DJP3E - PSYCHOLOGICAL STATISTICS

UNIT I: INTRODUCTION

Meaning and definition of statistics – origin, growth and characteristics – applications in psychology and limitations. Primary and Secondary Data: Differences and data collection methods

Unit II: DATA CLASSIFICATION & FREQUENCY DISTRIBUTION:

Classification and Tabulation: Objectives – types of classification – formation of continuous frequency distribution – uses of tabulation – parts of a table – types of tables – simple and complex tables – general purpose and special purpose tables; Diagrammatic and graphic Representation: General rules for construction– uses –Types - limitations of diagrams and graphs.

UNIT III: DESCRIPTIVE STATISTICS:

Averages: Concepts – requisites of a good average –mean, median and mode –merits and demerits – numerical computations; Dispersion: Concepts – types of measures– merits and demerits – numerical computations

UNIT IV: SAMPLING, PROBABILITY AND THEORETICAL DISTRIBUTIONS

Concept of population and sample – census – requisites of a sample - Random & Non Random sampling methods– sampling and non-sampling errors; Meaning of probability – theorems of probability - Poisson and normal distributions – Skewness and kurtosis

UNIT V: INFERENTIAL STATISTICS:

Parametric and Non-parametric tests: Meaning – Rules of using – Chi-square and contingency coefficients: Meaning and assumptions – numerical computations - Correlation and Regression: Meaning– correlation and regression coefficients – numerical computations;

TEXTBOOKS:

1. Verma, J. P., & Ghufran, M. (2012). Statistics for Psychology: A comprehensive text. Tata McGraw Hill Education, New Delhi.
2. Garrett, H.E. (1979): Statistics in Psychology and Education, 9th Indian Reprint, Bombay, wakils, Feffer and Simons Pvt. Ltd.

REFERENCE BOOKS:

1. Gupta, S.P. (2006): Statistical Methods, New Delhi: Sultan Chand and Sons.
2. Howell, D.C. (2002): Statistical Methods for Psychology, 5th edition, Australia Duxbury Publishers.
3. Howell, D.C. (2002): Statistical Methods of Psychology. 5th edition. Australia, Duxbury Publishers.
4. Minium, E.W., King B.M. and Bear. G. statistical Reasoning in psychology and Education N.Y: John Wiley & sons, end 2001.
5. Gravetter F.J. and Wallnay L.B. Essentials of statistics for the Bahavional sciences N.Y. West Publishing com., 1995.

UNIT I: INTRODUCTION

1.1 Meaning and definition of statistics

The word ‘Statistics’ and ‘Statistical’ are all derived from the Latin word Status, means a political state. The theory of statistics as a distinct branch of scientific method is of comparatively recent growth. Research particularly into the mathematical theory of statistics is rapidly proceeding and fresh discoveries are being made all over the world.

Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis. Statistics is concerned with the systematic collection of numerical data and its interpretation. The word ‘ statistic’ is used to refer to

1. Numerical facts, such as the number of people living in particular area.
2. The study of ways of collecting, analysing and interpreting the facts.

Statistics is defined differently by different authors over a period of time. In the olden days statistics was confined to only state affairs but in modern days it embraces almost every sphere of human activity. Therefore a number of old definitions, which was confined to narrow field of enquiry were replaced by more definitions, which are much more comprehensive and exhaustive. Secondly, statistics has been defined in two different ways – Statistical data and statistical methods. The following are some of the definitions of statistics as numerical data.

Statistics are numerical statement of facts in any department of enquiry placed in relation to each other. - A.L. Bowley

Statistics are the classified facts representing the conditions of people in a state. In particular they are the facts, which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement.

Statistics are measurements, enumerations or estimates of natural phenomenon usually systematically arranged, analysed and presented as to exhibit important interrelationships among them.

Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis. It is clear that the definition of statistics by Croxton and Cowden is the most scientific and realistic one. According to this definition there are four stages:

1. **Collection of Data:** It is the first step and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of data such as census, sampling, primary, secondary, etc., and the investigator should make use of correct method.
2. **Presentation of data:** The mass data collected should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular or diagrammatic or graphic form.
3. **Analysis of data:** The data presented should be carefully analysed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, regression etc.,
4. **Interpretation of data:** The final step is drawing conclusion from the data collected. A valid conclusion must be drawn on the basis of analysis. A high degree of skill and experience is necessary for the interpretation.

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

1.2 Origin, Growth and Characteristics

1.2.1 Characteristics of Statistics

i. *Statistics are aggregate of facts:* A single age of 20 or 30 years is not statistics, a series of ages are. Similarly, a single figure relating to production, sales, birth, death etc., would not be statistics although aggregates of such figures would bestatistics because of their comparability and relationship.

ii. *Statistics are affected to a marked extent by a multiplicity of causes:* A number of causes affect statistics in a particular field of enquiry, e.g., in production statistics are affected by climate, soil, fertility, availability of raw materials and methods of quick transport.

iii. *Statistics are numerically expressed, enumrated or estimated:*The subject of statistics is concerned essentially with facts expressed in numerical form—with theirquantitative details but not qualitative descriptions. Therefore, facts indicated by terms such as ‘good’, ‘poor’ are not statistics unless a numerical equivalent, is assigned to each expression. Also this may either beenumerated or estimated, where actual enumeration is either not possible or is very difficult.

iv. *Statistics are numerated or estimated according to reasonable standard of accuracy:* Personal bias and prejudices of the enumeration should not enter into the counting or estimation of figures, otherwise conclusions from the figures would not be accurate. The figures should be counted or estimated according to reasonable standards of accuracy. Absolute accuracy is neither necessary nor sometimes possible in social sciences. But whatever standard of accuracy is once adopted, should be used throughout the process of collection or estimation.

v. *Statistics should be collected in a systematic manner for a predetermined purpose:* The statistical methods to be applied on the purpose of enquiry since figures are always collected with some purpose. If there is no predetermined purpose, all the efforts in collecting the figures may prove to be wasteful. The purpose of a series of ages of husbands and wives may be to find whether young husbands have young wives and the old husbands have old wives.

vi. *Statistics should be capable of being placed in relation to each other:* The collected figure should be comparable and well-connected in the same department of inquiry. Ages of husbands are to be compared only with the corresponding ages of wives, and not with, say, heights of trees.

1.2.2 Functions of Statistics

The functions of statistics may be enumerated as follows :

(i) To present facts in a definite form : Without a statistical study our ideas are likely to be vague, indefinite and hazy, but figures helps as to represent things in their true perspective. For example, the statement that some students out of 1,400 who had appeared, for a certain examination, were declared successful would not give as much information as the one that 300 students out of 400 who took the examination were declared successful.

(ii) To simplify unwieldy and complex data : It is not easy to treat large numbers and hence theyare simplified either by taking a few figures to serve as a representative sample or by taking average to give a bird's eye view of the large masses. For example, complex data may be simplified by presenting them in the form of a table, graph or diagram, or representing it through an average etc.

(iii) To use it as a technique for making comparisons : The significance of certain figures can be better appreciated when they are compared with others of the same type. The comparison between two different groups is best represented by certain statistical methods, such as average, coefficients, rates, ratios, etc.

(iv) To enlarge individual experience : An individual's knowledge is limited to what he can observe and see; and that is a very small part of the social organism. His knowledge is extended n various ways by studying certain conclusions and results, the basis of which are numerical investigations. For example, we all have general impression that the cost of living has increased.

But to know to what extent the increase has occurred, and how far the rise in prices has affected different income groups, it would be necessary to ascertain the rise in prices of articles consumed by them.

(v) To provide guidance in the formulation of policies : The purpose of statistics is to enable correct decisions, whether they are taken by a businessman or Government. In fact statistics is a great servant of business in management, governance and development. Sampling methods are employed in industry in tacking the problem of standardisation of products. Big business houses maintain a separate department for statistical intelligence, the work of which is to collect, compare and coordinate figures for formulating future policies of the firm regarding production and sales.

(vi) To enable measurement of the magnitude of a phenomenon : But for the development of the statistical science, it would not be possible to estimate the population of a country or to know the quantity of wheat, rice and other agricultural commodities produced in the country during any year.

1.2.3 Importance of Statistics

These days statistical methods are applicable everywhere. There is no field of work in which statistical methods are not applied. According to A L. Bowley, ‘A knowledge of statistics is like a knowledge of foreign languages or of Algebra, it may prove of use at any time under any circumstances”. The importance of the statistical science is increasing in almost all spheres of knowledge, e g., astronomy, biology, meteorology, demography, economics and mathematics. Economic planning without statistics is bound to be baseless.

1.3 Applications in psychology

Psychological statistics is application of formulas, theorems, numbers and laws to psychology. Statistical Methods for psychology include development and application statistical theory and methods for modeling psychological data. These methods include psychometrics, Factor analysis, Experimental Designs, Multivariate Behavioral Research.

Statistics allow psychologists to:

Organize Data: When dealing with an enormous amount of information, it is all too easy to become overwhelmed. Statistics allow psychologists to present data in ways that are easier to comprehend. Visual displays such as graphs, pie charts, frequency distributions, and scatterplots make it possible for researchers to get a better overview of the data and to look for patterns that they might otherwise miss.

Describe Data: Think about what happens when researchers collect a great deal of information about a group of people. The U.S. Census is a great example. Using statistics, we can accurately describe the information that has been gathered in a way that is easy to understand. Descriptive statistics provide a way to summarize what already exists in a given population, such as how many men and women there are, how many children there are, or how many people are currently employed.

Make Inferences Based Upon Data: By using what's known as inferential statistics, researchers can infer things about a given sample or population. Psychologists use the data they have collected to test a hypothesis or a guess about what they predict will happen. Using this type of statistical analysis, researchers can determine the likelihood that a hypothesis should be either accepted or rejected.

1.4 Limitations of statistics

Statistics with all its wide application in every sphere of human activity has its own limitations. Some of them are given below.

i. *Statistics is not suitable to the study of qualitative phenomenon:* Since statistics is basically a science and deals with a set of numerical data, it is applicable to the study of only these subjects of enquiry, which can be expressed in terms of quantitative measurements. As a matter of fact, qualitative phenomenon like honesty, poverty, beauty, intelligence etc, cannot be expressed numerically and any statistical analysis cannot be directly applied on these qualitative phenomenons. Nevertheless, statistical techniques may be applied indirectly by first reducing the qualitative expressions to accurate quantitative terms. For example, the intelligence of a group of students can be studied on the basis of their marks in a particular examination.

ii. *Statistics does not study individuals:* Statistics does not give any specific importance to the individual items, in fact it deals with an aggregate of objects. Individual items, when they are taken individually do not constitute any statistical data and do not serve any purpose for any statistical enquiry.

iii. *Statistical laws are not exact:* It is well known that mathematical and physical sciences are exact. But statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.

iv. *Statistics table may be misused:* Statistics must be used only by experts; otherwise, statistical methods are the most dangerous tools on the hands of the inexpert. The use of statistical tools by the inexperienced and untraced persons might lead to wrong conclusions. Statistics can be easily misused by quoting wrong figures of data. As King says aptly ‘statistics are like clay of which one can make a God or Devil as one pleases’ .

v. *Statistics is only, one of the methods of studying a problem:* Statistical method do not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy or religion into consideration. Thus the statistical study should be supplemented by other evidences.

1.6 Primary and Secondary Data

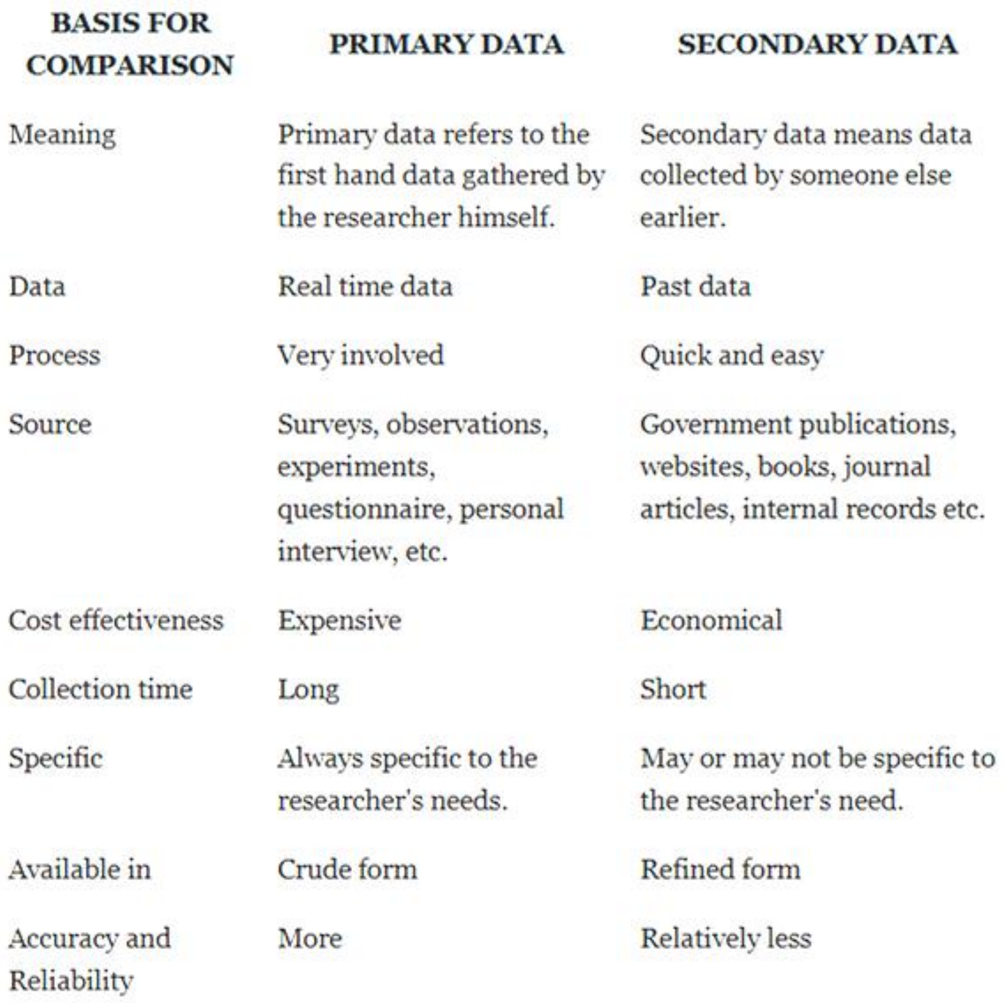
Primary data is data originated for the first time by the researcher through direct efforts and experience, specifically for the purpose of addressing his research problem. Also known as the first hand or raw data. Primary data collection is quite expensive, as the research is conducted by the organisation or agency itself, which requires resources like investment and manpower. The data collection is under direct control and supervision of the investigator.

The data can be collected through various methods like surveys, observations, physical testing, mailed questionnaires, questionnaire filled and sent by enumerators, personal interviews, telephonic interviews, focus groups, case studies, etc.

Secondary data implies second-hand information which is already collected and recorded by any person other than the user for a purpose, not relating to the current research problem. It is the readily available form of data collected from various sources like censuses, government publications, internal records of the organisation, reports, books, journal articles, websites and so on.

Secondary data offer several advantages as it is easily available, saves time and cost of the researcher. But there are some disadvantages associated with this, as the data is gathered for the purposes other than the problem in mind, so the usefulness of the data may be limited in a number of ways like relevance and accuracy.

There are many differences between primary and secondary data, which are discussed in this article. But the most important difference is that primary data is factual and original whereas secondary data is just the analysis and interpretation of the primary data. While primary data is collected with an aim for getting solution to the problem at hand, secondary data is collected for other purposes.



