no. 7 times the student says 'yes' and 3 times the student says 'No'. Is this result different from what would have been obtained by merely guessing ?

Table : Computation of χ^2 with Yates Correction

	Yes	No	Total
fo	7	3	10
fe	5	5	10
(fo-fe)	2	-2	
Correction (fo-fe-.5)	1.5	1.5	
$(fo-fe)^2$	2.25	2.25	
$(fo-fe)^2/ fe$	.45	.45	

If we does not apply the Yates correction than value of χ^2 will be different. This can be illustrated in the following table.

Table : Chi-square calculation

	Yes	No	Total
fo	7	3	10
fe	5	5	10
(fo-fe)	2	-2	
$(fo-fe)^2$	4	4	
$(fo-fe)^2/ fe$	4/5 =.20	4/5 = .20	

$\chi^2=\sum[(fo-fe)^2/fe]$

$X^2 = .2+.2 = 0.4$

d.f.=(3-1)(2-1)

d.f.=2

on 1 df the values at .05 is 3.84 and on .01 is 6.635.

In these examples although both the χ^2 are non significant but if we does not apply the correction than the probability of significance of χ^2 increased. We can conclude that failure to use the correction causes the probability of its being called significant considerably increased.

Yates Correction in 2×2 contingency

Example

60 Students 50 from urban areas and 10 from rural area were administered attitude scale to measure the attitude towards dating. Some of them expressed positive attitudes and some expressed negative attitudes. Whether the attitudes of the students will be effected by their belonging to rural or urban areas. The data are given in the following table.

	Positive	Negative	Total
Urban	20 (A)	30(B)	50
Rural	3©	7(D)	10
Total	23(A+C)	37(B+D)	N=60(A+B+C+D)

In this type of 2×2 table you can calculate chi-square in a different method which is given below:

$$\chi^2 = N(|AD-BC|- N/2)^2/(A+B(C+D)(A+C)(B+D)$$

$$= [60\{|20\times7-30\times3|\} -60/2]^2/ (50)(10)(23)(37)$$

$$= [60[|140-90|\}-30]^2/60$$

$$=60\times400/425500$$

$$=24000/42500$$

$$= .056$$

Entering the table for chi-square we find that for df= 1 the value of Chi-square at .05 level is be 3.841. The obtained value is less than the table value and hence we retain the null hypothesis and conclude that the attitude towards dating is not influenced by the person belonging to rural or urban areas.

2.4 LET US SUM UP

The Chi-square test is used for two broad purposes. It is used as a test of goodness of fit and second as a test of independence.

As a test of goodness of fit χ^2 tries to determine how the observed frequencies are different from the frequency expected theoretically by hypothesis of equal probability and hypothesis of normal probability.

The formula of χ^2 is

$$\chi^2=\sum[(fo-fe)^2/fe]$$

As a test of independence χ^2 is applied for testing the relationship between two variables in two ways. In the 2 × 2 contingency table there are four cells A,B,C,D, and χ^2 is calculated with the help of following formula:

$$\chi^2= N\{(AD-BC) –N/2/\}^2 / (A+B(C+D)(A+C)(B+D)$$

When in any cells entries are less than 5, Yate's correction is applied and following formula is used to calculate χ^2

$$\chi^2=\sum[\{(fo-fe) -.5\}^2/fe]$$

Yates Correction in 2 ×2 fold contingency table

When we have 2×2 fold contingency table (where df=1) sometimes it happens that in any one of the cell frequencies are less than five than Yates correction for continuity have to be applied.

The formula is

$$\chi^2= N(|AD-BC|)-N/2/(A+B(C+D)(A+C)(B+D)$$

The vertical lines {AD-BC} mean that the difference is to be taken as positive

2.5 UNIT END QUESTIONS

1) In 100 tosses of a coin, 40 heads and 60 tails were observed. Test the hypothesis that coin is fair using a significant level of (1) .05 (b) .01.

2) 42 lectures were rated by their principal in terms of their professional competency. The results are as below

Very effective	satisfactory	poor
16	20	6

3) Does the distributions of rating differ significantly from that to be expected if professional competency is normally disturbed in our population of lecture.

4) A groups of 100 college students were administered an attitude scale . The distribution of score on item 4 is shown below.

Strongly agree	Agree	Indifferent	Disagree	Strongly disagree
23	18	26	18	16

Based on the results obtained discuss the following:

i) On the basis of equal probability hypothesis whether obtained answers are different from expected answer.

ii) On the basis of normal probability hypothesis test, whether there is difference between observed frequency and expected.

5) The following table represent the number of boys and the number of girls who express their response on whether there should be co-education in school.

	Strongly agree	**Agree**	**In different**	**Disagree**	**Strongly disagree**
Boys	25	30	10	25	10
Girls	10	15	5	15	15

Do these data indicate significant sex difference in attitude towards co education.

6) 200 students express their, response on a five point scale to the statement "Maths should be compulsory for all the of College students. The responses are given in the following table.

Students	**Strongly agree**	**Agree**	**In different**	**Disagree**	**Strongly disagree**
Medical College	12	18	4	8	54
Engineering College	48	22	10	8	98
Law College	10	4	12	10	48

Test whether there exists significant difference in opinion of the students studying different subjects.

7) Depression inventory was administered on 190 persons. The table below shows the number of depressives and normals who choose each of the three possible answers to an item on the inventory.

	Yes	No	Undecided
Normals	14	66	10
Depressive patient	27	6	17

Based on the results indicate if this item is differentially responded by the two groups? Test the independence hypothesis.

8) In a study equal number of boys and girls were asked to express their preference for lecture method and discussion method. The data are given below:

Students	Preferred Lecture Method	Preferred Discussion Method
Boys	31	19
Girls	24	16

Determine whether gender of the students determine the preference of teaching methodology.

9) A student is asked to respond to 8 objective type questions. This response is shown in the following table:

Yes	No
6	2

Is this result different from what is expected.

Answers

1) **(i) F, ii), T, (iii) F, iv) F**

2) χ^2 =4 significant at .05, not significant at .01 level

3) χ^2 = 15.66 significant at.01

4) χ^2 = 3.20 not significant b χ^2 =148.35 significant at .01

5) χ^2 = 7.03not significant

6) χ^2 = 34.20 significant at.01

7) χ^2 = 4.14 not significant

8) χ^2 = 1.98 not significant

9) χ^2 = 1.12 not significant

2.6 GLOSSARY

Critical Value : Table value according to the given df value.

Expected Frequency : The expected scores in a given class.

Observed Frequency : The frequency actually obtained from the performance of an experiment.

Yate's Correction : The correction of -.5 in expected frequency.

Contingency table : A table having rows and columns where in each row corresponds to a level of one variable and each column to a level of another variable.

2.7 SUGGESTED READINGS

Kerlinger, Fred N. (1963). *Foundation of Behavioural Research*, (2nd Indian reprint) Surjeet Publication, New Delhi.

Garrett, H.E.(1971). *Statistics in Psychology and Education*, (Sixth Indian edition Vakils, Feffer and Simons Pvt. Ltd.

UNIT 3 SIGNIFICANCE OF THE DIFFERENCES BETWEEN MEANS (T-VALUE)

Structure

3.0 INTRODUCTION

In psychology some times we are interested in research questions like Do the AIDS patients who are given the drug AZT show higher T-4 blood cell counts than patients who are not given that drug? Is the error rate of typist the same when work is done in a noisy environment as in a quiet one? Whether lecture method of teaching is more effective than lecture cum discussion method?