BASIS FOR COMPARISON	PRIMARY DATA	SECONDARY DATA
Meaning	Primary data refers to the first hand data gathered by the researcher himself.	Secondary data means data collected by someone else earlier.
Data	Real time data	Past data
Process	Very involved	Quick and easy
Source	Surveys, observations, experiments, questionnaire, personal interview, etc.	Government publications, websites, books, journal articles, internal records etc.
Cost effectiveness	Expensive	Economical
Collection time	Long	Short
Specific	Always specific to the researcher's needs.	May or may not be specific to the researcher's need.
Available in	Crude form	Refined form
Accuracy and Reliability	More	Relatively less

1.7 Methods of Data Collection:

Methods of Collection of Primary Data

The primary methods of collection of statistical information are the following :

1. Direct Personal Observation,
2. Indirect Personal Observation,
3. Schedules to be filled in by informants
4. Information from Correspondents, and
5. Questionnaires in charge of enumerators

The particular method that is decided to be adopted would depend upon the nature and availability of time, money and other facilities available to the investigation.

1. Direct Personal Observation: According to this method, the investigator obtains the data by personal observation. The method is adopted when the field of inquiry is small. Since the investigator is closely connected with the collection of data, it is bound to be more accurate. Thus, for example, if an inquiry is to be conducted into the family budgets and giving conditions of industrial labour, the investigation himself live in the industrial area as one of the industrial workers, mix with other residents and make patience and careful personal observation regarding how they spend, work and live.

2. Indirect Personal Observation: According to this method, the investigator interviews several persons who are either directly or indirectly in possession of the information sought to be collected. It may be distinguished form the first method in which information is collected directly from the persons who are involved in the inquiry. In the case of indirect personal observation, the persons from whom the information is being collected are known as witnesses or informants. However it should be ascertained that the informants really passes the knowledge and they are not prejudiced in favour of or against a particular view point. This method is adopted in the following situations:

a) Where the information to be collected is of a complete nature.

b) When investigation has to be made over a wide area.

c) Where the persons involved in the inquiry would be reluctant to part with the information.

3. *Schedules to be filled in by the informants:* Under this method properly drawn up schedules or blank forms are distributed among the persons from whom the necessary figure are to be obtained. The informants would fill in the forms and return them to the officer incharge of investigation. The Government of India issued slips for the special enumeration of scientific and technical personnel at the time of census. These slips are good examples of schedules to be filled in by the informants.

The merit of this method is its simplicity and lesser degree of trouble and pain for the investigator. Its greatest drawback is that the informants may not send back the schedules duly filled in.

4. *Information from Correspondents:* Under this method certain correspondent are appointed in different parts of the field of enquiry, who submit their reports to the Central Office in their own

manner. For example, estimates of agricultural wages may be periodically furnished to the Government by village school teachers.

The local correspondents being on the spot of the enquiry are capable of giving reliable information.

But it is not always advisable to place much reliance on correspondents, who have often got their own personal prejudices. However, by this method, a rough and approximate estimate is obtained at a very low cost. This method is also adopted by various departments of the government in such cases where regular information is to be collected from a wide area.

Sources of Secondary Data

There are number of sources from which secondary data may be obtained. They may be classified as follow. :

1. Published sources, and
2. Unpublished sources.

i. *Published Sources:* The various sources of published data are:
    a) Reports and official publications of-
        1. International bodies such as the International Monetary Fund, International Finance Corporation, and United Nations Organisation.
        2. Central and State Governments- such as Report of the Patel Committee, etc.
    b) Semi Official Publication. Various local bodies such as Municipal Corporation, and Districts Boards.
    c) Private Publication of—
        1. Trade and professional bodies such as the Federation of India, Chamber of Commerce and Institute of Chartered Accountants of India.
        2. Financial and Economic Journals such as “Commerce”, ‘Capital’ etc.
        3. Annual Reports of Joint Stock Companies.
        4. Publication brought out by research agendas, research scholars, etc.

ii. *Unpublished Sources:* There are various sources of unpublished data such as records maintained by various government and private offices, studies made by research institutions, scholars, etc., such source can also be used where necessary.

UNIT II: DATA CLASSIFICATION & FREQUENCY DISTRIBUTION:

Statistics deals with numerical data. The word 'data' is plural, the singular being ‘datum’ meaning fact. Usually in statistics, the term 'data' means evidence or facts describing a group or a situation. In practical sense, data indicates numerical facts such as measures of height, weight, scores on tests like achievement tests, intelligence tests, creativity tests or scores on any measuring instrument.

The data in its original from is known as raw data. That is, they are taken as such without any classification or organization. For example the marks obtained by ten students in an English achievement test are given as

20, 30, 35, 15, 10, 22, 36, 38, 21, 16.

These marks (scores) constitute the raw data. But the raw data will not be meaningful to the investigator or the data will be too large to manipulate. For getting a clear picture about the data and making the analysis possible, one has to classify and tabulate the information. Then raw data are grouped together or organized in such a way that the features of the data are revealed. In grouped data, the individual score has no meaning, but the characteristics of the total data is revealed. Grouping held the investigator to know about the distribution and makes the analysis easier.

2.1 Classification and tabulation

The collected data after editing (avoiding incomplete or incorrect information) is classified. Classification is the process of grouping of related data into classes. It is the first step of tabulation. Classification helps to organize the data in a tabular form. This can be done in two ways; either as discrete frequency table or continuous frequency table.

Frequency table

A frequency table is an arrangement of raw data revealing frequency of each score or of a class.

For preparing a discrete frequency distribution (table), first place all the possible values of the variable from lowest to highest. Then put a vertical line (tally) against the particular value for each item in the given data. Usually blocks of five bars are prepared by crossing the four bars already marked by the fifth one. Finally count the number of bars and write the frequency.

Example:

The achievement scores of 50 students are given below.

48 42 37 35 41 27 28 30 27 31
43 41 38 34 42 28 31 37 28 35
21 27 36 28 26 31 38 28 27 30
42 37 35 41 36 34 37 29 28 40
43 41 42 36 37 41 37 28 26 27

This data can be converted into discrete frequency table by the steps described below.

1. First list individual scores that are included in the data without repetition in ascending order.
2. Then put a tally mark against each score when ever it occurs in the raw data.
3. Add the tally marks which represent the frequency of each data.
4. Write the sum of the frequencies which will be equal to the total number of scores.

Marks	**Tally marks**	**Frequency**
21	*I*	1
26	*II*	2
27	~~*IIII*~~	5
28	~~*IIII*~~ *II*	7
29	*I*	1
30	*II*	2
31	*III*	3
34	*II*	2
35	*III*	3
36	*III*	3
37	~~*IIII*~~ *1*	6
38	*II*	2
40	*I*	1
41	~~*IIII*~~	5
42	*IIII*	4
43	*II*	2
48	*I*	1
	Total	**50**

In the case of discrete variables, one can prepare such types of frequency distribution. But, discrete frequency table again may be lengthy especially when the data contains a large number of scores and a long series. For proper handling of data, adequate organization will be needed. For this, numerical data will be grouped into some groups or classes and the frequency of each class is found out by putting tally marks against each class when an individual score belongs to that class.

The systematic steps of forming a frequency distribution are given below.

Step 1: Calculate the range

Range is the difference between the lowest and highest score in the set of data. That is Range = H – L where H is the highest score and L is the lowest score.

Step 2: Determination of class interval

The number and size of the classes are to be decided. The number of classes is decided according to the number of scores included. Usually the number of classes is limited to 20, but if number of items are small, say 50, the number of classes he class size or class interval denoted as 'i' is calculated using the forr

$$i = \frac{Range}{Number\ of\ classes}$$

where 'i' should be taken as a whole number and hence nearest approximate number can be taken as the class interval.

Step 3: Writing the classes

Classes are written from lowest to highest from bottom to top. The lowest class isprepared so that it contains the lowest score in the set and the last class is written so that the highest number is included in that class.

Step 4: Marking tally and writing the frequency

Mark tally against each class for items belonging to that class. In the next column, write the frequency, which represents the number of items in that class. The total of the third column 'frequency' should be equal to the number of data. In the above example, the least score is 21 and largest is 48.

$$\text{Hence Range} = 48 - 21$$

$$= 27$$

Here the total number of data is 50 and we can classify it into approximately 10 classes. Then the class interval will be

$$i = \frac{Range}{Number\ of\ classes}$$

= 27 / 10 = 2.7

Class interval may be taken as a whole number, and hence ‘i’ can be approximated to 3.

Now we have to write the classes. The first class should include 21 and the last class should include 48.

Classes	**Tally marks**	**Frequency**
47 – 49	*I*	1
44 – 46		0
41 – 43	*~~IIII~~ ~~IIII~~ I*	11
38 – 40	*III*	3
35 – 37	*~~IIII~~ ~~IIII~~ II*	12
32 – 34	*II*	2
29 – 31	*~~IIII~~ I*	6
26 – 28	*III ~~IIII~~ ~~IIII~~*	14
23 – 25		0
20 – 22	*1*	1
		N = 50

Here 20- 22 is the first class and it is taken so that the lowest score 21 is included in it. Similarly the last class is 47-49 which include the highest value 48.

The class interval of the class 20-22 is three because 20, 21 and 22 are included in this class. The value 20 is known as the lower limit of that class ad 22 as the upper limit. The midpoint of the class is 21 which is the average of the upper and lower limits of the class $[\frac{(20 + 22)}{2}]$

In this frequency distribution it is assumed that there is no scores between 22 and 23, 25 and 26 and so on. But if the variable measured is a continuous variable, one cannot take the classes like this and hence should convert the classes into actual classes. This is done by bridging the gap between the upper limit of a class and lower limit of the next class. For this 0.5 is added to the upper limit of the classes and 0.5 in subtracted from the lower limit of each class. The new frequency table of the above data will be,

Actual classes	**Frequency**
46.6 – 49.5	1
43.5 – 46.5	0
40.5 – 43.5	11
37.5 – 40.5	3
34.5 – 37.5	12
31.5 – 34.5	2
28.5 – 31.5	6
25.5 – 28.5	14
22.5 – 25.5	1

	50

In this frequency distribution, it is assumed that the lower limits are included in the class but the upper limits are excluded from that class. Thus, in the class 46.5 - 49.5, the value 46.5 is included in the class, but 49.5 is excluded from the class. The class interval will not change by this procedure. That is class interval ‘i’ of the class 46.6 – 49.5 = 49. 5 – 46.5

2.2 Diagrammatic and graphic Representation

Statistical data may be displayed pictorially through various types of diagrams, graphs and maps. Diagrams and graphs are convincing and appealing ways in which statistical data may be presented. These presentations will be more interesting to the common man than the frequency tables.

Diagrams and graphs are important because,

- They provide a bird's eye-view of the entire data and therefore the information is easily understood. Large number of figures may be confusing, but the pictorial presentation makes the data simple to understand and interesting to the readers.
- They are attractive to the eye. When figures are over looked by the common men, pictures create greater interest.
- They have great memorising effect. The impressions created by diagrams are long lasting than that made by data in tabular form.

- They facilitate comparison of data. Quick and accurate comparison of data is possible through diagrammatic presentation.
- They bring out hidden facts and relationship and help in analytical thinking and investigation.

2.3 Graphs of Frequency Distribution

A frequency distribution can be presented graphically in the following ways.

1. Histogram
2. Frequency polygon
3. Frequency curve
4. Ogives

2.3.1 Histogram

Histogram is a set of vertical bars whose areas are proportional to the frequencies represented. It is the most popular method of presenting a frequency distribution.

Construction of a histogram

While constructing a histogram, the variable is always taken on the X-axis and the frequencies on the Y-axis. Each class is marked on the X-axis by taking an appropriate scale to represent the class interval. Then rectangles are erected at each class with height as the frequency of that class. The area of each rectangle is proportional to the frequency of that class and the total area of the histogram is proportional to the total frequency.

[Note: While drawing graphs, it is conventional to take the scales on X-axis and Y-axis so that the height of the graph is 75% of its width.]

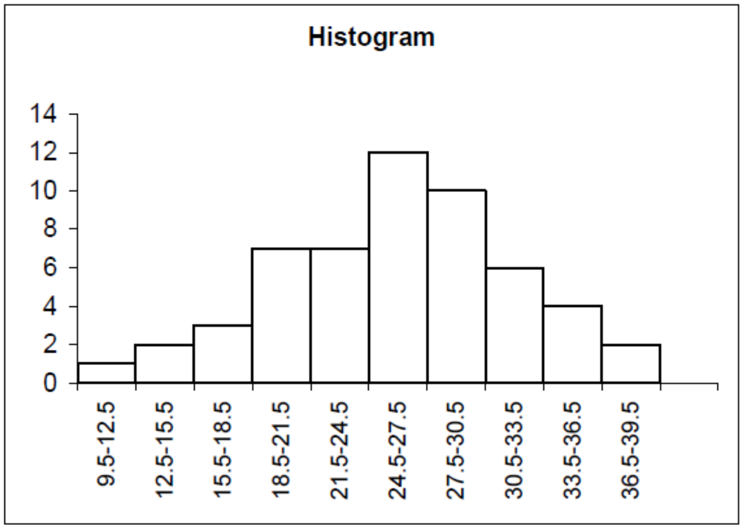
Illustration: Draw histogram for the following data.

Classes	**Frequency**
37-39	2
34-36	4
31-33	6
28-30	12
25-27	7
22-24	7
19-21	7
16-18	3
13-15	2
10-12	1

Here the classes are not continuous and hence have to be converted into actual classes. For this reduce 0.5 from each lower limit and add 0.5 to each upper limit.

The actual classes will be

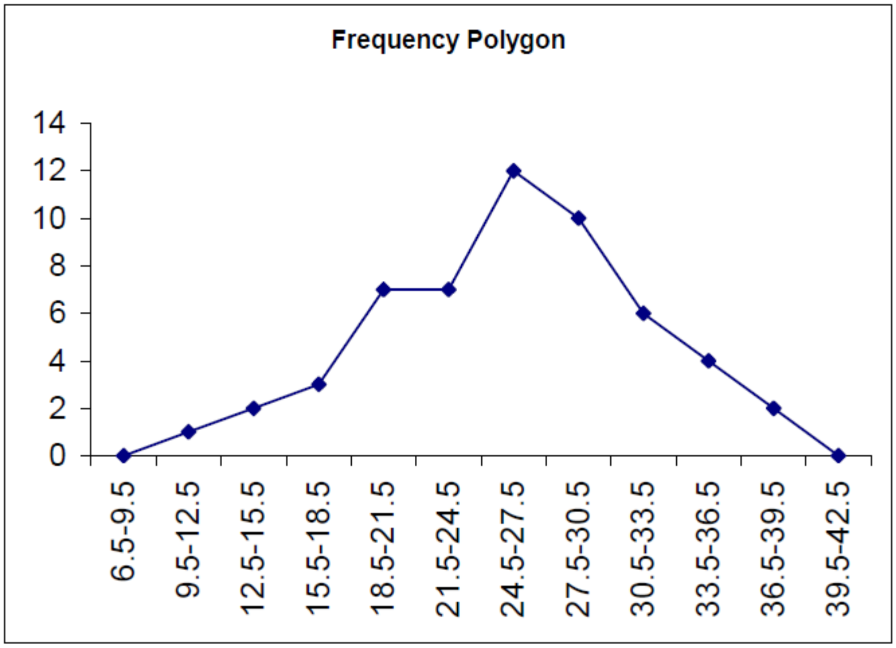
Classes	**Frequency**
36.5-39.5	2
33.5-36.5	4
30.5-33.5	6
27.5-30.5	10
24.5-27.5	12
21.5-24.5	7
18.5-21.5	7
15.5-18.5	3
12.5-15.5	2
9.5-12.5	1
	54



2.3.2 Frequency Polygon

A frequency polygon is a graph of frequency distribution. In a histogram, if the midpoints of the upper bases of the rectangles are connected by straight lines, we will get a frequency polygon. It is assumed that the area under the polygon is proportional to the total frequency. We can construct a frequency polygon directly from the frequency distribution without drawing the histogram. For this, as a first step calculate the midpoints of the classes and mark them on the Xaxis. Then plot the frequency corresponding to each point and join all these points by straight lines. Join the two ends to the X-axis, the left end to the mid point of the class before the first class and the right end to the midpoint of the class after the last class.