Consider the question whether lecture method of teaching is more effective than discussion method. For this investigator divides the class in to two groups. One group is taught by lecture method and other by discussion method. After a few months researchers administer an achievement test for both the group and find out the mean achievement scores of the two groups

say M_1 and M_2. The difference between these two mean is then calculated. Now the questions is whether the difference is a valid difference or it is because of sampling fluctuation or error of sampling. Whether this difference is significant or not significant. Whether on the basis of this difference, could we conclude that one method of teaching is more effective than the other method.

These question can be answered by the statistical measures which we are going to discuss in this unit. To test the significance of difference between mean we can use either the t-test or Z test. When the sample size is large, we employ Z test and when sample is small, then we use the t test. In this unit we are concerned with t test. We will get acquainted with the various concepts related, to computation and description of t test.

3.1 OBJECTIVES

After reading this unit, you will be able to:

- Understand the need and importance of the significance of the difference between means;
- Know about what is null hypothesis, standard error, degrees of freedom, level of significance, two tailed and one tailed test of significance, type I error and type II error; and
- Calculate the significance of difference between mean (t-test) when groups are independent, when there are correlated groups, groups matched by pair and groups matched by mean and standard deviation.

3.2 NEED AND IMPORTANCE OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN MEANS

In psychology sometimes we are interested in knowing about the significance of the differences between populations. For example we are interested to discover whether ten year old boys and ten year old girls differ in their linguistic ability. Or we want to find out if children from high SES perform and score better academically than children from low SES. We may also try to find out at times, if two groups of persons coming from different background differ in their agility factor. Thus many questions are asked and to be answered in psychology for which one of the measures we use is the Mean.

Let us take the first question on linguistic ability of boys and girls. First we randomly select a large sample of boys and girls (large sample means the group comprises of 30 or more than 30 persons.). Then we administer a battery of verbal test to measure the linguistic ability of the two groups, compute the mean scores on linguistic ability test of the two groups. Let us say the obtained mean scores for boys and girls are M_1 and M_2. Now we try to find the difference between the two means. If we get a large difference($M_1 - M_2$) in favour of the girls then we can confidently say that girls of 10 years of age are significantly more able linguistically than 10 years old boys. On the contrary if we find small difference between two means then we would conclude that ten years old girls and boys do not differ in linguistic ability.

An obtained mean is influenced by *sampling fluctuation* or *error of sampling* and whatever differences are obtained in the means, it may be due to this sampling error. Even mean of population 1 and mean of the population 2 may be the same but because of sampling error we may find the difference in the range of 2 samples drawn from two populations. In order to test the significance of an obtained difference we must first have a standard error(SE) of the difference. Then from the difference between the sample mean and standard error of difference we can say whether the difference is significant or not. Now the question arises what do we mean by significant difference? According to Garrett (1981) a difference is called *significant* when the probability is high and that it cannot be attributed to chance that is (Temporary and accidental factors) and hence represent a true difference between population mean.

A difference is non significant when it appears reasonably certain that it could easily have arisen from sampling fluctuation and hence imply no real or true differences between the population means.

3.3 FUNDAMENTAL CONCEPTS IN DETERMINING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN MEANS

3.3.1 Null Hypothesis

This is a useful tool in testing the significance of differences. Null hypothesis asserts that there is no true difference between the two population means, and the difference found between the sample mean is therefore, accidental or unimportant (Garreet 1981). In the course of a study or an experiment, the null hypothesis is stated so that it can be tested for possible rejection. For example to study the significant difference in linguistic ability of 8 years old girls and boys we select random sample of girls and boys and administer a battery of verbal test, compute the means of the two groups. In this study the null hypothesis may be stated thus: there exists no significant difference between the linguistic ability of boys and girls. If this null hypothesis is rejected then we can say one group is superior to the other.

3.3.2 Standard Error

The primary objective of statistical inference is to make generalisation from a sample to some population of which the sample is part. Standard error measures (1) error of sampling and (2) error of measurement. Suppose we have knowledge of the true mean, means of the population, we randomly select 100 representative sample from the population and compute their mean and standard deviations. The standard deviation obtained from this representative sample is known as standard error of the mean. The standard error of the mean can be calculated by the following formula:

SEm or $\sigma_m = \sigma/\sqrt{N}$

Where

σ= The standard deviation of the sample mean

N= The number of cases in the sample.

If the standard error of measurement is large it shows considerable sampling error.

3.3.3 Degrees of Freedom

When a statistics is used to estimate a parameter the number of degrees of freedom (d.f) available depends upon the restriction placed upon the observations. One d.f. is lost for each restriction imposed. For example we have five scores as 5,6,7,8 and 9 the mean is 7 and deviation of our scores from 7 are -2, -1, 0, 1 and 2. The sum of these deviations is zero. In consequence if any four deviations are known the remaining deviations may be automatically determined. In this way, out of the five deviations, only four (N-1) are free to vary as, the condition that " the sum equals to "Zero" impose restriction upon the independence of the 5^{th} deviation. Originally there were 5(N=5) degrees of freedom in computing the mean because all the observation or scores were independent. But as we made use of the mean for computing standard deviation we lost one degree of freedom.

Degrees of freedom varies with the nature of the population and the restriction imposed. For example in the case of value calculated between means of two independent variables, where we need to compute deviation from two means, the number of restrictions imposed goes up to two consequently d.f. is (N-1+N-2).

3.3.4 Level of Significance

Whether a difference between the means is to be taken as statistically significant or not depends upon the probability that the given difference could have arisen "by chance". The researcher has to take a decision about the level of significance at which he will test his hypothesis.

In social sciences .05 and .01 level of significance are most often used. When we decide to use .05 or 5% level of significance for rejecting a null hypothesis it means that the chances are 95 out of 100 that is not true and only 5 chances out of 100 the difference is a true difference.

In certain types of data, the researcher may prefer to make it more exact and use .01 or 1% level of significance. If hypothesis is rejected at this level it shows the chances are 99 out of 100 that the hypothesis is not true and only 1 chance out of 100 it is true. The level of significance which the researcher will accept should be set by researcher before collecting the data.

3.3.5 Two Tailed and One Tailed Tests of Significance

In many situations we are interested in finding the difference between obtained mean and the population mean. Our null hypothesis states that the M_1 amd M_2 do not vary and the difference between them is zero. (i,e. Ho:M_1-M_2=0). Whether this difference is positive or negative we are not interested in the direction of such a difference. All that we are interested is whether there is a difference. For example we hypothesised that two groups will differ from each other we don't know which group will have higher mean scores and which group lower. This a *non directional hypothesis* and it gives rise to a two-tailed hypotheses test. In other words the difference may be in either direction and thus is said to be non directional.

In many experiments our primary concern is with the direction of the difference rather than with its existence in absolute term. For example if we are interested to determine the gain in vocabulary resulting from additions by weekly reading assignment. Here we are interested in finding out the gain in vocabulary. To take another example, if we say that training in yoga will

reduce the degree of tension in persons, then we are clearly stating that there will be a reduction in the tension. In cases like this we make use of the *one tailed* or *non directional* test to test the significance of difference between the means.

3.3.6 Errors in Making Inferences

If the null hypothesis is true and we retain it or if it is false we reject it, we had made a correct decision. But sometimes we make errors. There are two types of errors Type I error and Type II error.