Illustration Classes	**Frequency**	**Mid Point= (Upper limit + Lower limit)/2**
36.5-39.5	2	38
33.5-36.5	4	35
30.5-33.5	6	32
27.5-30.5	10	29
24.5-27.5	12	26
21.5-24.5	7	23
18.5-21.5	7	20
15.5-18.5	3	17
12.5-15.5	2	14
9.5-12.5	1	11



Frequency polygon has some advantages over histogram. Some of them are

- More than one frequency distribution can be presented as frequency polygon in the same graph facilitating comparison. But histogram of different distributions are to be drawn in different graph papers.
- Frequency polygon is simpler than histogram.
- Frequency polygon gives a better idea about the nature of the distribution than histogram.

2.3.3 Frequency Curve

A frequency curve can be drawn by joining the points plotted for a frequency polygon by a smooth curve. The curve is drawn freehand so that the total area under the curve is approximately the same as that under the polygon.

The smoothed frequency curve can be drawn by calculating the smoothed frequency of each class and plotting the points based on the smoothed frequencies. Smoothed frequency of a class is calculated by adding the three consecutive frequencies and dividing by three. ie.,

$$\textit{Smoothed frequency} = \frac{\textit{Frequency of the given class} + \textit{frequenicies of the two adjacent classes}}{3}$$

2.3.4 Ogives (Cumulative Frequency Curve)

The curve obtained by plotting cumulative frequencies is called a cumulative frequency curve. There are two types of cumulative frequency -less than and greater than. To calculate the less than cumulative frequency, upper limits are taken into consideration and the number of scores less than the upper limit of each class is calculated.

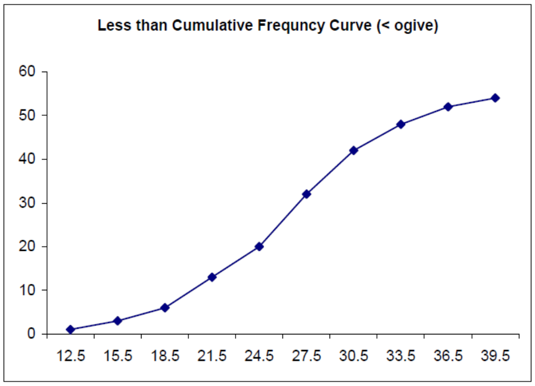
In the case of greater than cumulative frequency, number of scores greater than the lower limit of each class is calculated. The graph drawn with upper limits of the classes on X-axis and less than cumulative frequencies on Y-axis is called a less than cumulative frequency curve.

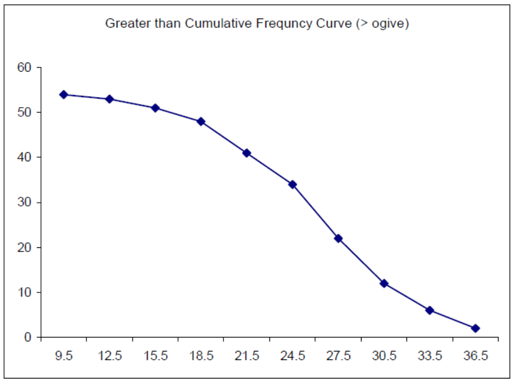
The graph drawn with lower limits of the classes on X-axis and greater than cumulative frequencies on Y-axis is called a greater than cumulative frequency curve.

These curves help the reader to determine the number of cases above or below a given value. They help to determine graphically values of median and quartiles.

Less than cumulative frequency curve and greater than cumulative frequency curve.

Upper Limit	<c.f(less than cumulative frequency)	Lower limits	>c.f (greater than cumulative frequency)
39.5	54	36.5	2
36.5	52	33.5	6
33.5	48	30.5	12
30.5	42	27.5	22
27.5	32	24.5	34
24.5	20	21.5	41
21.5	13	18.5	48
18.5	6	15.5	51
15.5	3	12.5	53
12.5	1	9.5	54





If the cumulative frequencies are converted into the corresponding percentages the respective curves will be called as less than Ogive (read an Ojive) and greater than ogive. Ogives help the investigator to determine percentiles and deciles directly from the graph.

These curves have much uses in research but they are not simple to interpret.

2.4 Limitations of Graphs

Diagrams and graphs are powerful mode of presenting frequency distribution, but they can not always substitute the tabular presentation. While selecting graph or diagram for presenting the data, utmost care must be taken to select the most appropriate one for the given purpose.

Some limitations of graphs and diagrams are

- They present only approximate values.
- They are representing only a limited amount of information.
- They can be easily misinterpreted.
- They are used for explaining quantitative data to the common man, for a statistician they are not much helpful for analysis of data.

UNIT III: DESCRIPTIVE STATISTICS:

3.1 Averages

Characteristics of a good measure of central tendency

Measure of central tendency is a single value representing a group of values and hence is supposed to have the following properties.

1. Easy to understand and simple to calculate: A good measure of central tendency must be easy to comprehend and the procedure involved in its calculation should be simple.

2. Based on all items: A good average should consider all items in the series.

3. Rigidly defined: A measure of central tendency must be clearly and properly defined. It will be better if it is algebraically defined so that personal bias can be avoided in its calculation.

4. Capable of further algebraic treatment: A good average should be usable for further calculations.

5. Not be unduly affected by extreme values: A good average should not be unduly affected by the extreme or extra ordinary values in a series.

6. Sampling stability: A good measure of central tendency should not be affected by sampling fluctuations. That is, it should be stable.

The most common measures of central tendency are

- Arithmetic mean
- Median and
- Mode

Each average has its own advantages and disadvantages while representing a series of number. The details of these averages are given below.

3.1.1 Arithmetic Mean

The most useful and popular measure of central tendency is the arithmetic mean. It is defined as the sum of all the items divided by the number of items. Mean is usually denoted as M or X.

Mean for raw data:

Let x1 x2, x3 ....xn be the n scores in a group. Then its Arithmetic mean X is calculated as

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum x}{n}$$

Where Σ *x* denote the sum of 'n' items.

For example if 10, 15, 16, 9, 8, 11, 12, 17, 18, 14 are the marks obtained by ten students in a unit test, its arithmetic mean will be

$$\overline{X} = \frac{10+15+16+9+8+11+12+17+18+14}{10} = \frac{130}{10} = 13$$

That is '13' is a single score that can be used to represent the given marks of 10 students.

Arithmetic mean for Grouped data.

If the data is given as a discrete frequency table, then Arithmetic mean

$$\overline{X} = \frac{\sum fx}{N}$$

where X - the score, f - frequency of that score and N - Total frequency.

Calculate A.M. for the following data.

Score	**Frequency**
14	2
21	4
23	3
28	4
35	2

Here the data reveals that 14 occur 2 times, 21 occur 4 times 23 occur 3 times and so on. When the formula for A.M is used

$$\overline{X} = \frac{14x2 + 21x4 + 23x3 + 28x4 + 35x2}{2+4+3+4+2} = \frac{363}{15} = 24.2$$

Again

x	**f**	**fx**
14	2	28
21	4	84
23	3	69
28	4	112
35	2	70
	-----	--------
N =	15	$\sum fx$ 363

$$\overline{X} = \frac{\sum fx}{N} = \frac{363}{15} = 24.2$$

A.M is the same when we use different methods to calculate the mean.

If the data is presented as a continuous series (classes and frequencies), then A.M is calculated by using the formula

$$\overline{X} = \frac{\sum fx}{N}$$

where

X - midpoint of a class

f - frequency of that class

N - Total frequency

Classes	**Frequency**
40-50	15
30-40	20
20-30	10
10-20	15
0-10	10
N =	**70**

To calculate mean, midpoint of each class is to be calculated. For this upper limit and lower limit of each class is added and divided by two.

Thus midpoints of the classes will be

$$X$$

$$\frac{40+50}{2} = 45$$

$$\frac{30+40}{2} = 35$$

$$\frac{20+30}{2} = 25$$

$$\frac{10+20}{2} = 15$$

$$\frac{0+10}{2} = 5$$

These midpoints are then multiplied by the corresponding frequencies. i.e, X x f

45 x 15 = 675

35 x 20 = 700

25 x 10 = 250

15 x 15 = 225

5 x 10 = 50

Now, sum of these Xf, ie Σ∫x , is computed and divide Σ∫x by the total frequency. Thus,

$$\overline{X} = \frac{1900}{70} = 27.14$$

Short-cut Method (Assumed mean Method)

To make the calculation more easy, we can assume a value as Assumed mean and using the following formula. Arithmetic mean can be calculated. $\overline{X} = \frac{A \div i\sum \int d}{N}$

where

A - Assumed mean

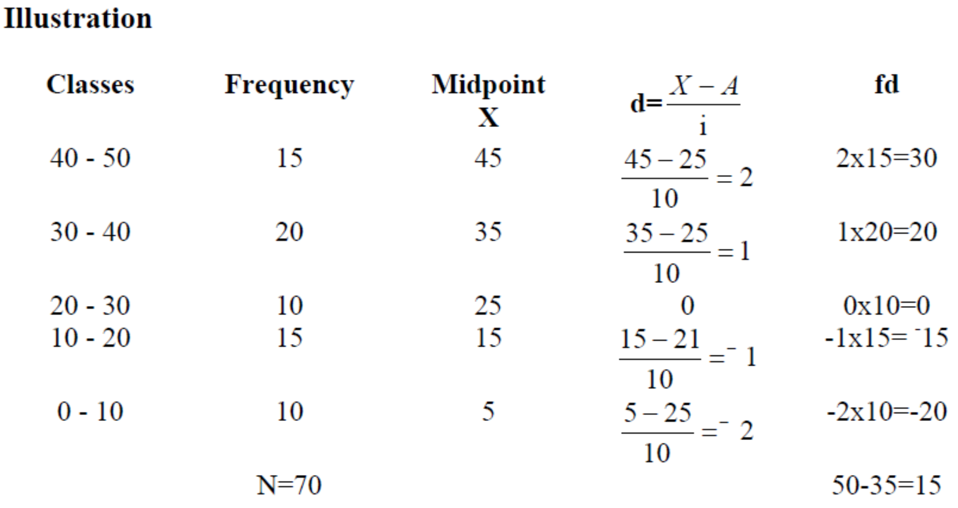
i - Class interval

f - Frequency of each class

d = (X – A) / i

X - Midpoint of the class

N - Total frequency



Illustration

Classes	Frequency	Midpoint X	$d = \frac{X - A}{i}$	fd
40 - 50	15	45	$\frac{45-25}{10} = 2$	2x15=30
30 - 40	20	35	$\frac{35-25}{10} = 1$	1x20=20
20 - 30	10	25	0	0x10=0
10 - 20	15	15	$\frac{15-21}{10} = ^{-}1$	-1x15= ⁻15
0 - 10	10	5	$\frac{5-25}{10} = ^{-}2$	-2x10=-20
	N=70			50-35=15

Take 25 as the assumed mean and calculate ‘d’ using the formula d = (X – A) / i

Multiply each ‘d’ with the corresponding frequency. Add these values (care should be taken as there will be positive and negative values. Add the numbers with same sign and subtract the smaller from the larger and put the sign of the larger number).

$$\text{A.M} \quad \overline{X} = A + \frac{i\sum fd}{N}$$

$$= 25 + \frac{10x15}{70}$$

$$= 25 + \frac{150}{70}$$

$$= 25+2.14$$
$$= 27.14$$

Merits

Arithmetic mean is the most widely used average**.** It has many advantages. Some of them are:

- It is simple to understand and easy to calculate
- It takes into account all the items of the series
- It is rigidly defined and is mathematical in nature
- It is relatively stable
- It is capable of further algebraic treatement
- Mean is the centre of gravity of the series, balancing the values on either side of it and hence is more typical

Demerits:

As mean considers all items of the series for its calculation, the value is unduly affected by extreme values (highest or least values). For example, consider the data 10, 30, 35, 36, 34. Here the mean is 145/5=29. a signle value 10, reduced the A.M to 29. similarly even a single higher value will increase the mean of the set of data. Also, when a single item is missing or if the classes are open ended, it is not possible to calculate mean. In distributions, which are highly deviating from normal distribution, mean will not be a suitable measure to represent the data.

3.1.2 Median

Median is the central value in a series, when the measures are arranged in the order of magnitude. One-half of the items in the distribution have value less than or equal to the median value and one-half have a value greater than or equal to median. That is median is the middle value in a distribution and it splits the set of values (observations) into two halves. Median is a positional average and not a value calculated from every items of the series. Median is the value so that equal number of items lie on either side of it.

When the number of observations is a series is odd, it will be easy to calculate median.Arrange the items in the order of magnitude, take the $[(n+1)/2]^{th}$ item in the series. It will be the median.

For example, consider the data

210, 121, 98, 81, 226, 260, 180, 167, 140, 138, 149.

Re arrange the values according to magnitude.

81, 98, 121, 138, 140, 149, 167, 180, 210, 226, 260.

As there are 11 observations take $[(11+1)/2]^{th}$ item ie 6^{th} item in the series which is 149. Median of the given set of measures is 149 and number of observations to the left of 149 is the same as that to the right.

But when the number of observations is even, median is the average of the two middle position values. That is, when 'n' is even, median is the average of $(n/2)^{th}$ and $[(n/2)+1]^{th}$ items after arranging in the order of magnitude.

Consider the values, 269, 247, 272, 282, 254, 266 when arranged in the order of magnitude,

247, 254, 266, 269, 272, 282 the middle values are $(n/2)^{th}$ and $[(n/2)+1]^{th}$ values i.e, 3^{rd} and 4^{th} values.

266 and 269 respectively median will be the average of 266 and 269

i.e, $(266+269)/2 = 267.5$

Median of a frequency distribution

Discrete series

For a discrete series, calculation of median involves the following steps.

Step 1: Arrange the values in the order of magnitude.

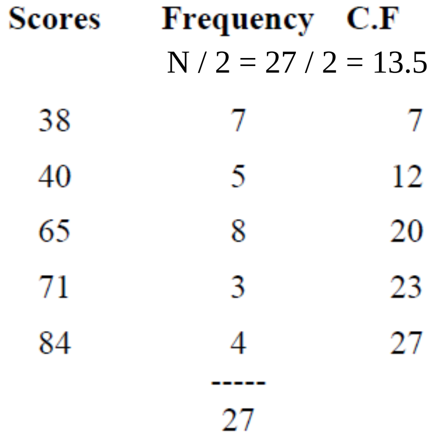
Step 2: Write the cumulative frequency, ie the number of observations less than that value.

Step 3: Find the (n / 2)th item in the series. This can be done by looking for the item with cumulative frequency equal to or greater than N / 2,

Score	**Frequency**
84	4
38	7
71	3
65	8
40	5

	27

Re arranging the series



Scores	**Frequency**	**C.F**
38	7	7
40	5	12
65	8	20
71	3	23
84	4	27

	27	

N / 2 = 27 / 2 = 13.5

'F' Take the 14th item in the series. When cumulative frequency is observed we can see that 14th item will be 65 as the c.f. of the previous item 40 is 12 which less than N/2.

∴ Median is 14th item = 65

Continuous series

For a continuous series median is calculated using the formula

$$\text{Median} = l + \frac{i(N - m)}{f}$$ where

l = lower limit of the median class

i = Class interval of the median class

N = Total frequency;

m = Cumulative frequency of the class below the median class

f - Frequency of the median class.

Median class is that class for which the cumulative frequency is equal to or greater than N / 2 for the first time in the distribution. It is the class in which median will fall. (The classes must be written as exact classes).

Classes	Frequency	Exact classes	c.f.
41-50	5	40.5-50.5	50
31-40	15	30.5-40.5	45
21-30	8	20.5-30.5	30
11-20	17	10.5-20.5	22
1-10	5	0.5-10.5	5

	50		

$$\frac{N}{2} = \frac{50}{2} = 25$$

The class with c.f. 25 or above is 20.5 - 30.5 (c.f = 30) and hence it is the median class. The class just before the median class is 10.5-20.5 with c.f. 22.

$$\therefore \text{ Median} = l + \frac{\frac{i(N-m)}{2}}{f} \qquad l = 20.5$$

$$i = 10$$

$$= 20.5+10\,\frac{(25-22)}{8} \qquad \frac{N}{2} = 25$$

$$= 20.5+10\text{x}\frac{3}{8} \qquad m = 22$$

$$= 20.5+\frac{308}{} \qquad F = 8$$

$$= 20.5 + 3.75$$

$$= 24.25$$

Merits

- Median is not affected by extreme values. Median of the data 10, 11, 12, 13, 14 is 12 and that of 0, 10, 12, 14, 100 is also 12. Hence if we know that the distribution contains extreme values, median will be more representative than mean.
- In open ended classes and if the data is incomplete, if the relative position of the missing data is known, median can be calculated.
- Median can be calculated graphically (by drawing the ogives and taking the x-co-ordinate of the meeting points of the two curves.
- Median is simple to understand and easy to calculate.

Demerits

- Calculation of median need re arrangement of data and hence will be difficult if number of observations is large.
- It is only a [positional average and is not based on all other items.
- It can not be used for further algebraic treatment.
- Less stable than arithmetic mean.
- Median is not mathematically defined or rigidly defined. As the number of observations becomes odd or even, the position of median changes.

3.1.3 Mode

The mode is that value in a series of observations which occurs with the greatest frequency.

Example

5, 3 8, 5, 13, 9, 5, 11, 5, 8, 10, 8, 5, 6, 5

In this set of observations, 5 is repeated six times and is the most frequent item in the series taken other values. Hence mode in this case is 5.

Mode is the value which occurs most often in the data. It is the value at the point around which the items tend to be most heavily concentrated.

In a raw data, mode can be found out by counting the number of times the various values repeat themselves and finding the value occuring maximum number of times. If there are two or more values with the highest frequency, mode is ill-defined or all the values with highest frequency are modal values. Then the distribution is bi-modal or multimodal.

If all the items of a series are not repeating or repeating the same time, one can say that there is no mode for the distribution.

For grouped data

In a discrete series, mode is the item with highest frequency.

For a continuous series, mode can be calculated by the formula.

$$\text{Mode} = L + \frac{i\,f_2}{f_1 + f_2} \text{ where}$$

L = Exact lower limit of the modal class (the class with highest frequency).

i - Class interval

f1 - Frequency of the class just below the modal class (preceding the modal class)

f2 - Frequency of the class just above the modal class (succeeding the modal class)

Empirical formula

Mode can be estimated from the values of mean and median using the formula.

Mode = 3 Median - 2 Mean

Illustration

Raw data 9, 8, 6, 11, 12, 9, 9, 8, 6, 13 here 9 repeats three times and no other value repeat that much. That is 9 is the most frequent item in the series. Hence mode of these observations is 9.

If the series is

9, 8, 6, 11, 6, 9, 9, 8, 6, 13, 6 and 9 are occuring most frequently and there are two modes 6 and 9.

If the values are 9, 8, 6, 11, 13 there is no mode.

Discrete series

Marks :	10	15	20	25	30	35
Frequency:	8	12	30	27	18	9

In this frequency distribution, 20 is the mark with maximum frequency and therefore mode is 20.

Continuous series

Class	**Frequency**	**Exact class**
41 - 50	5	40.5 - 50.5
31 - 40	15	30.5 - 40.5
21 - 30	8	20.5 - 30.5
11 - 20	17	10.5 - 20.5
1 - 10	5	0.5 - 10.5

10.5 - 20.5 is the modal class as the frequency is maximum for that class.

$$\text{Mode} = L + \frac{i f_2}{f_1 + f_2} \text{ where}$$

l = 10.5, f1 - frequency of the preceding class ie frequency of the class 0.5 - 10.5 = 5

f2 = frequency of the succeeding class

= frequency of the class 20.5 - 30.5

= 8

i = 10

$$\therefore \text{ Mode} = 10.5 + \frac{10 \times 8}{5+8} = 10.5 + \frac{80}{13} = 10.5 + 6.15$$

$$= 16.65$$

Supposed in a frequency distribution Mean = 44.8 and Median = 44. Then mode can be taken as

Mode = 3 medium - 2 mean

= 3 x 44 - 2 x 44.8

= 132 - 89.6 = 42.4

Merits

Mode is the simplest measure of central tendency.

- It gives quickest measure of central tendency. Therefore when a quickest, but approximate, value to represent a group of observations is needed, mode can be used.
- Mode is not affected by extreme values.

- In open ended classes, mode can be calculated.
- Even qualitative data can be described through mode. (When we say the consumer preferences of a product, modal value is used instead of median or mean which are not even meaningful).

Demerits

- Mode is not rigidly defined. In all cases we can not calculate a unique mode. It may be bimodal or multimodal. That is mode is ill-defined.
- It is not capable of further algebraic calculations.
- It is not based on all items of a series.
- Mode is less used in quantitative data as mean and median are more representative of the distribution.

When to use Mean, Median and Mode:

Arithmetic mean is the most reliable and accurate measure of central tendency. It is more stable than median or mode and is less affected by sampling fluctuations. When we need a reliable, more accurate measure to represent the data, mean can be used. If we want to compute more statistics like standard deviation, correlation etc, Mean is recommended.

But when there are extreme values in the set of data, mean will not be a true representation. If extreme values exist, median will give more representation of data than mean.

If we want to find the middle most value in the series, median is calculated. If the classes are open-ended, or some values are missing, but their relative position is known, mean cannot be calculated, where median becomes the most reliable measure of central tendency.

When a crude (rough) measure of central tendency is needed or if we want to know the most often recurring value, mode is calculated. Mode can be easily obtained from the graphs like histogram, frequency polygon or frequency curve.

It should be borne in mind that these values, mean, median or mode are values representing a group of values. That is average or measure of central tendency is a sing value representing a group of values and hence it must be properly interpreted, otherwise will arrive at wrong decisions. It lies between the lowest and highest values in the series. Some times the average need not be a value in the series. This may also lead the individual to wrong interpretation. For example, the average size of a family is 4.6 is absurd as a family cannot have a size of 4.6.

Two or more set of values may have the same measure of central tendency but differ in their nature. Therefore while comparing distributions measures of central tendency will not give complete picture of the distributions.

Also, if the data is not having a clear single concentration, of observations, an average will not be meaningful.

3.2 Measure of dispersion

Measure of dispersion can be defined as the degree to which numerical data tend to spread about an average value. More clearly, dispersion measures the extent to which the items vary from some central value.

The dispersion or scatter or variability of a set of values can be calculated mainly through four measures

1. Range
2. Quartile deviation
3. Mean deviation and
4. Standard deviation

These four measures give idea about the variability or dispersion of the values in a set of data.

3.2.1. Range:

Range is the simplest measure of dispersion. It is the difference between the smallest and largest items of a distribution.

ie Range = H - L where H - the highest measure in the distribution and L - lowest measure in the distribution.

It is a very rough measure of dispersion, considering only the extreme values.

Merits

Range is simple to understand and easy to computer. When a quick rather than a very accurate picture of variability is needed, range may be computed.

Demerits

It is not based on each item of the series. Range is affected by sampling fluctuation. Range does not give any idea about the features of the distribution in between the extreme values. In open ended distribution range cannot be calculated. (In a frequency distribution range is the difference between the upper limit of the highest class and the lower class).

3.2.2. Quartile Deviation (Semi inter quartile range)

Quartile Deviation Q.D. = $(Q_3 – Q_1) / 2$

where Q1 and Q3 are the 1st and 3rd quartiles of the distribution. The first quartile Q1 is that value below which 25% of the distribution fall. The third quartile Q3 in the point below which 75% cases lie.

$$Q_1 = l_1 + \frac{i\left(\frac{N}{4} - m_1\right)}{f_1}$$

where l1 – lower limit of the first quartile class.

i – class interval

N – total frequency

m1 – c.f upto the first quartile class.

f1 – frequency of the first quartile class

$$Q_3 = l_3 + \frac{i\left(\frac{3N}{4} - m_1\right)}{f_3}$$

m3 – c.f upto the third quartile class.

f3 - frequency of the third quartile class.

First quartile class in the class with c.f. greater than or equal to N / 4 th item.

Third quartile class in the class with C.f. greater than or equal to 3N / 4 th item.

For raw data

20, 18, 16, 25, 28, 32, 15

Arranging according to magnitude, the series becomes.

15, 16, 18, 20, 25, 28, 30, 32

$\frac{N}{4}$ th item = $\frac{8}{4}$ 2^{nd} item = 16. $\therefore$ $Q_1 = 16.$

$\left(3\frac{N}{4}\right)^{th}$ item = 3x2 = 6^{th} item = 28. $\therefore$ $Q_3 = 28.$

$$\text{Q.D} = \frac{Q_3 - Q_1}{2} = \frac{28-16}{2} = \frac{12}{2} = 6$$

For discrete Series

Marks :	10	20	30	40	50	60
Frequency:	4	7	8	10	6	5
c.f. :	4	11	19	29	35	40

$$Q_1 = \left(\frac{N}{4}\right)^{th}_{item} = \frac{40}{40} = 10^{th}\,\text{item} \therefore Q_1 = 20$$

$$Q_3 = \left(\frac{3N}{4}\right)^{th}_{item} = 3x10 = 30^{th}\,\text{item} \therefore Q_3 = 50$$

$$\therefore \text{Q.D} = \frac{Q_3 - Q_1}{2} = \frac{50-20}{2} = \frac{30}{2} = 15$$

For continuous frequency distribution

Class	**Frequency**	**c.f**
40 - 5 - 50.5	5	50
30.5 - 40.5	15	45
20.5 - 30.5	8	30
10.5 - 20.5	17	22
05 - 10.5	5	5

	50	
	= = =	

Q_1 class is the class with c.f. greater than or equal to $\frac{N}{4}\left(\frac{50}{4}=12.5).\right)$ Q_1 class is 10.5 - 20.5.

III^{rly} Q_3 class - the class with c.f. $\geq \frac{3N}{4}(=37.5)$ is 30.5-40.5

$$\therefore\ Q_1 = l_1 + \frac{i\left(\frac{N}{4}-m1\right)}{\underline{f}}$$

$$= 10.5 + 10\ \frac{(12.5-5)}{17} = 10.5 + \frac{10x7.5}{17}$$

$$= \ 10.5 + \frac{75}{17} = 10.5 + 4.4 = 14.9$$

$$Q_3 = l_3 + \ i\left(\frac{3N}{4} - m_3\right)/f_3$$

$$= 30.5 + 10\frac{(37.5-30)}{15}$$

$$= \ 30.5 + \frac{10x7.5}{15}$$

$$= \ 30.5 + \frac{75}{15} = 30.5 + 5 = 35.5$$

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

$$= \ \frac{35.5-14.9}{2} = \frac{20.6}{2} = 10.3$$

Merits

1. It is superior to range as a measure of dispersion. Range considers the extreme values whereas Q.D considers the range of middle 50% of cases.
2. In open ended class Q.D can be computed.
3. Q.D. is not affected by extreme values.

Demerits

1. It is not capable of further mathematical calculation.
2. It is not based on all observations.
3. It is affected by sampling fluctuations.

3.2.3. Mean Deviation

Mean deviation can be defined as the mean of deviations of all the separate scores in the series taken from their mean. M.D is the simplest measure that rally takes into account the variation of observations from an average (measure of central tendency).

Mean deviation is also termed an average deviation and can be used for finding variation with respect to median or more instead of mean. (The sum of deviations of observations from their Median in the minimum when signs are ignored).

Mean deviation for Raw data

If x1, x2, x3… xn are in observations of a series, Mean Deviation.

$$\text{M.D} = \frac{1}{n}\sum |X - \overline{X}|$$

$$\text{Or} = \frac{1}{n}\sum |D| = \frac{\sum |D|}{n}$$

Where $|D| = |X - \overline{X}|$, $\overline{X}$ = A.M. of the series.

Note: $|x| = x$ if x is positive

- x if x is negative For example |2|=2, | -2| = 2.

(If mean deviation is small, the distribution is highly compact or uniform).

If 15, 21, 26, 13, 14, 18, 28, 25, 12

$$\text{Mean} = \frac{15+21+26+13+14+18+25+12}{8} = \frac{144}{8} = 18$$

| D | : |15-18|, |21-18|, |26-18|, |13-18|, |14-18|, |18-18|. |25-18|, |12-18|

$$\sum |D| = 3+3+8+5+4+0+7+6 = 36$$

$$\therefore \text{MD} = \frac{\sum |D|}{n} = \frac{36}{8} = 4.5$$

Discrete Series

In the case of a frequency distribution M.D can be calculated by using the formal.

$$\text{M.D} = \frac{\sum f\left|\text{X} - \overline{\text{X}}\right.}{\text{N}} \text{ or } = \frac{\sum f|\text{D}|}{\text{N}}$$

Where , $|\text{D}| = |\text{X} - \overline{\text{X}}|$ f – frequency.

X:	10	20	30	40	50	60
F:	4	7	8	10	6	5

$$\text{X} = \frac{\sum fX}{N} = \frac{10x4 + 20x7 + 30x8 + 40x10 + 50x6 + 60x5}{4 + 7 + 8 + 10 + 6 + 5}$$

$$= \frac{40 + 140 + 240 + 400 + 300 + 300}{40}$$

$$= \frac{1420}{40} = 35.5$$

$$\left|\text{X} - \overline{\text{X}}\right|: 10 - 35.5 |,| 20 - 35.5 |,| 30 - 35.5 |,| 40 - 35.5 | .150 - 35.5 |,| 60 - 35.5 \|$$

$f \mid D \mid$:: 25.5x4+5.5x7+5.5x8+4.5x10+14.5x6+24.5x5

$$\sum f|D| : 102 + 108.5 + 44 + 45 + 87 + 122.5 = 631.5$$

$$\text{M.D} = \frac{631.5}{40} = 15.79$$

For a Continuous Series

$$\text{M.D} = \frac{\sum f \mid D \mid}{N}$$ Where $|\text{D}| = |\text{X} - \overline{\text{X}}|$, X -midpoint of the class.

Classes	**Frequency**	**Midpoint (X)**	**d= $\frac{\text{X-A}}{\text{i}}$**	**fd**
40.5-50.5	5	45.5	2	10
30.5-40.5	15	35.5	1	15
20.5-30.5	8	25.5	0	0
10.5-20.5	17	15.5	-1	-17
0.5-10.5	5	5.5	-2	-10
	---- 50			----- -2

Assumed Mean = 25.5

$$\text{Mean } \mathrm{X} = \mathrm{A} + \mathrm{i}\,\frac{\sum \mathrm{fd}}{\mathrm{N}} = 25.5 + 10\mathrm{x}\frac{-2}{50}$$

$$= 25.5 - \frac{2}{5} = 25.5 - 0.4$$

$$= 25.1 \quad —$$

$\lvert\mathbf{D}\rvert = \lvert \mathrm{X} - \overline{\mathrm{X}} \rvert$	$\mathbf{f}\,\lvert\mathbf{D}\rvert$
20.4	102
10.4	156
0.4	3.2
9.6	163.2
19.6	-------
	522.4

$$\mathrm{M.D} = \frac{\sum f \mid D \mid}{N} = \frac{522.4}{50} = 10.45$$

Merits

1. It is simple to understand and easy to calculate.
2. It is based on all observations.
3. It is rigidly defined.
4. It is not unduly affected by extreme values.
5. It is statistically stable.

Demerits

Mean deviation is a non-algebraic measure as the formula includes the absolute deviations. It cannot be used for further algebraic treatment.

Mean deviation is used as a measure of dispersion in small samples and the results are presented before the public with less statistical background. But for higher statistical purposes, mean deviation is no recommended as a measure of dispersion.