A Type I error, also known as alpha error, is committed when null hypothesis (Ho) is rejected when in fact it is true.

A Type II error, also known as beta error, is committed when null hypothesis is retained and in fact it is not true. For example suppose that the difference between two population means (μ_1- μ_2) is actually zero, and if our test of significance when applied to the sample mean shows that the difference in population mean is significant we make a Type I error. On the other hand if there is true difference between two population mean, and our test of significance based on the sample mean shows that the difference in population mean is " not significant' we commit a type II error.

3.4 METHODS TO TEST THE SIGNIFICANCE OF DIFFERENCE BETWEEN MEANS OF TWO INDEPENDENT GROUPS (t-test)

3.4.1 Testing Significance of Difference Between Uncorrelated or Independent Means

The step used to find out significance of differences between independent mean are as below:

Step 1. Computation of the mean

Step 2. Computation of the combined standard deviation by using the following formula:

$x_1=X_1-M_1$ (deviation of scores of first sample from its mean)

$x_2= X_2-M_2$ (deviation of scores of second sample from its mean)

$SD = \sqrt{(\Sigma x_1^2+ \Sigma x_2^2)/(N_1-1)+(N_2-1)}$

Step 3. Computation of the Standard error of the difference between two means by using following formula:

$SED= SD / \sqrt{(1/N_1+1/N_2)}$

Step 4. To Compute the t value for the different in two independent sample mean. The following formula is used to determine *t* value is

$t = (M_1-M_2)-0/SED$

Step 5. Find out the degree to freedom. The (df) degree of freedom is calculated using the following formula

$df=(N_1-1)+(N_2-1)$

Step 6. We then refer to table of t (can be found in any statistic's book) distributions with the calculated degree of freedom df and read the t value

given under column .05 and .01of two tailed test. If our computed t value is equal or greater than the critical t value given in table then we can say that t is significant. If the computed value is lesser than given value then we will say that it is non-significant.

Let us illustrate the whole process with the help of example

Example: An interest test was administered to 6 boys and 10 girls. They obtained following scores is the mean difference between two groups significant ?

Table: Scores of boys and girls and the t value calculation

Scores of boys			**Scores of girls**		
X1	x1	x^2	X2	x2	$x2^2$
28	-2	4	20	-4	16
35	5	25	16	-8	64
32	2	4	25	1	1
24	-6	36	34	10	100
36	-4	16	20	-4	16
35	5	25	28	4	16
			31	7	49
			24	0	0
			27	3	9
			15	-9	81
$\sum X1=180$		$\sum x1^2= 110$	$\sum X2=240$		$\sum x2^2 =352$

Calculation

$M_1=(\sum X_1/N1)$

$M_1=180/6 =30$

$M_2=(\sum X_2 / N_2)$

$M_1=240/10 =24$

$SD= \sqrt{} [(\sum x^2_1+\sum x^2_2)/(N_1-1)+(N_2-1)]$

$SD = \sqrt{} [(110+352)/(6-1)+(10-1)] = 5.74$

$SEd= SD [\sqrt{}((N_1+N_2)/(N_1N_2)]$

$= 5.74 / \sqrt{} [(16)/(60)]$

$= 5.74 \times .5164 = 2.96$

$t= (M_1-M_2)-0/SE_D$

$= (30-20)-0/2.96 = 2.03$

$df=(N_1-1)+(N_2-1)$

$df=(6-1)+(10-1) =14$

Entering the value in table on 14 d.f. We get 2.14 at the .05 and 2.98 at the.01 level, since our t is less than 2.14, therefore we will say that the mean difference between boys and girls is non significant.

Let us take another example

Example

On an academic achievement test 31 girls and 42 boys obtained the following scores.

Table: Mean scores and Standard deviation

	Mean	SD	N	df
Boys	40.39	8.69	31	30
Girls	35.81	8.33	42	41

Is the mean difference in favour of boys and significant at the .05 level.

First we will compute the pooled standard deviation by the following formula

$SD = \sqrt{[(SD_1)^2 \times (N_1-1) + (SD_2)^2 \times (N_2-1)]/(N_1+N_2)}$

SD_1= Standard Deviation of group 1 i.e. 8.69

N_1= Number of subject is group 1 i.e. 31

SD_2= Standard devotion of groups 2 i.e. 8.33

N2= Number of subject in group 2 i.e. 42

$SD = [(8.69)^2 \times (31-1)+(8.33)^2 \times (42-1)/(N_1-1)+(N_2-1)]$

SE_D= SD $[((N_1+N_2)/(N_1N_2)]$

SE_D= 8.48 $[(31+42)/(31X42)]$

$t= (M_1-M_2)-0/SE_D$

$= (40.39-35.81)-0/2.01 = 2.28$

$df=(N_1-1)+(N_2-1)$

$df=(31-1)+(42-1) =71$

Entering Table with 71 df we find t entries of 2.00 at the .05 level and 2.65 at the .01 level. The obtained t of 2.28 is significant at .05 level but not at the .01 level. We may say boys academic achievement is better in comparison to that girls.

3.5 SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO CORRELATED MEANS

3.5.1 The Single Group Method

In the previous section we discussed the problem of determining the significance of difference between mean obtained by two independent groups of boys and girls.

Some time we have single group and we administer the same test twice. For example if we are intending to find out the effect of training on the students' educational achievement, first we take the measures of subject's educational achievement before training , then we introduce the training programme, and again we take the measure of educational achievement. In this we have single group and administer educational achievement test twice. Such type of design is known as single group design. In order to get the significance of the

difference between the means obtained in the before training and after training we use the following method.

SED or $\sigma_D = [\sigma^2_{M1} + \sigma^2_{M2} - 2r\,\sigma_{M1}\,\sigma_{M2}]$

Where

σ_{M1}= Standard error of the initials test

σ_{M2}=Standard error of the finals test

r= coefficient of correlation between scores on initial test and final test

$t=(M_1-M_2)-0/\sigma_D$

Let us illustrate the above formula will the help of following example.

Example

At the beginning of the session an educational achievement test in maths was given to 100 IX grade students. Their mean was 55.4 and SD was 7.2. After six months an equivalent from of the same test was given and the mean was 56.9 and SD was 8.0. The correlation between scores made on the first testing and second testing was .64. Has the class made significant progress in maths during six month? We may tabulate our data in the following manner

Table: Scores in the initial and final test of students

	Initial Test	Final Test
No. of students	100	100
Mean	55.4	56.9
SD	7.2.	8.0
Standard error of the mean	0.72	0.80
r12	.64	

Calculation

SED or $\sigma_D = [\sigma^2_{M1} + \sigma^2_{M2} - 2r\,\sigma_{M1}\,\sigma_{M2}]$

SEd= $[(.72)^2+(.80)^2-2x.64x.72x.80]$

= $[.5184+.6400-.7373] = .649$

$t= (M_1-M_2)-0/SE_D$

$= 1.5-0/.649 = 2.31$

$df=(N_1-1)=(100-1)$, df=99

From Table we look at the df = 99, and find that the value at .05 level is 1.98 and at .01 level is 2.63. Our obtained value is 2.31 therefore this value is significant at .05 level and not on .01 level. Here we can say that class made substantial improvement in mathematical achievement in six months.

3.5.2 Difference Method

When groups are small then we must prefer the difference method to that given above.

Let us illustrate the use of this method with the help of following example.