3.2.4. Standard Deviation

Karl Pearson introduced the concept of standard deviation in 1823. It is defined as the square root of the mean of the squared deviations from the arithmetic mean. Standard deviation is usually represented by the small Greek letter ‘σ’ (sigma). (Σ-the symbol to denote ‘sum’ is the capital Greek letters sigma).

$$\text{Thus S. D } \sigma = \sqrt{\frac{\sum (x-\overline{x})^2}{n}}$$

where X – individual score

$\overline{X}$ = arithmetic mean

n = number of observation.

For the raw data

15, 21, 26, 13, 14, 18, 25, 12

Mean = 18

Sum of squares of deviation from

$$\sum (X-\overline{X})^2 = (-3)^2 + 3^2 + 8^2 + (-5)^2 + (-4)^2 + 0 + 7^2 + (-6)^2$$

$$= 9+9+64+25+16+49+36$$

$$= 208$$

$$\therefore \sigma = \sqrt{\frac{208}{8}} = \sqrt{26} = 5.10$$

In a discrete series

$$\sigma = \sqrt{\frac{\sum f(X-\overline{X})^2}{N}}$$

Marks:		10	20	30	40	50	60
Frequency:	4	7	8	10	6	5	

$$\overline{X} = 35.5$$

$$(X-\overline{X})^2 = (-25.5)^2 \quad (15.5)^2 \quad (-5.5)^2 \quad (4.5)^2 \quad (14.5)^2 \quad (24.5)^2$$

$$: 650.2; \quad 240.25, 30.25 \quad 20.15, \quad 210.25, 60.25$$

$$\sum f\left(X-\overline{X}\right)^2 : 4x650.25+7x240.25+8x30.25+10x20.5+6x210.25+5x60.25.$$

$$= 2601+1681.75+242+205+1261+201.25$$

$$= 6292.5$$

$$\sigma \quad = \sqrt{\frac{\sum f\left(X-\overline{X}\right)^2}{N}} = \sqrt{\frac{6292.5}{40}} = \sqrt{157.3}$$

$$= 12.54$$

For a continuous series

$$\sigma = \sqrt{\frac{\sum f\left(X\overline{X}\right)^2}{N}} \text{ or } \sigma = \frac{i}{N}\sqrt{N\sum fd^2\left(\sum fd\right)^2}$$

Where $d = \dfrac{X-A}{i}$ A – assumed mean.

(Shortcut method reduces the complexity of Calculation)

Classes	Frequency	X	$d = \dfrac{X-A}{i}$	d^2	fd	fd^2
40.5-50.5	5	45.5	2	4	10	20
30.5-40.5	15	35.5	1	1	15	15
20.5-30.5	8	25.5	0	0	0	0
10.5-20.5	17	15.5	-1	1	-17	17
0.5-10.5	5	5.5	-2	4	-10	20
	------				----	-----
	50				-2	72

Take assumed mean A = 25.5

$$\sigma = \frac{i}{N}\sqrt{N\sum fd^2 - \left(\sum fd\right)^2}$$

$$= \frac{10}{50}\sqrt{50x72-(-2)^2}$$

$$= \frac{10}{50}\sqrt{3600-4}$$

$$= \frac{1}{5}\sqrt{3596} \qquad = \frac{1}{5}x59.96$$

$$= 11.99$$

Merits

SD is well defined and is based on all observation. It is less influenced by sampling fluctuation. The only measure of dispersion that can be used for further mathematical treatment is standard deviation. It is the most stable, reliable measure of dispersion.

Demerits

It is not easy to compute or simple to understand.

Variance

Variance is the square of standard deviation. That is variance is the average of each score's squared difference form mean.

If one distribution is more spread out than another, its variance will be larger as the deviation scores will be larger.

Variance = σ^2

$$= \frac{\sum \left(x - \overline{x}\right)^2}{n}$$

(for law data)

$$\text{and variance} = \frac{\sum f\left(x - \overline{x}\right)^2}{N}$$

(for frequency distribution)

Coefficient of Variation:

The relative measure of Variation based on standard deviation is coefficient of variation (developed by Karl Pearson).

Coefficient of Variation

$$C.V = \frac{\sigma}{\overline{X}} x100$$

While comparing the variance of two or more series C.V is preferred. If CV is greater for a distribution it is less uniform or stable than the other.

UNIT IV: SAMPLING, PROBABILITY AND THEORETICAL DISTRIBUTIONS

4.1 Concept of population and sample

The 'population' is all the people about whom a researcher wishes to make a statement. (The phrase 'target population' is sometimes used). The hypothesis is a statement about the population not the sample. A sample is defined as a relatively small set (subset) of individuals, groups, objects, items and events selected from the respective population.

The aim is to select a representative sample from the target population—a small group who represent the target population in terms of characteristics such as age, IQ, social class, relevant experiences and so on. The importance of representativeness is required to be able to generalize from the sample to the target population. A sample that is not representative is described as biased, i.e., leading in one direction. A biased sample means that any generalization lack external validity.

Population: A population is defined as the total and complete set of individuals, groups, objects, items and events that the researcher is interested in and is studying them. A population refers to all the data that a researcher is interested in.

For example, if we were to study the employment patterns of college graduates from India, our population would include every person who graduated from colleges across India. A population is the full set of all the possible units of analysis. The population is also sometimes called the universe of observations. The population is defined by the researcher, and it determines the limits of statistical generalization.

Suppose you wish to study the impact of corporate image advertising in large corporations. You might define the unit of analysis as the corporation, and the population as "Fortune 500 Corporations" (a listing of the 500 largest corporations in the United States compiled by Fortune magazine).

Any generalizations from your observations would be valid for these 500 corporations, but would not necessarily apply to any other corporations. Alternatively, you might define the universe as "Fortune 1000 Corporations." This includes the first 500, but expands the universe, and your generalizations, by adding an additional 500 corporations that are somewhat smaller. When all the members of the population are explicitly identified, the resulting list is called a sampling frame. The sampling frame is a document that can be used with the different selection procedures described below to create a subset of the population for study. This subset is the sample.

Census: If we actually measure the amount of advertising for each of 1000 corporations, we will be conducting a census of the variable. In a census, any statements about the variable (advertising column inches, in this case) are absolutely correct, assuming that the measurement of the variable is not in error. Suppose we conduct a census of image advertising done by Fortune 1000 corporations in two different years. To summarize the advertising done in each year, we calculate the average number of column inches of advertising done each year (by adding together all 1000 measurements for a year together, and dividing the sum by 1000). If this figure is 123.45 column inches for the first year and 122.22 column inches for the second year, we can say, with perfect confidence, that there was less image advertising done in the second year. The difference may be small, but it is trustworthy because all units (corporations) in the population are actually observed. Thus, when we examine every member of a population, difference we observe is a real one (although it may be of trivial size).

Sample: A sample is defined as a relatively small set (subset) of individuals, groups, objects, items and events selected from the respective population.

For example, instead of studying every college graduate across India is difficult and not feasible, which would be a very hectic thing and would cost a lot of time and money, we could select a sample of recent college graduates (graduated in the last year or so) and then we can study them and then using appropriate methods can generalize the findings to the respective larger population.

The aim of all psychological research is to be able to make valid generalizations about behavior. In a research project only a small number of participants are studied because you could never study the whole population. Psychologists use sampling techniques which maximize generalizability. Sampling is not limited to participants. Sampling techniques are also used in observational studies where term time sampling and event sampling are used.

4.1.1 Population Parameters And Sample Statistics

A parameter is a numerical characteristics of population, but a statistic is numerical characteristics of a sample. Statisticians use Greek letter to symbolize population parameters and English letter to symbolize sample statistics. Some of these symbols are illustrated below.

Name	**Sample Statistic**	**Population Parameter**
Mean		μ (mu)
Variance	SD^2	σ^2 (sigma squared)
Standard deviation	SD	σ (sigma)
Correlation	r	ρ (rho)
Proportion	p	π (pi)
Regression coefficient	b	β (beta)

4.2 Sampling distribution

A sampling distribution is the most important concept to be understood for initiating in the arena of inferential statistics. The concept of the sampling distribution makes it possible to form the probability statements in inferential statistics. One can define a sampling distribution as: a theoretical frequency distribution of the values of the chosen statistic, when random samples of the same size are drawn from a population. When the statistic drawn from such repeated random samples from a population are plotted with their relative frequency of occurrence then the resulting distribution is called a sampling distribution of that statistic.

The most popular and applied sampling distribution is the sampling distribution of the means. The process of building a sampling distribution of means involves picking up random samples of same size repeatedly from the mother population. For each of the chosen samples the mean is calculated. Once these *n* number of means are obtained then they are plotted for their relative frequency of occurrence. The resulting graph is the depiction of the sampling distribution of means.

The sampling distribution of means is normally distributed, with the mean value equivalent to the population mean. The spread or the variance is different from the population.

The normal curve shape of the sampling distribution of means owes to the central limit theorem: "*the random sampling distribution of the means tends towards a normal distribution irrespective of the shape of the population of the observations sampled; the approximation to the normal distribution improves as the sample size increases.*"

4.2.1 Random & Non Random sampling methods

Sampling is a process of selecting samples from a group or population to become the foundation for estimating and predicting the outcome of the population as well as to detect the unknown piece of information. A sample is the sub-unit of the population involved in your research work. There are a few advantages and disadvantages associated with the sampling process.

Advantages of Sampling: sampling can save cost and human resources during the process of research work. In ICT, sampling does not cause much constraint such as heavy use of tools and technology in predicting the research output.

Disadvantages of Sampling: A researcher may not find the information about the population being studied especially on its characteristics. The research can only estimate or predict them. This means that there is a high possibility of error occurence in the estimation made. Sampling process only enables a researcher to make estimation about the actual situation instead of finding the real truth. If you take a piece of information from your sampling population, and if your reasoning is correct, your findings should also be accurate to a certain degree.

When selecting a sample, it is very important for a researcher to consider the possibility of error during the selection process. In the field of ICT, sampling has little significance because the main purpose of ICT research is to explore or describe diversity in technology, phenomenon and issues.

Another factor is the nature of ICT research which focuses on qualitative approach. Qualitative approach does not make an attempt to quantify or determine the extent of diversity. A researcher can select a sample and describe his/ her inquiry based on the research problem. Then, the study proceeds based upon the obtained sample.

You must always remember that qualitative research has a characteristic called saturation point. Saturation point is where a researcher reaches the limit of obtaining information after many attempts to get new information. When you find you are not obtaining new information, it is assumed you have reached the saturation point. Again, saturation point is subjective judgement which a researcher always decide about it in the entire research process.

4.2.3 Sampling Terminologies

The community, families living in the town with smart homes form the population or study population and are usually denoted by the letter ***N***.

The sample group of elderly people or senior citizens and disable people in the vicinity of the smart home community is called ***sample***.

The number of elderly people or senior citizens and disabled people you obtain information to find their average age is called the ***sample size*** and is usually denoted by letter ***n***.

The way you select senior citizens and disabled people is called the ***sampling design*** or ***strategy***.

Each citizen or disabled people that becomes the basis for selecting your sample is called the ***sampling unit*** or ***sampling element***.

A list identifying each respondent in the study population is called ***sampling frame***. In case when all elements in a sampling population cannot be individually identified, you cannot have a sampling frame for the study population.

Finally, the obtained findings based on the information of the respondents are called ***sample statistics***.

4.2.4 Sample Size and Selection

Most of the new researchers always wonder about the sample size that needs to be selected. You must remember that the larger the sample for your research, the better outcome you can evaluate at the end of the research process. The larger the sample, the more likely the sample mean and standard deviation will become a representation of the population mean and standard deviation. For instance, in IT survey, the sample size required depends on the statistical outcome needed for the findings. The following are some guidelines to decide on how large a sample should be:

- When the selected sample needs to be segregated into smaller clusters involving comparisons of clusters, a large sample would be appropriate.
- The longer the duration of a study, the higher the number of subjects that will drop out. To reduce attrition, a researcher should keep demands on subjects to the minimum, to fully inform the subject about the study and research, and make frequent communication with subjects to maintain the interest.
- A larger sample is needed when the population is highly heterogeneous on the variables being studied so that different characteristics can be identified. If members of the population is less, then a small sample size would do to obtain the necessary characteristics.

Selecting a Sample

The objective of selecting a sample is to achieve maximum accuracy in your estimation within a given sample size and to avoid bias in the selection of the sample. This is important as bias can attack the integrity of facts and jeopardize your research outcome.

There are also factors that may influence the degree of certainty in inferences drawn from a sample for research study. As we know, the size of samples influence findings such that large samples have more certainty than those based on smaller ones. Therefore, the larger the sample, the researcher will obtain more accurate findings.

Another factor is the extent of variation in the sampling population where the greater the variation in the population will have greater uncertainty with respect to its characteristics. Therefore, it is crucial for a researcher to bear these in mind especially when selecting a sample for her/his respective research work.

4.3 sampling and non-sampling errors

The standard deviation of the sampling distribution is referred to as the standard error. The standard error can be described as a special kind of standard deviation and is derived from the sample standard deviation only with the formula: $\boldsymbol{\sigma_x = \sigma_x / \sqrt{n}}$**.** The smaller the standard error lesser is the variability in a sampling distribution.

4.4 Sampling Techniques

Sampling techniques often depend on research objectives of a research work. Generally there are two types of sampling techniques that are widely deployed. These techniques are:

4.4.1 Probability Sampling:

This sampling technique includes sample selection which is based on random methods. The techniques that are based in this category are random sampling, stratified sampling, systematic sampling and cluster sampling.

i. *Random Sampling:*

Random sampling is used to increase the probability of the sample selected. By deploying this technique, each member of a population stands a chance to be selected. Let’s say you are interested to survey the usage of ecommerce application in business-to-consumer.

The sample you select needs to represent the types of e-commerce application and its usage. Due to financial and time constraints you are unable to survey the usage of all types of e-commerce application across the Indian network (N= 100,000). Therefore you decide to confine the study to e-commerce application for merchandise products in India (n=10,000) which is called the accessible population.

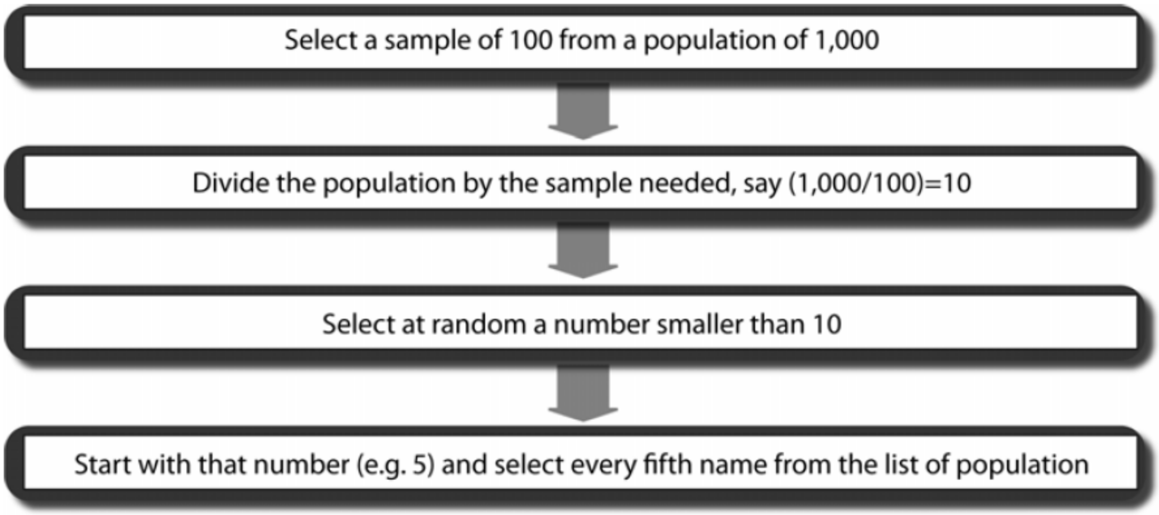
From this accessible population, a sample of 100 e-commerce application is retrieved. How do we randomly select sample? It is understood that random sample is a procedure in which all individuals in the defined population have an equal and independent chance to be selected in the sample design. In the above example, the number of e-commerce application on merchandise products across India network is 10,000 and you may intend to draw a sample of 100. When you select the first application, it has 1:10,000 chances of being selected. Once the first application selected, the remaining will be 9,999 so that each application has 1:9,999 of being selected as second case. Therefore, once each case is selected, the probability of being selected next changes because the population of selection has become one case smaller each time.

ii. Stratified Sampling:

In some IT surveys, a researcher may want to ensure individuals with certain characteristics are included in the sample to be studied. For such cases, stratified sampling is used. In this sampling design, a researcher will attempt to stratify population in such a way that the population within a stratum is homogeneous with respect to the characteristics on the basis of which it is being stratified. You must bear in mind that it is important for the characteristics chosen as the basis of stratification, are clearly identifiable in the population. For example, it is much easier to stratify the population on the basis of gender rather than age or income group.

iii. Systematic Sampling:

Systematic sampling also known as Âmixed samplingÊ category since it has both random and non-random sampling designs. A researcher has to begin by having a list names of members in the population, in random approach. Figure below shows the order of the sampling.


Select a sample of 100 from a population of 1,000
Divide the population by the sample needed, say (1,000/100)=10
Select at random a number smaller than 10
Start with that number (e.g. 5) and select every fifth name from the list of population

This sampling method is good as long as the list does not contain any hidden order. Systematic sampling is frequently used in ICT research and survey, especially in selecting specified number of records from computer documents.

iv. Cluster Sampling:

In cluster sampling, the unit of sampling is not referring to an individual entity but rather a group of entities. For example, in an organisation there are 25 departments and in each department there are an estimated 20 IT administrators. You need a sample of about 100 staff but this would mean going to many departments if random sampling approach is used. Using cluster sampling, you may select 5 departments randomly from a total of 25 departments. You study all the staff in the 5 departments you chose. The advantage that can be highlighted here is: it saves cost and time especially if the population is scattered. The disadvantage is that it is less accurate compared to other techniques of sampling discussed.

4.4.2 Non-Probability Sampling

In some research scenarios, it is not possible to ensure that the sample will be selected based on random selection. Non-probability sampling is based on a researcher’s judgement and there is possibility of bias in sample selection and distort findings of the study. Nonetheless, this sampling technique is used because of its practicality. It can save time and cost, and at the same time, it is a feasible method given the spread and features of a population. Some common sampling methods are quota sampling, purposive sampling and convenience sampling.

i. *Quota Sampling:*

The main reason directing quota sampling is the researcherÊs ease of access to the sample population. Similar to stratified sampling, a researcher needs to identify the subgroups and their proportions as they are represented in the population. Then, the researcher will select subjects based on his/ her convenience and judgement to fill each subgroup. A researcher must be confident in using this method and firmly state the criteria for selection of sample especially during results summarisation.

ii. *Purposive Sampling:*

This sampling method is selected on the basis that members conform to certain stipulated criteria. You may need to use your own judgement to select cases to answer certain research questions. This sampling method is normally deployed if the sample population is small and when the main objective is to choose cases that are informative to the research topic selected. Purposive sampling is very useful in the early stages of an exploratory study. One of the disadvantages of this technique is that the sample may have characteristics different from population characteristics.

iii. Convenience Sampling:

Using this sampling method, a researcher is free to use anything that they could find in the research outline. The sample is selected based on preferences and ease of sampling respondents. This sampling is easier to conduct and less expensive. However, it has poor reliability due to its high incidence of bias. In ICT, convenience sampling seems to be dominant especially in cases of organisations that conduct web surveys, mail their responses to a survey questions and SMS their opinions to a question. Although convenience sampling can cater to a lot of data, it is not reliable in terms whether the sample represents the real population or not.

4.5 Meaning of probability

Probability can be defined as the ratio of the frequency of a single outcome to the total number of possible outcomes. As an example, suppose we toss a coin. Two outcomes are possible: the coin comes up heads or tails. The probability of obtaining heads (one outcome) is the ratio of that single outcome to the two possible outcomes, or 1:2, or 1/2, or .50, or 50%. If the coin is balanced, the probability of obtaining tails is exactly the same as that of obtaining heads—it is also .50.

If we roll a single die, 6 possible outcomes can occur: the spots can come up as 1, 2, 3, 4, 5 or 6. The probability associated with each individual outcome is therefore equal to 1/6 or .1667. In the simple examples above, each outcome has an identical probability of occurring (i.e., each has equal likelihood). The balanced physical structure of the coin or die does not favor heads over tails, or 1 over 6, or 2 over 5, etc. But the individual probabilities within a set of outcomes are not always equal. Let’s look at what happens when you throw two dice. Since each die can come up with 1 to 6 spots showing, we have 11 sums that represent the outcome of throwing two dice: the numbers 2 (1+1), 3 (2+1), 4 (1+3 or 2+2), … 11, 12. But the probability of getting each number is not equal to 1 / 11, since there are differing numbers of combinations of values of the dice that may produce a number.

If you count the different combinations that the dice might show, you'll see that there are 36 unique pairs of values. Each of these pairs are equally probable, since they are the result of the action of two independent dice, each of which possess equally probable individual outcomes. Since each pair of numbers represents a unique outcome, we can divide the number of unique outcomes (1) by the total number of possible outcomes (36) to get the probability of any single pair of numbers.

But we can get this result in an way that is easier than listing all combinations. Let's consider a specific instance: What is the probability of getting the pair that has a [3] on Die 1 and a [5] on Die 2? If we throw one die, we know the probability of getting a [3] is 1/6. So, we will expect to have to throw the dice 6 times in order to get the first number of the pair on Die 1. If we then consider Die 2, we will expect to get a [5] (or any other given number) only on one of every six throws. So it should take us 6 times 6, or 36 throws before we can expect Die 1 to come up [3] and Die 2 to come up [5].

This is the same probability (1/36) that we find if we list all possible combinations, but we can find it by simply multiplying two simple probabilities. This is called the multiplicative rule of probabilities, and it can be stated like this:

"The probability of two independent events oc curring jointly (i.e., within the same observation) is the product of the individual probabilities of each event." Using the multiplicative rule, we don't have to list all 36 pairs of outcomes to compute the probability of a particular outcome. We know the probability of Die 1 coming up [3] is 1/6 and of Die 2 coming up [5] is 1/6, so we can just compute:

Prob of [3&5] = 1/6 x 1/6 = 1/36

We can use the multiplicative rule to compute other probabilities. For example, what is the probability of throwing four heads in a row with a coin?

Prob [4 heads] = 1/2 x 1/2 x 1/2 x 1/2 = 1/16

What is the probability of the sum of the dice being 5? We see that there are 4 combinations of Die 1 and Die 2 values that sum to 5. Since there are 36 unique pairs of values, the probability of getting a sum of 5 is 4/36, or 1/9. We expect to get a sum of 5 about once in every nine throws of the dice. We can compute this in another way by using the additive rule of probabilities. This rule states:

“The probability that one of a set of independent outcomes will occur is the sum of the probabilities of each of the independent outcomes.”

Applying this rule to the simple question above, we get:

Prob[sum of 5] = Prob[1&4]+Prob[2&3]+Prob[3&2]+Prob[4&1]
= 1/36 + 1/36 + 1/36 + 1/36
= 4/36 = 1/9

Another example: “What is the probability of getting either a sum of 5 or a sum of 6?”

Prob[sum 5 or sum 6] = Prob[sum 5] + Prob[sum 6]
= 4/36 + 5/36
= 9/36 = 1/4

Restated, we expect to get a sum of 5 or 6 about one-fourth of the time. The rules can be combined to compute the probability of complex situations.

4.6 Normal Distribution

It is the general tendency of the quantitative data to take the symmetrical bell shaped form. This tendency may also be stated in the form of a “principle” as follows: measurement of many natural phenomena and many mental and social traits under certain conditions *tend* to be distributed symmetrically about their means in proportions which approximate those of the normal probability distribution.

Theoretically, the normal curve is a bell-shaped, smooth, mathematically defined curve that is highest at its center. From the center it tapers on both sides approaching the X-axis asymptotically (meaning that it approaches, but never touches, the axis). In theory, the distribution of the normal curve ranges from negative infinity to positive infinity. The curve is perfectly symmetrical with no skewness.

The cases where the data is distributed around a central value with no bias towards right or left side draws closer to a Normal Distribution as shown in the figure 1.

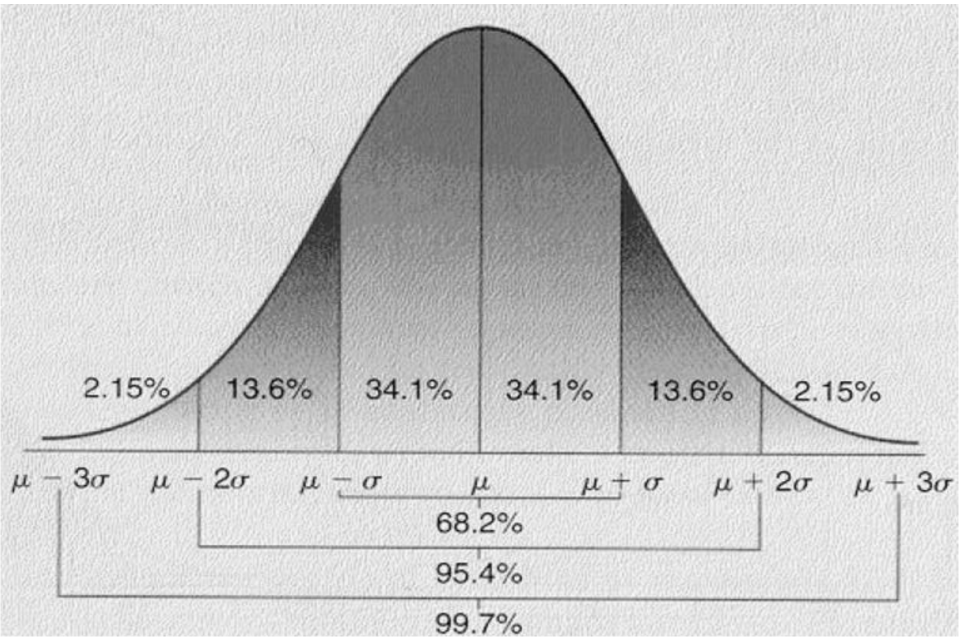


Figure 1: Normal Distribution

The normal distribution can be completely described by two important descriptive measures:

1) Mean

2) Standard deviation

<u>Properties of the Normal Probability Curve (NPC)</u>

Physical Properties:

- The NPC is a **bell shaped continuous curve**. It is sometimes also called as a "bell curve" because its shape closely resembles a bell.
- The NPC is a one-peaked curve. The highest frequency is obtained by only a particular event or score and therefore it can be called as **uni-modal curve.**
- The NPC tapers equally on both the sides of its peak it shows no skewness or in other words it is a **zero skewed curve.**
- The NPC is an open ended curve i.e. it does not touch the X axis from both the sides. The term used for this property is **“asymptotic”**.
- It is symmetrical about the midpoint of the horizontal axis.

Statistical properties:

- The total area under a Normal Probability Curve is always one. Also it is known what area under the curve is contained between the central point (mean) and the point where one standard deviation falls. In fact, working in units of one standard deviation, we can calculate any area under the curve.
- The point about which it is symmetrical is the point at which the mean, median and model all fall. Thus Mean = Median = Mode.
- According to the empirical rule for normal distribution, 50% of the scores occur above the mean and 50% of the scores occur below the mean.

  Approximately 34% of all scores occur between the mean and 1 standard deviation above the mean.

  Approximately 34% of all scores occur between the mean and 1 standard deviation below the mean.

  Approximately 68% of all scores occur between the mean and + 1 standard deviation.

  Approximately 95% of all scores occur between the mean and + 2 standard deviations.

  Approximately 99.74% the data will fall within 3 standard deviation of the mean.

4.7 Skewness and kurtosis

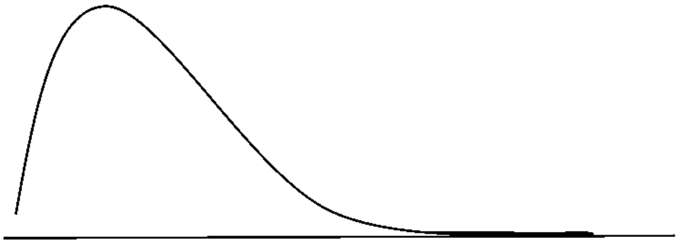
It is not that all the variables occur in the form of normal distribution. There are two main ways in which a distribution can deviate from normal: (i) lack of symmetry (called **skew**) and (ii) pointyness (called **kurtosis**). The empirical data deviate from normalcy too. The deviations from normalcy can be measured by:

4.7.1 Skewness:

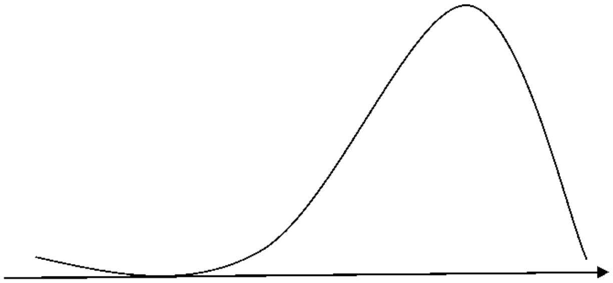
Although the NPC is scattered symmetrically around its mean at the centre but sometime the spread is not equal on both sides of the mean this is skewness. The mean, median and mode coincide to give the NPC a perfect shape, so that the left and right sides of the curve are balanced. If these three statistic do not coincide in a distribution then the symmetry is disturbed and the distribution appears to shift either to the left or to the right. In a perfectly normal distribution the skewness is zero, but if the scores are massed at the lower side the spread of the curve is gradually towards the right side, this is called positive skewness. The negative skewness occurs when the scores are clustered towards the higher end of the distribution, consequently, the spread of the curve is more towards the left. The index of skewness is given by the formula:

Skewness (sk) = 3(mean – median) ÷ Standard deviation

A distribution has a positive skew when relatively few of the scores fall at the high end of the distribution. In positive skewness; the curve tapers to the right with scores loaded at the left.



A distribution has a negative skew when relatively few of the scores fall at the low end of the distribution. The negative skewness is characterized by the curve tapering at the left while the scores are loaded at the right.

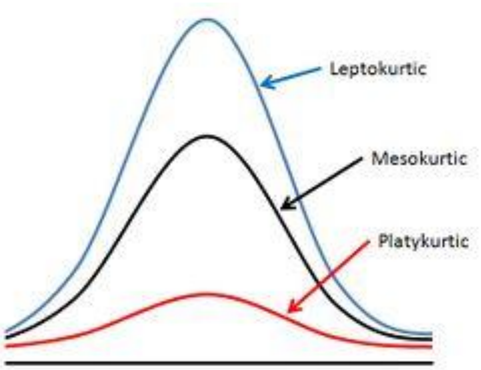


4.6.2 Kurtosis:

Distributions also vary in their kurtosis. Kurtosis refers to the degree to which scores cluster at the end of the distribution (known as the tails) and how pointy a distribution is (but there are other factors that can affect how pointy the distribution looks). Thus, kurtosis refers to the degree of flatedness or peakedness in a distribution.

The Normal Probability Curve is a mesokurtic curve. Distributions are generally described as platykurtic (relatively flat), leptokurtic (relatively peaked), or somewhere in the middle i.e., mesokurtic. , A distribution with positive kurtosis has many scores in the tails (a so called heavy-tailed distribution) and is pointy. This is known as **leptokurtic** distribution. In contrast, a distribution with negative kurtosis is relatively thin in the tails (has light tails) and tends to be flatter than normal. This distribution is called platykurtic. Ideally, we want our data to be normally distributed (i.e., not too skewed, and not too many or too few scores at the extremes).

Kurtosis (ku) = Q ÷ (P90 -- P10)



UNIT V: INFERENTIAL STATISTICS:

The statistical analysis of behavioral research data popularly employs parametric and nonparametric tests. The distinction between the tests is based on the assumptions of population parameter and the type of data to be analyzed.