Example

Ten subjects were tested on an attitude scale. Then they were made to read some literature in order to change their attitude. Their attitude were again measured by the same scale. The results of the initials and final testing are as under.

Table: Results of initial and final testing of attitude

Initial condition	Final condition	Difference Cond 2- cond 1 (Mean = 8)	X = D-M	x^2
50	62	12	4 (12-8)	16
42	40	-2	-10(-2--8)	100
35	30	-5	-13(-5-8)	169
51	61	10	2 (10-8)	4
42	52	10	2 (10-8)	4
26	35	9	-1 (9-8)	1
41	51	10	2 (10-8)	4
42	52	10	2 (10-8)	4
60	68	8	0 (8-8)	0
70	84	14	-6 (14-8)	36
55	63	8	0 (8-8)	0
38	50	12	4 (12-8))	16
Total		$\sum D = 96$		$\sum x^2 = 354$

Note: The sum of scores of final condition is more than sum of initial condition therefore we subtract scores of initial condition from scores of final condition (Final condition –Initial condition) and add the score to find ΣD.

Mean = $(\sum D)/N$

Mean = 96/12=8

$SD_D = \sqrt{[(\sum x^2)/N-1]}$

$SD_D = \sqrt{[354/11]} = 5.67$

$SE_{MD} = SD_D/ (N)$

$SE_{MD} = 5.67/ (12)$

$t = MD-0/SE_{MD}$

t=8-0/1.64

t=4.88

d.f.=12-1

Entering in table with 11 df, we find t entries of 2.20 and 3.11 at the .05 and at the .01 levels. Our t of 4.88 is far above the .01 level. We can conclude that subjects attitude changed significantly from initial to final condition.

3.5.3 The Method of Equivalent Groups

In experiments when we want to compare the relative effect of one method of treatment over another we generally take two groups, one is known as experimental group and the other is known as control group. Here we have two groups not a single group. For the desired results these two groups need to be made equivalent. This can be done by (i) Matched pair technique or (ii) Matched groups technique. These are explained below

i) Matched pair technique: In this techniques matching is done by pair. Matching is done on variables which are going to affect the results of the study like age, intelligence, interest, socio-economic status.

ii) Matched groups technique: In this technique instead of person to person matching, matching of groups is carried out in terms of Mean and S.D.

3.5.4 Matching by Pair

Formula for calculation of standard error of difference between mean is:

SED or $\sigma_D = [\sigma^2_{M1} + \sigma^2_{M2} - 2r\, \sigma_{M1}\, \sigma_{M2}]$. Here

σ^2_{M1}= Standard error of mean 1

σ^2_{M2}= Standard error of mean 2

r=correlation between the two groups scores

$t = (M_1 - M_2) - 0 / SE_D$

Example

Two groups X and Y of Children 72 in each group are paired child to child for age and intelligence. Both groups were given group intelligence scale and scores were obtained. After three weeks experimental group subjects were praised for their performance and urged to try better. The control groups did not get the incentive. Group intelligence scale again was administered on the groups. The data obtained were as follows. Did the praise affect the performance of the group or is there a significant difference between the two groups.

Table:

	Experimental group	**Control group**
No. of children in each group	72	72
Mean score of final test	88.63	83.24
SD of final test	24.36	21.62
Standard error of the mean of final test	2.89	2.57
Correlation between experimental and control group scores	.65	

SED or $\sigma_D = [\sigma^2_{M1} + \sigma^2_{M2} - 2r\, \sigma_{M1}\, \sigma_{M2}]$

SE_D or $\sigma_D = [(2.89)^2 + (2.57)^2 - 2x.65x\ 2.89x2.57]$

$= 2.30$

$t = (88.63-83.24)-0/2.30 = 2.34$

d.f. = 72-1= 71

If we see the table we find that at 71 d.f. the value at .05 is 2.00 and at .01 is 2.65. The obtained t is 2.34 therefore this value is significant at .05 level and not at .01 level. On the basis of the results is can be said that praise did have significant effect in stimulating the performance of children.

3.5.5 Groups Matched for Mean and SD

When groups matched in terms of mean and S.D., the following formula is used to calculate 't'.

$$SE_D = \sqrt{[\sigma^2_{M1} + \sigma^2_{M2}\,(1-r^2]}$$

$$t = (M_1-M_2)-0/\ SE_D$$

SE_D= Standard error of difference

σ^2_{M1} = Standard error of mean 1

σ^2_{M2} = Standard error of mean 2

r= Correlation between final scores of two tests

The above formula can be illustrated by the following example

Example

The 58 students of academic college and 72 students of technical college were matched for mean and SD upon general intelligence test. Then the achievement on a mechanical aptitude test was compared. The Questions is do the two groups enrolled in different courses differ in mechanical ability?

	Academic	**Technical**
No. of children in each group	58	72
Mean on Intelligence GTest (Y)	102.50	102.80
SD on Intelligence Test Y	33.65	31.62
Mean on Mechanical Aptitude (X)	48.52	53.51
SD on Mechanical Aptitude X	10.60	15.36
r 12	.50	

$$SE_D = [\sigma^2_{M1\,/\,N1]} + [\sigma^2_{M2\,/\,N2\,]}\text{-}(1\text{-}r^2)$$

Therefore

SE_D or $\sigma_D = [(10.60)^2/58 + (15.35)_2/72]\text{-}[1\text{-}.25] = 1.97$

$t = (53.61\text{-}48.52)\text{-}0/1.97 = 2.58$

$d.f.=(N_1\text{-}1)+(N_2\text{-}1)=(58\text{-}1)+(72\text{-}1)=128$

Entering the value in table we find that on 125 df (which is near to 128) the value are 1.98 at .05 level and 2.62 at .01 our obtained value is 2.58. This is significant at .05 level. We may say that two group differ in mechanical aptitude.

3.6 LET US SUM UP

In the field of psychology some time we are interested in testing the significance of difference between two sample means.

The sample may comprise of two independent groups and single groups tested twice. Some time we have two groups matched by pair or matched for means and S.D. The process of determining the significance of difference between

the Means is different in different conditions. We may broadly summarise the procedure of calculating significance of differences between Means as under.

- Establish a null hypothesis.
- Decide a suitable level of significance .05 or .01
- Determine the standard error of the difference between means of two samples.
- Compute the value of 't'
- Find out the degrees of freedom.
- Determine the critical value of t from the 't' table.
- If the computed value is same or more than the value given in the table then it is taken to be significant if the computed valued is less than the given value it is considered as non significant.
- When the t-value is significant we reject the null hypothesis and when 't' value is not significant we retain the null hypothesis.

3.7 UNIT END QUESTIONS

1) Given below are some statements. Indicate whether the statement is true or false

   i) We commit a Type I error when we reject a null hypothesis when it is really true. (T/F)

   ii) In testing a hypothesis, one can make three types of error. (T/F)

   iii) An exercise in hypothesis testing enables us to draw conclusions about the estimated parameters. (T/F)

   iv) For a given level of significance, we find that as the sample size increases, the critical values of t get closer to zero. (T/F)

   v) If the standardised sample mean exceeds the critical clue, we should accept Ho. (T/F)

2) Differentiate between

   1) Null hypothesis and alternative hypothesis.
   2) One tailed test and two tailed test.
   3) Type I error and Type II error.

3) Write short notes on the following

   Concept of Standard error

   Level of significance

4) The marks obtained in math's by 10 boys and 10 girls are given in the following table. Find out whether their is any difference between the mean of boys and girls.