- The normalcy of the population of scores from which the sample is drawn is to be normally distributed for the parametric tests but the non-parametric tests are not bound by this.
- The parametric tests are applied on data measured in quantitative scale of measurement. The non- parametric tests on the other hand are more suitable for the data in the qualitative scales.

It is seen that the parametric statistical tests enjoy an edge over the non-parametric tests in terms of “power of a test”, parametric ones being more powerful. However, it is important to meet the assumptions of the parametric tests for their optimal application to the data. The non-parametric tests on the other hand have limited scope and applications. Their use is found in data that do not meet the assumptions of the parametric tests. In the present module the reader will be explained the concept of the two major types of statistical tests, parametric and non-parametric, along with their applications.

5.1 Parametric Tests

The name “parametric test” itself seems to draw from the fact that the tests depend heavily upon the parameters obtained and estimated from the data. The tests share two basic characteristics.

- The first characteristic being the assumption about the nature of parameters which have been taken on the selected variables.
- The second characteristic is the tests involve numeric calculations which require the estimation of the population parameters from the sampled data.

So, the population parameters and their estimation from the sample is the key to the parametric tests. These tests are also referred to as the assumption based tests and therefore, should be used only when their requirements are met. However, there are aberrations in meeting the assumptions and still under certain circumstances parametric tests are found to be suitable.

Some of the most popularly used parametric tests are t-test, f-test, z-test, ANOVA etc.

5.1.1 Assumptions of parametric test:

The parametric tests are referred to as assumptions-based tests and they frequently require the following assumptions to be fulfilled before their implementation to the data.

- The data should be independent i.e. no observation drawn from and individual can be repeated. So, the selection of an individual to a sample is not dependent upon the selection of any other individual.
- The observations should be drawn from a population that has a normal distribution of the variable being quantified. So, observations should be extracted from only normally distributed populations.
- The samples drawn from a population must follow homogeneity of variance i.e. samples of a population should have approximately same variance. There should not be any significant difference in the variances of the samples drawn from a population. This assumption holds more strongly under the circumstances when the sample size is small.
- The variable being quantified must be in higher scales of measurement i.e. interval or ratio. The data in the qualitative scales namely, nominal and ordinal is not suitable for parametric tests.
- The data should be continuous in nature to qualify for parametric tests.

It should be kept in mind that the above discussed assumptions are the preliminary requirements for parametric tests. Additional criteria do exist for certain tests under this class.

One must also know that researchers do ignore some of these assumptions and therefore interfere with the validity of the study. It should also be kept in mind that when the sample size is large certain violations in the assumptions do not matter very much. The tests still enjoy considerable statistical power.

Statisticians are also less concerned with the criterion of scale of measurement being used to quantify the variable. It is not wise to apply parametric tests to the nominal and ordinal data. However, if the data is in ordinal scale of measurement and has sufficient levels (e.g. 7 or more as in Likert scale) and other assumptions are met one can consider applying parametric test.

It is important to note that if the criteria of a parametric test are not met by the data then it will be inappropriate to use that test. Although it should be kept in mind that such data would be suitable for non-parametric test applications, as they have their own set of criteria to be taken care of.

In spite of the violations of the assumptions the parametric test they have been found to be reasonably accurate. But in certain conditions the robustness of these tests can be badly affected, this may mislead the results and the probability of the type I errors may markedly increase.

Although there are statistical procedures that may indicate violations in assumptions of parametric tests, they are themselves dependent on same hypothesis testing procedures. So, similar problems associated with statistical power may crop up there too. It should be kept in mind that if a large sample is used then even a small and unimportant degree of violations can be detected. Therefore such tests that detect violations of assumptions should be employed to know how far the data is from ideal conditions of parametric applications.

Levels of Statistical Significance

Establishing the statistical significance helps the researcher in assuring that the statistic obtained is reliable enough. In no way does it guarantee that the findings or the results obtained are important or they ensure any decision making use. The significance level is just a critical probability point that facilitates the decision making about how important and acceptable is the difference between the observed and table values of the statistic and whether the hypothesis based on that should be accepted or rejected.

The level of significance represented by ά (alpha) is a decision criterion that is used in deciding that the obtained value of a statistic has a low chance of occurring by chance if the null hypothesis is correct or accepted. The probability value represented as p is another way to put the observed or computed level of significance. A low p value points at the fact that there is small chance of a statistical statement to be true. The behavioral science research conventionally take the level of significance as .05 and, 01.

One-tailed & two-tailed tests

One tailed tests or directional tests: Here the alternative hypothesis Ha states that the parameter differs from the stated values of Ho in one particular direction (critical region is in one tail of sampling distribution).

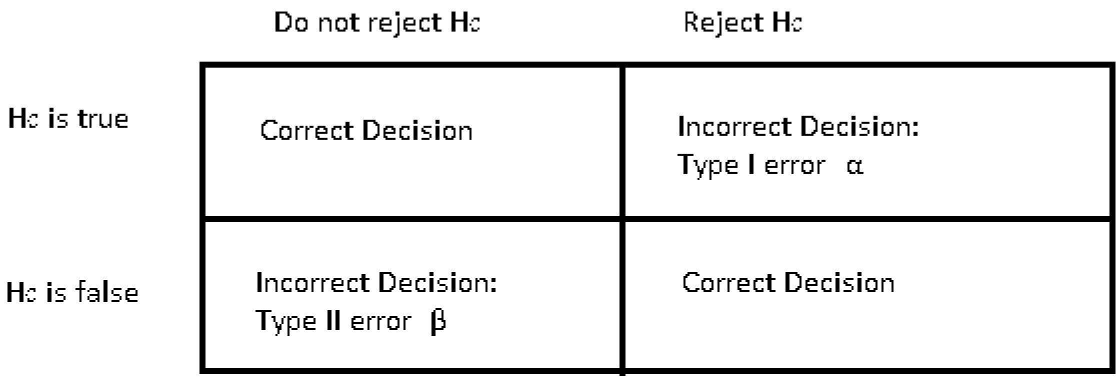
Two tailed tests or Non-directional tests: Here the alternative hypothesis Ha states that the parameter may be less or greater than the stated values of Ho (critical region is divided between both the tails of sampling distribution).

Types of errors

The concept of hypothesis testing rests on probability theory. The data is drawn on a sample and the observations made on the sample are used to draw inferences about the population from which the sample has been taken. The entire idea judges the probability that might be associated with the sample statistic had it been the population parameter. So, now the statements that cannot be made with certainty about a population can be made on the basis of the sample statistic with some degree of error. The hypothesis testing procedure is therefore fraught with some errors always. The researchers using sampling distribution of a statistic can make two types of decision errors: Type I error and Type II error. They are referred to as decision errors because even after following the right procedures one ends up with wrong decisions.

i. *Type I error:* A type I error is a statistical condition that prevails when one rejects the null hypothesis when it is actually correct. So one incorrectly rejects a relationship which actually holds true in the population. The probability of the type I error is the ά alpha or the significance level. An example would demonstrate this better: a patient is tested to have an ailment when actually he does not suffer from it.

ii. *Type II error:* The second type of decision error is the type II error. When a researcher accepts the null hypothesis when it is actually false then the type II error occurs. The data gathered on the sample is unable to detect the difference between the statistic when it is actually present in the population. The probability of a type II error is given by β, beta level or beta rate. The type II error is detected by a number of factors like: sample size and significance level. The type II error is just like concluding that the patient does not have an ailment when he actually suffers from it. When comparing two means, concluding the means were different when in reality they were not different would be a Type I error; concluding the means were not different when in reality they were different would be a Type II error.



	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	Incorrect Decision: Type I error α
H_0 is false	Incorrect Decision: Type II error β	Correct Decision

Remember that if the null hypothesis is true, it should not be rejected, but if the null hypothesis is false, it should be rejected.

5.2 Non-parametric tests:

When the data are in qualitative scales of measurement and application of parametric tests is not suggested, the non-parametric tests come a researchers' rescue. Non-parametric tests are those statistical tests where there is no knowledge about the population and its parameters. Even in such situations it is required and is essential to test the hypothesis about the population parameters. Non-parametric tests do not recommend any conditions about the population parameters. As they do not need compliance to distribution characteristics, the tests are also called "distribution free" tests.

The nonparametric tests cater to the data that is obtained in the qualitative scales. The data is in the forms of either frequency counts or rankings. For the reason that the data is in lower scales it is thought that non parametric tests are low on precision and prone to type II error (accepting the null hypothesis even when it is false). It is perhaps for this reason only that non parametric tests are low on popularity and are used only when the data does not comply with the assumptions of parametric test. However, statisticians do have counter arguments that as the non-parametric tests are not dependent upon assumptions (which are often flouted) for their validity they are more dependable.

Some of the popularly used non- parametric tests are Mann-Whitney U test, Rank sum test, Kruskal-Wallis test, Kendall’s Tau etc.

Non- parametric tests are used when

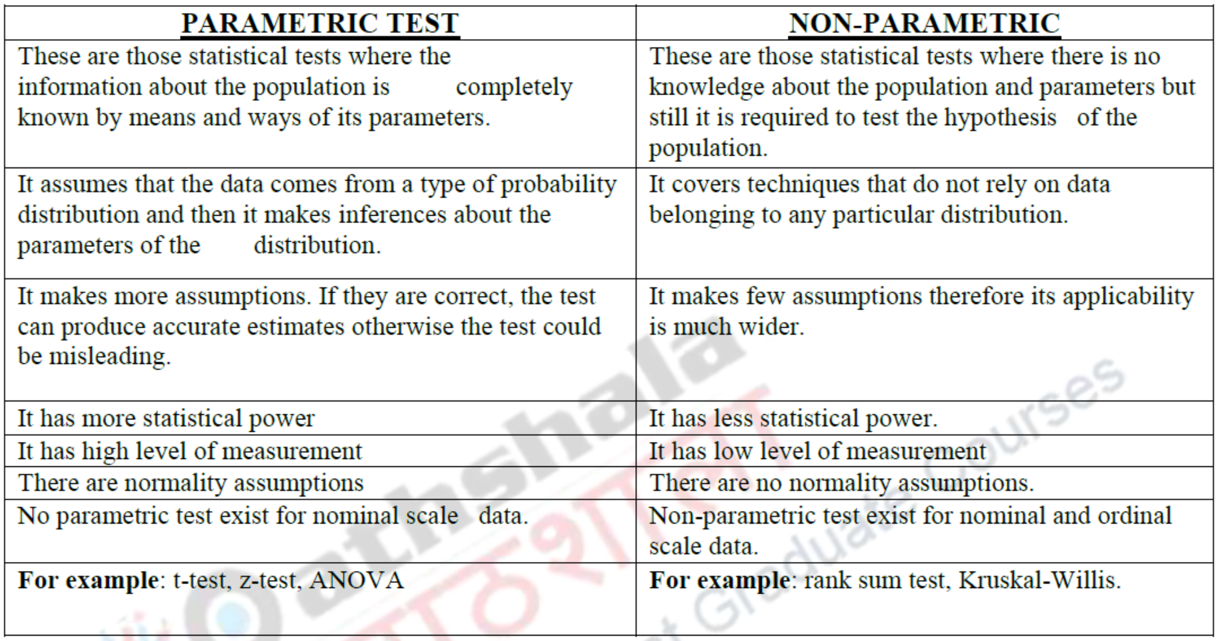
- Sample size is small
- Assumptions of normalcy about the population is doubtful
- Data is in lower scales of measurement

5.2.1 Assumptions of non-parametric test

These tests are also referred to as “assumptions free tests” as they do not stress heavily on their conditional applications. Although the tests do expect certain characteristics to be fulfilled by the data, they are not rigid and elaborate. However, it does not mean that the tests are completely assumption free. There are certain pre-requisites associated with these tests too. Some of the requirements for the application of non-parametric tests are:

- The population from which the research samples are drawn may not be normal.
- The variables are measured in nominal scale (frequency counts) or ordinal scale (ranking).
- Observations must be independent.

The differences between parametric and non- parametric tests



PARAMETRIC TEST	**NON-PARAMETRIC**
These are those statistical tests where the information about the population is completely known by means and ways of its parameters.	These are those statistical tests where there is no knowledge about the population and parameters but still it is required to test the hypothesis of the population.
It assumes that the data comes from a type of probability distribution and then it makes inferences about the parameters of the distribution.	It covers techniques that do not rely on data belonging to any particular distribution.
It makes more assumptions. If they are correct, the test can produce accurate estimates otherwise the test could be misleading.	It makes few assumptions therefore its applicability is much wider.
It has more statistical power	It has less statistical power.
It has high level of measurement	It has low level of measurement
There are normality assumptions	There are no normality assumptions.
No parametric test exist for nominal scale data.	Non-parametric test exist for nominal and ordinal scale data.
For example: t-test, z-test, ANOVA	**For example**: rank sum test, Kruskal-Willis.

Advantages and disadvantages of non-parametric tests over parametric tests

Although parametric tests enjoy greater popularity in the field of statistical applications, the non-parametric tests have their own utility and advantages. Let us compare the advantages and disadvantages of both in terms of derivations, applications, assumptions and statistical power.

- *Ease of application and simplicity:* as the non-parametric tests are concerned with data in basic scales it is easier and simpler to apply them. The parametric tests on the other hand need mathematical knowledge and are difficult to comprehend by the researchers with no mathematical background.
- S*impler or no assumptions compliance:* the non-parametric tests in comparison to parametric ones are based on fewer and flexible assumptions. As the compliance to a number of assumptions about the population parameters is not needed the non-parametric tests enjoy wider scope of applications.
- *Violation in assumptions:* the parametric tests have rigid assumptions to adhere to and thus, are difficult to comply. These violated assumptions effect the validity of the results when the tests are used despite of the assumptions. The nonparametric tests are less susceptible to violations, which if detected can be easily rectified also. But if all the assumptions of the parametric tests are fulfilled they enjoy a clear edge over the non-parametric ones.
- *Scales of measurement:* the parametric tests with their higher levels of measurement (interval & ratio) use more enriched and profound data that is conducive to higher order statistical analysis. This presents refined, reliable and more powerful data analysis opportunities to the researcher.
- *Sample size:* one of the main factors in use of the two tests is the size of the sample drawn too. The parametric tests with small sample size are a waste of data, as said by some statisticians. The non-parametric tests even with small sample size (even if it is less than 10) are efficient, easy to apply and prove better than parametric equivalents.

Disadvantages of non-parametric test

- Non parametric tests are not as efficient as parametric tests.
- Non parametric tests have less statistical power than parametric tests.

5.3 Chi-square and contingency coefficients:

Chi-Square (χ^2), introduced by Karl Pearson in 1900, is the discrepancy between **observed (O)** frequencies (the frequencies that are actually obtained from our random samples) and **expected (E)** frequencies (the frequencies expected according to the hypothesis) of nominal variables, in which subjects are grouped in categories or cells. It proves especially useful in conditions that require comparing empirically obtained results with the theoretically expected results on some hypothesis.

There are two basic types of chi-square analysis, the **Goodness of Fit Test**, used with a single nominal variable, and the **Test of Independence**, used with two nominal variables. Both types of chi-square use the same formula. The general formula for chi-square is –

$$\chi^{2} = \Sigma(\frac{11_{1}1\;1_{1}1^{1}}{1_{1}})$$

where,

11 = observed frequency

11 = expected frequency

Thus, the value of χ^2 equals the sum of O-E differences squared and divided by E. The more O differs from E, the larger the χ^2 is. When χ^2 values exceed the appropriate critical value, it is declared significant. Chi-Square is used on variables that are measured on a nominal or ordinal scale. Although the chi-square test is popularly conducted in terms of frequencies, it is best viewed as a test about proportions. The calculated value of chi square can be evaluated with the appropriate degrees of freedom {df= (r-1) (C-1), where r stands for rows and c for columns in the data table} in the chi square table to make inferences about the hypothesis.