Scores of boys	Scores of girls
7	7
5	6
6	5
5	8
6	9
6	8
7	8

5) One group of boys (N=20) and one groups of girls (N=22) were tested on verbal ability test. They got following scores:

	Boys	Girls
Mean	34.56	30.56
SD	5.68	6.98

Do the groups differ on the verbal ability test?

6) Ten persons are tested before and after the experimental procedure, their scores are given below. Test the hypothesis that there is no change.

Before	After
60	72
52	50
61	71
36	45
45	40
52	62
70	78
51	61
80	94
65	73
72	82

7) We take two groups one from technical classes and other from non technical class and each group is compared on numerical learning test. Both groups have been matched on in terms of means and standard deviation on the basis of scores on general Intelligence test. Do the groups differ in terms of mean numerical ability?.

The data are as under:

	Non technical	Technical
N	58	72
Mean of Numerical reasoning test	48.52	53.61
SD on Numerical reasoning test	10.61	15.35
r between Intelligence and Numerical reasoning test	.50	

8) A Vocabulary test was administered to a random sample of 8 students of section A and and 7 students of section B of Class IX of a school . The scores are:

Scores on Section A	16	14	12	12	10	8	6	4
Scores on Section B	14	8	7	6	4	4	1	2

Is the difference between the means of two groups significant at o.05 level?

9) In the first trial of a practice period, 25 twelve-year-olds have a mean score of 80.00 and a SD of 8.00 upon a digit-symbol learning test. On the tenth trial, the mean is 84.00 and the SD is 10.00. The r between scores on the first and tenth trials is .40. Our hypothesis is that practice leads to gain.

Answers 1

i) True (ii) False, (iii) True, (iv) True, (v) False

Answers 4 t= 1.66 non significant

Answers 5 t= 1.99 non significant

Answers 6 t= 4.88 significant at .01 level

Answers 7 t= 2.57 significant at .05 level

Answers 8 t= .80 not significant

Answers 9 t= 2.00 significant at .05 level

3.8 GLOSSARY

Null Hypothesis : A zero difference hypothesis. A statement about the status Quo about a population parameter that is being tested.

Alternative Hypotheses: A hypothesis that takes a value of population parameter different from that used in the null hypothesis. It states that there is a difference in the groups on a certain characteristic that is being tested.

Type I error : An error caused by rejecting a null hypothesis when it is true.

Type II : An error caused by failing to reject a null hypotheses when it is not true.

One tail test : A statistical hypothesis test in which the alternative hypothesis is specified such that only one direction of the possible distribution of values is considered. It would state there will be an increase in the performance of students after training.

Two tailed Test : A statistical hypothesis test in which the alternative hypothesis is stated in such way that it included both the higher and the lower values of a parameter than the value specified in the null hypothesis, It would state that there will be a difference (can be an increase or decrease) in the group that undergoes training.

Significance level : The probability that a given difference arises because of a chance factor or it is a true difference.

Standard error : The standard deviation of a sampling distributions.

3.9 LIST OF FORMULA

t test for two independent group

$t = (M_1-M_2)-0/SED$

$SED = SD/\sqrt{(1/N_1+1/N_2)}$

$SD = \sqrt{(\Sigma x_1^2+ \Sigma x_2^2)/(N_1-1)+(N_2-1)}$

t-test for single group (difference method)

$Mean = (\sum D)/N$

$SD_D = \sqrt{[(\sum x^2)/N-1]}$

$SE_{MD} = SD_D/\sqrt{(N)}$

$t = MD-0/SE_{MD}$

t –test for groups matched for mean and S.D

$t = (M_1-M_2)/SE_D$

SED or $\sigma_D = [(\sigma^2_{M1}+ \sigma^2_{M2})(1-r^2)]$

t –test for groups matched for paired

SED or $\sigma_D = [\sigma^2_{M1}+ \sigma^2_{M2}- 2r \sigma_{M1} \sigma_{M2}]$

$t= (M_1-M_2)-0/SE_D$

3.10 SUGGESTED READINGS

Garrett ,H.E. (1981) *Statistics in Psychology and Education*, Bombay, Vakils, Feffer and Simons Ltd.

UNIT 4 NORMAL DISTRIBUTION: DEFINITION, CHARACTERISTICS AND PROPERTIES

Structure

4.1 INTRODUCTION

The word probability is a part of our daily lives. We use it quite frequently in our day to day life. We ask such questions. How likely it is that I will get an A grade in this exam ? It is likely to rain heavily this evening. How likely it is that a price of equity shares of company will increase in the next few days?

While common man answers these questions in a vague and subjective way, the researcher attempts to give the answer to these questions in a more objective and precise way. There are different types of probability distributions. Normal distribution is a kind of probability distribution,. If we look around us we will see that persons differ in terms of attributes like intelligence, interest, height, weight etc. Take for example intelligence, it can be seen that majority of us possess average intelligence i.e. IQ between 85-110 and the persons who have above average i.e. IQ 145 and above or below average IQ i.e. less than 70 etc., will be very few. Similarly if we see the height of the persons we will find that the height of the maximum number of persons range between 5.2 to 6 feet. The number of persons having height less than 5 feet and more than 6 feet is relatively very few. Similar type of trend we will find in the biological field, Anthropometrical data, social and economic data. If we plot these variations or data in the form of a distribution

or put it in the form of graph, we would get distribution known as normal curve or normal distribution. In this unit we will discuss, what is probability, different types of probabilities and the concept of normal distribution.

4.2 OBJECTIVES

After completion of this unit, you will be able to understand:

- Define probability and explain different types of probability;
- Describe meaning of normal distribution;
- Identify the characteristics of the normal distribution;
- Analyse the properties of the normal distribution; and
- Apply the normal distribution.

4.3 DEFINITIONS OF PROBABILITY

In the previous units you have learned the way to describe variables. We generate hypothesis, collect the data, categorise the data and summarise the data by computing measures of central tendency and variability.

Our interpretation and conclusions about variables are based on what we observed. But here our approach will be some what different. We will first suggest certain theories, propositions or hypothesis about variables, which will then be tested using the data we observe. The process of testing hypothesis through analysis of data is probability.

According to Beri (2007), **"**Probability is the chance that a particular event will occur." (What is the chance of getting a head when a coin is tossed.). To take another example, A company has launched new product what is the chance that it will be successful?

According to Levin and Fox (2006): "The term probability refers to the relative likelihood of occurrence of any given outcome or event."

Probability associated with an event is the number of times an event can occur relative to the total number of times any event can occur.

Probability of an outcome or event = The total number of times the occurrence of the event / the total possible times an event can occur.

For example, if in a room there are three women and seven men, the probability that the next person coming out of the room is a woman would be 3 in 10.

Probability of a women coming out next = number of women in the room / total number of men and women in the room

= 3/10 = .30

The probability of an event not occuring is known as converse rule of probability.

4.3.1 Types of Probability

There are two types of probability, one based on theoretical mathematics and the other based on systematic observation.

Theoretical probabilities reflect the operation of chance along with certain assumption we make about the events. For example the probability of getting

a head on a coin flip is .5 (1/2 = .5). The probability of guessing the correct answer of five item multiple choice question is .20 (1/5).

Empirical probabilities are those for which we depend on observation to determine their value. For example, the probability that Indian team wins a cricket match is about .6 (6 out of 10 matches) a 'fact' we know from observing hundreds of games with various countries over a year.

In both form probabilities (P) varies from 0 to 1.0. In most situations, the percentage and not the decimal is used to express the level of measurement. For example 0. 5 probability means 50% chance.

A zero probability means impossible and 1.00 probability means certainty.

4.3.2 Probability Distribution

A probability distribution is directly analogous to a frequency distribution. The only difference is probability distribution is based on the theory (probability theory), while frequency distribution is based on empirical data. In a probability distribution, first we specify the possible values of a variable and calculate the probability associated with each.

There are three types of probability distribution: the Binomial distribution, the poisson distribution and the normal distribution. Here we are interested in normal distribution.

4.4 THE NORMAL DISTRIBUTION

The concept of normal distribution is very important in statistical theory and practice.

In 1973 the French mathematician Abraham de Moivere discovered the formula of the normal curve. In the 19th Century Gauss and Laplace rediscovered the normal curve independently Gauss was primarily interested in the problem of astronomy which led to the consideration of a theory of error of observation. In the middle of the 19th Century Quetelet promoted the applicability of the normal curve. He believe that the normal cure could be extended to apply to problem of anthropology sociology and human affair.

In the latter part of the 19th century Sir Francis Galton began the first serious study of individual differences and during his systematic study he found that most of the physical and psychological traits of human being conformed reasonably well to the normal curve. In this way he extended the applicability of the normal curve. Normal curve is also known as *Gaussion Curve and bell shaped curve*.

A normal curve is one which graphically represents normal distribution. A normal distribution is one in which majority of the cases falls in the middle of the scale and small number of cases are located at both extremes of the scale. In psychology most of the traits are normally distributed, for example, if we administer an intelligence test on randomly selected large sample, we will find that the greatest proportion of IQ scores fall between 85 and 115. We would see a gradual falling off of scores on either side with few 'geniuses' who score higher than 145 and equally few who score lower than 55.

So far as physical human characteristic is concerned, most adults would fall within the 5 to 6 feet range of height, with far fewer being either very short (less than 5 feet) or very tall (more than 6 feet).

Normal probability distribution is a continuous probability distribution. It represents the frequency with which a variable occurs when the occurrence of that variable is governed by the laws of chance.

The normal curve is a theoretical or ideal model that was obtained from a mathematical equation rather than from actually conducting research and gathering data.

The normal curve takes into account the law which states that greater is the deviation of an event from the mean value in a series the less frequently it occurs.

In social sciences we conduct the study on representative sample and not on the entire population. Therefore, in actual practice the slightly deviated or distorted bell shaped curve is also accepted as the normal curve.

Self Assessment Questions

1) Given below are statements, indicate in each statement whether it is true or false.

   i) A distribution where mean and median have different value is a normal distribution. T/F

   ii) The right and left tail of the normal curve touch the horizontal axis. T/F

   iii) A probability distribution is based on actual observation. T/F

   iv) Probability varies from 0 to 1. T/F

   v) The distribution are said to be skewed negatively when there are many individuals in a group with their scores higher than the average score of the group. T/F

2) Fill in the blanks :

   i) The mean median mode are _____ in normal curve.

   ii) A _____ indicates how for an individual raw scores falls from the mean of a distribution.

   iii) The _____ indicates how the scores in general scattee around the mean.

   iv) _____ cases lie between the mean and ±1s on the base line.

   v) To find the deviation from the point of departure (i.e.) mean _____ of the distribution is used as a unit of measurement.

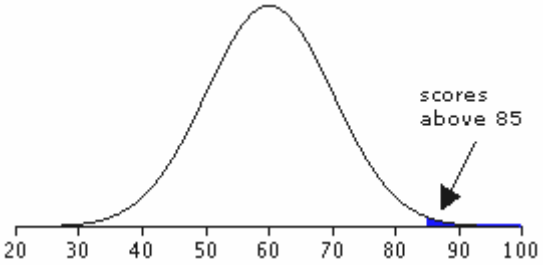
4.5 DEVIATION FROM THE NORMALITY

Although most of the variables in social sciences approximate the theoretical notion of normal distribution but some variables in social science do not conform to the theoretical notion of the normal distribution and they deviate from the normal distribution. This deviation from normality tends to vary in two ways.

4.5.1 Skeweness

Skeweness refers to lack of symmetry. A normal curve is perfectly symmetrical, there is a perfect balance between the right and left halves of the curve. For this curve, mean, median and mode are at the same point. A distribution is said to be 'skewed' when the mean and median fall at different

points in the distribution and the balance is shifted to one side or the other – that is to the left or to the right (Garrete 1981). Look at the figure given below.



Properties of a normal curve

- The normal curve is one of a number of possible models of probability distributions.
- The normal curve is not a single curve, rather it is an infinite number of possible curves, all described by the same algebraic expression:

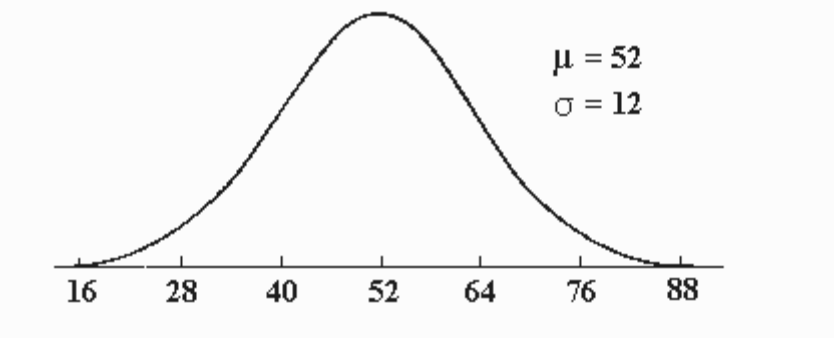
Similarity of Normal Curves of varied data include the following:

1) Shape
2) Symmetry
3) Tails approaching but never touching the X-axis, and
4) Area under the curve.
5) Bilaterally symmetrical
6) Most of the area under normal curve falls within a limited range of the number line.
7) All normal curves have a total area of 1.00 under the curve. This implies that the area in each half of the distribution is .50 or one half.

Drawing a normal curve

The standard procedure for drawing a normal curve is to draw a bell-shaped curve and an X-axis.

1) A tick is placed on the X-axis corresponding to the highest point (middle) of the curve.
2) Then, three ticks are placed to both the right and left of the middle point. These ticks are equally spaced and include all but a very small portion under the curve.
3) The middle tick is labeled with the value of □
4) Sequential ticks to the right are labeled by adding the value of σ .
5) Ticks to the left are labeled by subtracting the value of σ from □ for the three values.
6) For example, if M=52 and σ =12, then the middle value would be labeled with (52+ 12)= 64, then +12 = 76, and + 12 = 88, and the three points to the left would have the values (52 – 12) =40, then 28, and then 16. An example is presented below:



The two parameters, M and σ, each change the shape of the distribution in a different manner.

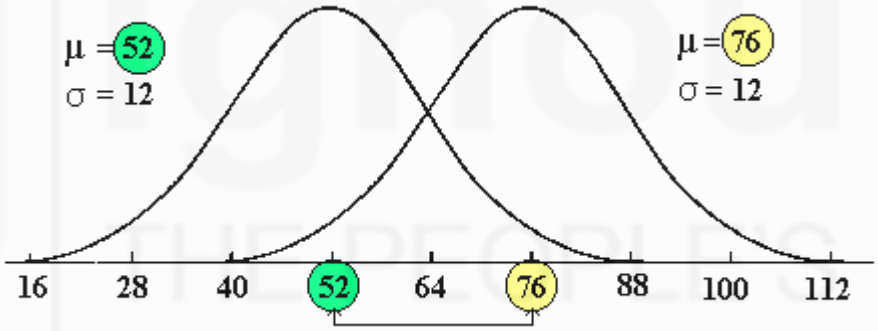
The first, M determines where the midpoint of the distribution falls.

Changes in M, without changes in σ, result in moving the distribution to the right or left.

That is, it depends on whether the new value of M was larger or smaller than the previous value.

At the same time, it does not change the shape of the distribution.

An example of how changes in M (μ) affect the normal curve are presented below:

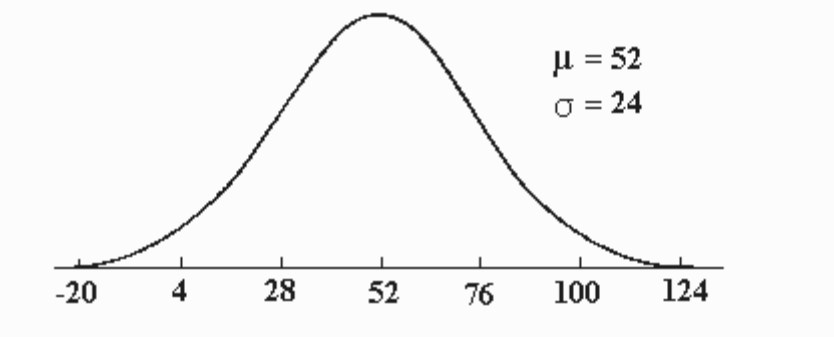


Changes in the value of σ , on the other hand, change the shape of the distribution without affecting the midpoint, because σ affects the spread or the dispersion of scores.

The larger the value of σ , the more dispersed the scores;

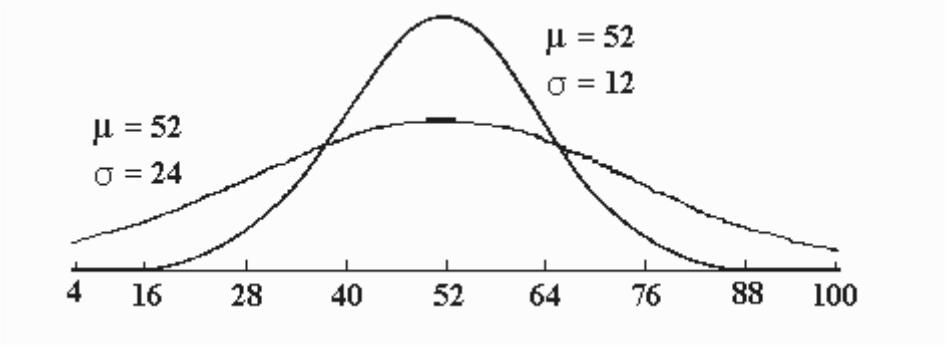
The smaller the value, the less dispersed.

The distribution below demonstrates the effect of increasing the value of σ



Suppose the second distribution was drawn on a rubber sheet instead of a sheet of paper and stretched to twice its original length in order to make the

two scales similar. Drawing the two distributions on the same scale results in the following graph:

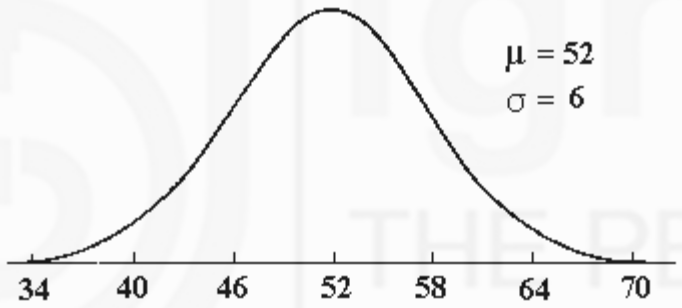


Note that the shape of the second distribution has changed dramatically, being much flatter than the original distribution.

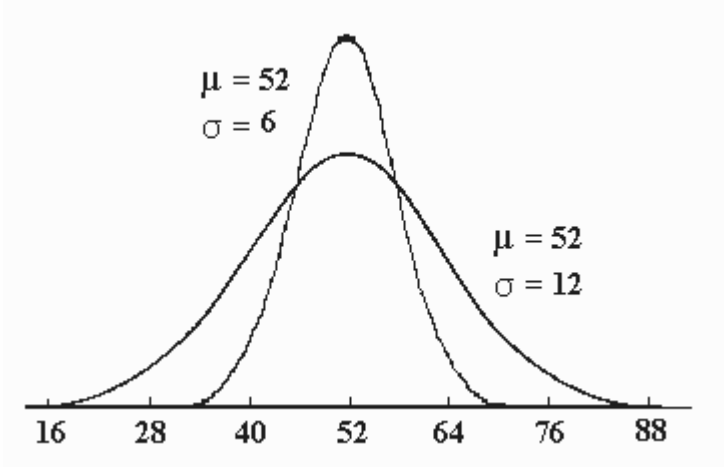
It must not be as high as the original distribution because the total area under the curve must be constant, that is, 1.00.

The second curve is still a normal curve; it is simply drawn on a different scale on the X-axis.

A different effect on the distribution may be observed if the size of σ is decreased. Below the new distribution is drawn according to the standard procedure for drawing normal curves:



Now both distributions are drawn on the same scale, as outlined immediately above, except in this case the sheet is stretched before the distribution is drawn and then released in order that the two distributions are drawn on similar scales:



Note that the distribution is much higher in order to maintain the constant area of 1.00, and the scores are much more closely clustered around the value of σ , or the midpoint, than before.

Skewness in a given distribution may be computed by the following formula.

3 (Mean - Median)

Skewness =

Standard deviation

In case when the percentiles are known, the value of skewness may be computed from the following formula :

$$S_k = \frac{P_{90} + P_{10} - P_{50}}{}$$

4.5.2 Kurtosis

The term Kurtosis refers to the peakedness or flatness of a frequency distribution as compared with the normal (Garrete 1981).

Kurtosis is usually of three types :

Platykurtic. A frequency distribution is said to be playkurtic, when it is flatter than the normal.

Leptokurtic. A frequency distribution is said to be leptokurtic, when it is more peaked than the normal.

Mesokurtic. A frequency distribution is said to be mesokurtic, when it almost resembles the normal curve (neither too flattened nor too peaked).

Self Assessment Questions

1) Discuss the various deviations from normality.

……..

……..

……..

……..

2) What is kurtosis. Enumerate the different types of kurtosis

……..

……..

……..

……..

3) How does change in the mean affect the normal curve?

……..

……..

……..

4.6 CHARACTERISTICS OF A NORMAL CURVE

The following are the characteristics of the normal curve.

Normal curves are of *symmetrical distribution*. It means that the left half of the normal curve is a mirror image of the right half. If we were to fold the curve at its highest point at the center, we would create two equal halves.

The first and third quartiles of a normal distribution are *equidistance from the median*.

For the curve the mean median and mode all have the same value.

In skewed distribution mean median and mode fall at different points.

The normal curve is *unimodal*, having only one peak or point of maximum frequency that point in the middle of the curve.

The curve is a *asymptotic*. It means starting at the centre of the curve and working outward, the height of the curve descends gradually at first then faster and finally slower. An important situation exists at the extreme of the curve. Although the curve descends promptly toward the horizontal axis it never actually touches it. It is therefore said to be asymptotic curve.

In the normal curve the *highest ordinate is at the centre*. All ordinate on both sides of the distribution are smaller than the highest ordinate.

A large number of scores fall relatively close to the mean on either side. As the distance from the mean increases, the scores become fewer.

The normal curve involves a *continuous distribution*.

4.7 PROPERTIES OF THE NORMAL DISTRIBUTION

In the following paragraphs we will discuss the properties of the normal distribution.

4.7.1 The Equation of the Normal Curve

$$y = N x^2 / s\ 2p^e\ 2s^2$$

Here :

x = Scores (expressed as deviation from the mean) laid off along the base line or x axis.

y = the height of the curve above the x.

N = Number of cases.

s = standard deviation of the distribution.

p = 3.1416 (the ratio of the circumstances of a circle to its diameter).

e = 2.7183 (base of the Napierian system of logarithms)

When N and s are known, then with the help of above formula we can compute (1) the frequency (or Y) of a given value x; and (2) the number between the points. But these calculations are rarely necessary as tables are available from which this information may be readily obtained.

4.7.2 Area Under the Normal Curve

It is important to keep in mind that the normal curve is an ideal or theoretical distribution (that is, a probability distribution). Therefore, we denote its mean

by m and its standard deviation by s. The mean of the normal distribution is at its exact center. The standard deviation (s) is the distance between the mean (m) and the point on the base line just below where the reversed S-shaped portion of the curve shifts direction.

To employ the normal distribution in solving problems, we must acquaint ourselves with the area under the normal curve : the area that lies between the curve and the base line containing 100% or all of the cases in any given normal distribution.

When normally distributed, it is seen that 34.13% cases lie between the mean and 1 s above the mean. In the same way we can say that 47.72% of the cases under the normal curve lie between mean and 2 s above the mean and 49.87% lie between the mean and 3 s above the mean.

The symmetrical nature of the normal curve leads us to make another important point. Any given sigma distance above the mean contains the identical proportion of cases as the same sigma distance below the mean.

Thus, if 34.13% of the total area lies between the mean and 1s above the mean, then 34.13% of the total area also lies between the mean and 1s below the mean; if 47.72% lies between the mean and 2s above the mean, then 47.72% lies between the mean and 2s below the mean; if 49.87% lies between the mean and 3s above the mean, then 49.87% also lies between the mean and 3s below the mean. In other words, 68.26% of the total area of the normal curve (34.13% + 34.13%) falls between - 1s and + 1s from the mean; 95.44% of the area (47.72% + 47.72%) falls between - 2s and + 2s from the mean; and 99.74%, or almost all, of the cases (49.87% + 49.87%) falls between - 3s and + 3s from the mean. It can be said, then, that six standard deviations include practically all the cases (more than 99%) under any normal distribution.

For example an intelligence test was administered on large randomly selected sample of girls. The obtained mean (m) was 100 and standard deviation was 10. Then we can say that 68.26% of the population would have IQ scores that falls between 90 (100-10) and 110 (100 + 10).

Moving away from the mean we would find that 99.74% of these cases would fall between score 70 and 30 (between - 3s to + 3s).

4.7.3 Table of Areas Under the Normal Curve

Suppose we want to determine the percent of total frequency that falls between the mean, and say a raw score located 1.40 s above the mean. A raw score 1.40s above the mean is obviously greater than 1s, but less than 2s from the mean. It means that this distance from the mean would include more than 34.13% but less than 47.72% of the total area under the normal curve. To determine the exact percentage within this interval we have to employ Table A given in any statistical book under the heading “area under the Normal curve”.

In Table A the total area under the curve is taken arbitrarily to the 10,000 because of the greater case with which fractional parts of the total area may then be calculated.

The first column of the table, x/s gives the distance that lie tenth of s measured off on the base line of the normal curve from the mean as origin. We have seen the deviation from the mean as $x = x - M$. If x is divided by s, deviation from the mean is expressed in s units. Such s deviation scores are often called Z scores (Z = x/s).

To find the number of cases in the normal distribution between the mean and 1s from the mean, go down the x/s column until 1.0 is reached and in the next column under .00 take the entry opposite 1.0 viz 3413. This figure mean 34.13% of the total frequencies falls between the mean and 1s. To find out the percentage of the distribution between the mean and 1.57s, go down the x/s column to 1.5 then across horizontally to the column headed .07 and take the entry 4418. This means that in a normal distribution 44.18% of the N lies between mean and 1.57 s.

Since the curve is bilaterally symmetrical, the entries in Table A apply to s distance measured in the negative or positive direction, which ever we need. For example to find out the percentage of the distribution between the mean and -1.26s take the entry in the column headed .06, opposite 1.2 in the x/s column. The entry is 3962 it means that 39.62% of the cases fall between the mean and -1.26s.

Self Assessment Questions

1) What are the characteristics of a normal curve?

……..

……..

……..

……..

……..

2) What are the properties of a normal curve?

……..

……..

……..

……..

……..

4.8 APPLICATION OF THE NORMAL CURVE

In psychological researches the normal curve has the main practical application given below :

- A normal curve helps in transforming the raw scores into standard scores.
- With the help of normal curve we can calculate the percentile rank of the given scores.
- A normal curve is used to find the limits in any normal distribution which include a given percentage of the cases.
- We can compare two distributions in terms of overlapping with the help of normal curve.
- A normal curve is used to determining the relative difficulty of test questions, problems and other test items.

- When the trait is normally distributed normal curve is used to separate a given group into subgroups according to capacity.

4.9 LET US SUM UP

In this chapter we introduced the concept of probability, indicated by the number of times an event can occur relative to the total number of times. A frequency distribution is based on actual observation whereas probability distribution is theoretical idea. There are three types of probability distribution i.e. Binomial distribution, the poisson distribution and normal distribution. It is a symmetrical bell shaped curve. It is not skewed. Normal curve can be used to determine the percent of the total area under the normal curve associated with any given sigma distance from the mean. Any given sigma distance above the mean contains the identical properties of cases as the same sigma distances below the mean.

4.10 UNIT END QUESTIONS

1) What is a normal curve ? Why is it named as Gaussion curve.
2) What do you understand by the term divergence from normality ? Point out the main types of such divergent curve.
3) Discuss the main characteristics of a normal curve.
4) What are the application of normal distribution

Answer:

1) (i) F (ii) F (iii) F (iv) T (v) T
2) (i) Identical (ii) Z score (iii) Standard deviation
   (iv) 68.26% (v) Standard deviation

4.11 GLOSSARY

Continuous random variable : A random variable than can assume any value within a given range.

Normal distribution : A symmerical bell shaped curve. The two tail of the curve never touch the horizontal axis.

Random variable : A variable that assume a unique numerical value for each of the outcomes is a sample space of a probability experiment.

Standard score : Known as Z scores also can be obtained by taking the deviation from mean and divided by standard deviation.

4.12 SUGGESTED READINGS

Beri G.C. (2007), *Business Stastics,* (2nd ed.) New Delhi, Tata MCgraw Hill.

Levin, J. & Fox, J.A. (2006) *Elementary Statistics in Social Research* (10th ed.) India, Pearson Education.