5.3.1 Assumptions of Chi-Square

As stated the chi square is a test of goodness of fit and independence. Therefore the test does not command many of the usual assumptions that statistical test are based on. Some of the assumptions required for the chi square test are:

- It is assumed that random sampling is used to collect the samples from populations about which inference is to be made.

- The observations are assumed to be independent of each other. This means that per subject, there should be only one observation.
- It is assumed that in repeated experiments, the observed frequencies will be normally distributed about the expected frequencies

The Goodness of Fit Test.

The Goodness of Fit Test is applied to a single nominal variable and determines whether the frequencies we observe in k categories fit what we might expect. The Goodness of Fit Test can be applied with equal or proportional expected frequencies. Equal expected frequencies are computed by dividing the number of subjects (N) by the number of categories (k) in the variable. The computed value of c2 value is compared to the appropriate critical value using α and df. The critical value is found in the Chi-square Table.

Chi-Square Test of Independence

The test of independence analyzes the relationship between two nominal variables. The procedure uses the special terms **independent** to mean *not related*, and **not independent** to mean *related*. The two nominal variables form a contingency table of cells.

The effect size for the chi square test is phi coefficient (Ǿ). This gives the degree of association of the nominal variables. The phi coefficient can be obtained by dividing the χ^2 of the sample with the sample size. The value ranges from a minimum of 0 to a maximum of 1. The phi coefficient of .5 according to Cohen (1988) convention is high and which means that the researcher can have confidence that the results are reliable and can be used for population predictions.

Let us take up an example considering five flavors of a chips brand being equally preferred by consumers. We are hence, exploring the possibility that there are an equal proportion of individuals that prefers each flavor i.e. 1/5.

Therefore,

Ho : P1 = P2 = P3 = P4 = P5 = 1/5

The alternate hypothesis is simply that the null hypothesis is not true.

Suppose we select 125 subjects to represent the entire population of consumers and the observed frequencies of preference are shown in the table below.

	Flavour 1	Flavour 2	Flavour 3	Flavour 4	Flavour 5
1_1	21	30	24	28	22
1_1	25	25	25	25	25

NOTE: 11 = (proportion hypothesized to characterize the category) * Sample Size

= 1/5 * 125

= 25

χ^2 = 0.64 + 1 + 0.04 + 0.36 + 0.36

= 2.4

The value calculated above is compared with the critical value of c2. The critical values of chi square can be taken from the statistical tables available for the purpose. The table gives the probability of exceeding the critical values (also referred to as the tabled values) of the chi square for the specified number of degrees of freedom (df). It is readily available in most of the statistics books.

If the χ^2 cal > χ^2crit , then the null hypothesis is rejected.

To determine the critical value, we need to first find out the df i.e. degrees of freedom. For problems pertaining to chi-square, the degrees of freedom will be C-1 (C is categories).

So, in the above example, the df comes out to be 5-1 = 4

Considering this value, we find that the critical value is 9.49 when $\alpha = 0.05$ and 13.28 when $\alpha = 0.01$

The χ^2 cal is therefore, lesser than χ^2 crit which means that the null hypothesis is not rejected. In other words the differences in the frequency of occurrence in the observed data can be attributed to chance errors rather than experimental conditions.

Hypothesis testing in the case of only two categories

When there are only two categories i.e. when df=1, a chi-square test is seen as a test about a single proportion. Suppose, for example, we wanted to know about consumers' preferences among just two flavors of chips and not five. We take a sample of 50 subjects selected randomly from the population and record their preferences among these two flavors. Since, there are only two categories, we may identify P as the proportion of people choosing Flavor 1 and (1-P) as the proportion of people choosing Flavor 2.

	Flavour 1	Flavour 2
1_1	20	30
1_1	25	25

Thus, we get the following null and alternative hypotheses-

Ho : P = .50

HA : P ≠ .50

Let us calculate the χ^2 for this problem:

$$\chi^2 = \Sigma(\frac{11_1 1\ 1_1 1^1}{1_1})$$

$$= \frac{(1\,1\,1\,1\,1)^1}{1\,1} + \frac{(1\,1\,1\ 1\,1)^1}{1\,1}$$

$$= \frac{(1\ 1\ )^1}{1\,1} + \frac{(1\ )^1}{1\,1}$$

$$= 1 + 1$$

$$= 2$$

<u>When to use Chi-square test?</u>

- The hypothesis predicts a difference between two conditions or an association between variables.
- The set of data must be independent (no individual should have a score in more than one ‘cell’)
- The data are in frequencies (i.e., nominal) and frequencies must not be percentages.

Lewis and Baker (1949) discussed the common errors that may crop up in application of chi square test to research. One of the commonly occurring error is very low expected frequency.

- The reasonable expected frequency in a chi square table was recommended to be at least 10, by the authors. Generally the test is unreliable when the expected frequencies fall below 5 in any cell. You need at least 20 participants for a 2 x 2 contingency table.
- Low expected frequency result in type 1 errors and also reduce power of the test

5.4 Correlation:

The concept of correlation coefficient is invaluable to quantitative research. The Pearson's product moment correlation coefficient is a linear relationship among two continuous variables and it has been explained in the module 22. Many a times the research data does not conform to the assumptions of linearity and continuity and hence, not suitable for Pearson's "r". So, other measures of correlation have been also been devised by statisticians.

5.4.1 Point- Biserial Correlation

The point biserial correlation coefficient is an estimate of product moment correlation under special circumstances when certain assumptions have been met. A research data may comprise of continuous as well as categorically measured variables. If the categorical variable is dichotomous then the relationship between the two variables can be established using point- biserial correlation.

The term dichotomous literally means cut into parts. So, if a variable can be measured in just two categories it can be called dichotomous. An example of a dichotomous variable is gender; "male" or "female" there is no underlying continuum between the two categories.

In point –biserial correlation coefficient (rpb) one of the variables is categorical and dichotomous into mutually exclusive categories and the other is continuous. So, if a study involves exploring the relationship between gender and anxiety point biserial correlation coefficient is an appropriate measure. The anxiety is a continuous variable whereas; gender is a categorical dichotomous variable.

Point bi serial correlation coefficient is useful in:

- When the data comprises of a continuous and other dichotomous variable
- The data is not suitable for pearson's product moment correlation coefficient
- Analysis of items of a test i.e. in item test correlation

The calculation of point biserial correlation coefficient using product moment formula

The point biserial correlation coefficient is a product moment correlation and can also be calculated using the formula:

$$r_{pb} \quad = \quad \frac{(N\Sigma XY) - (\Sigma X.\ \Sigma Y)}{\sqrt{\{[(N\Sigma Y^2) - (\Sigma Y)^2]\ [(\ N\Sigma X^2) - (\Sigma X)^2]\}}} \qquad \text{----------eqn. (2)}$$

Numerical Application of point bi- serial correlation

The researcher wishes to explore if anxiety scores obtained during a task performance were related to the gender of the group being tested. The scores on anxiety test as obtained by the males and females are presented under table I below:

Table I: the anxiety scores obtained by males (1) and females (0)

S.No.	Anxiety (X)	Gender(Y) Male (1) Female (0)	X^2	Y^2	XY
1	25	1	625	1	25
2	23	1	529	1	23
3	18	0	324	0	
4	24	0	576	0	
5	23	1	529	1	23
6	20	0	400	0	
7	19	0	361	0	
8	22	1	484	1	22
9	21	1	441	1	21
10	23	1	529	1	23
11	21	0	441	0	
12	20	0	400	0	
13	21	1	441	1	21
14	21	1	441	1	21
15	22	1	484	1	22
$N_{male} = 9$ $N_{female} = 6$	Total= 323		$\Sigma X^2 = 7005$	$\Sigma Y^2 = 9$	$\Sigma XY = 201$

M_{males}=201/9= 22.33 **p= 9/15= 0.6**

$M_{females}$= 122/6= 20.33 **q= 6/15 = 0.4**

M_{total} = 323/ 15 = 21.53 **σ_{total}**

Using equation 1 the point biserial correlation coefficient can be calculated as:

$$r_{pb} \quad = \mathbf{(22.33 - 20.33)/1.82\ \sqrt{(0.6 * 0.4)}}$$

$$= \qquad 0.54$$

The same value of point biserial correlation coefficient can also be obtained using product moment correlation formula given in equation (2) above.

$$\mathbf{r_{pb}} = ((\mathbf{15} * \mathbf{201}) - (\mathbf{323} * \mathbf{9}))/\surd(\{[(\mathbf{15} * \mathbf{9} - (\mathbf{9})]\ \ [(\mathbf{15} * \mathbf{7005}) - (\mathbf{323})^\wedge\mathbf{2}\ ]\})$$

$$= .54$$

Significance of point biserial r

The obtained value of point biserial r can put to test against the null hypothesis. The degree of freedom (n- 2) can be used to find the critical value of r from the table. If the r_{calc} is more than r_{crit} it can be taken as significant and reject the null hypothesis.

5.4.2 Biserial correlation

The concept of biserial correlation coefficient is quite similar to the point biserial correlation. The difference between the two pertains to the measurement of the categorical variable. The dichotomy in the variable is not true but has an underlying continuity in it. In other words, one may say that the variable in which dichotomy is assumed may be found to be continuous and normally distributed had more information been taken into account. Like, dividing the students into those who are graduates and those who are not, but one ignores small section of those who may have attempted graduation studies but unable to complete or are in the process of accomplishing it. So, the categorical variable cannot be said to be split in two categories of graduates and non- graduates rather an arbitrary split point is created to facilitate the grouping.

Whenever, the research requires establishing relationship between two variables one of which is continuous and the other is dichotomized by placing an arbitrary division, the biserial correlation is a suitable choice.

Bi serial correlation coefficient is useful in:

- When the data comprises of a continuous and other dichotomous variable (with arbitrary division imposed)
- The data is does not meet all the assumptions of pearson's product moment correlation coefficient
- Analysis of items of a test i.e. in item test correlation

Bi serial correlation coefficient is NOT useful in:

- The biserial r is not a very popular method as it is difficult to calculate when the data is not normal.
- Biserial r cannot be used in a regression equation
- Biserial r cannot be used for comparing with other correlation coefficients as it goes beyond the range of + 1.0 and -1.0.

Calculation formula for biserial correlation coefficient

The biserial correlation coefficient can be calculated using the following formula:

$$r_{\text{bis}} = \frac{M_p - M_q}{\sigma} * \frac{pq}{u} \quad \text{-----------eqn(3}$$

where,

Mp is the mean of the group in category 1

Mq is the mean of the group in category 2

σ is standard deviation of the group

p is proportion of group in category 1

q is proportion of group in category 2

u is the height of the normal curve ordinate dividing the two parts p and q.

Calculation formula for biserial correlation coefficient

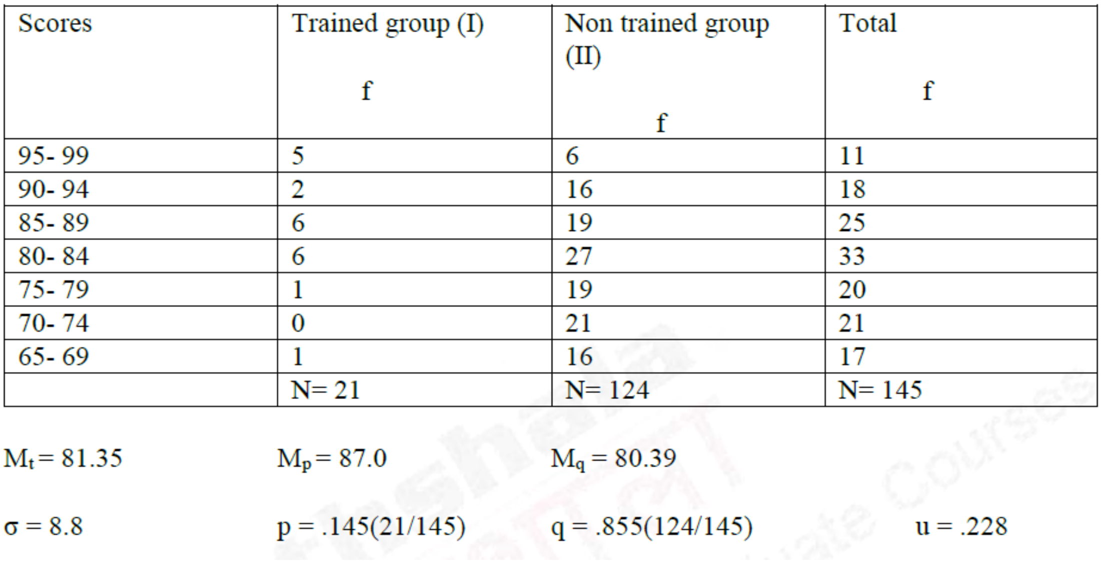
An alternative formula to calculate biserial r is also available, this is slightly convenient than the earlier one mentioned in equation 3.

$$r_{\text{bis}} = \frac{M_p - M_t}{\sigma} * \frac{p}{u} \quad \text{---------eqn(4)}$$

Numerical Application of bi- serial correlation

A school counselor wants to check out if a special training program she devised is related to performance in numerical ability test for students. Among the 145 students 21 were randomly chosen to undergo the training program. The two groups were of 21 students with training and 124 without training. The scores obtained are shown in table II.

Table II: Scores obtained on numerical ability test by trained and non- trained students.



Scores	Trained group (I) f	Non trained group (II) f	Total f
95- 99	5	6	11
90- 94	2	16	18
85- 89	6	19	25
80- 84	6	27	33
75- 79	1	19	20
70- 74	0	21	21
65- 69	1	16	17
	N= 21	N= 124	N= 145

$M_t = 81.35$ $M_p = 87.0$ $M_q = 80.39$

$\sigma = 8.8$ $p = .145(21/145)$ $q = .855(124/145)$ $u = .228$

The value of u is obtained from the table that gives the heights of the ordinates in a normal distribution of unit areas, with N= 1, M= 0 and SD= 1. As evident one can find u= .228.

$$r_{\text{bis}} = \frac{87-80.39}{8.8} * \frac{(0.145*0.855)}{0.228}$$

$$= .41$$

The same value of biserial correlation can be obtained using equation 4 also.

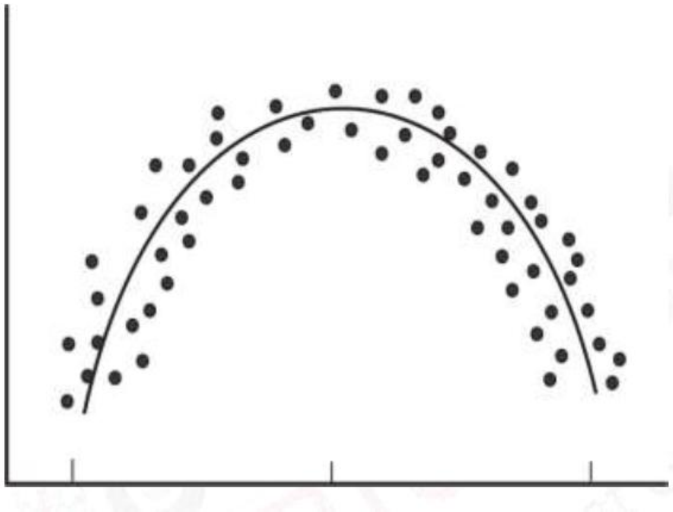
Significance of biserial r

The obtained value of point biserial r can put to test against the null hypothesis. The degree of freedom (n- 2) can be used to find the critical value of r from the table. If the rcalc is more than rcrit it can be taken as significant and reject the null hypothesis.

5.4.1 Curvilinear relationship

The pearson's product moment correlation is best suited to data that is linear in nature, otherwise error may occur. The research data however is not linearly related always. When the straight line of fit is not apt enough to describe a set of data it is said to be curvilinear or more simply a non-linear relationship. In such situations r is not an adequate measure of correlation. In such data a curve rather than a line of best fit is used to describe the degree of relationship that the variables share. Fitting a curve instead of straight line would give a better and more accurate measure of correlation among the variables.

One often encounters non- linearly related variables in data measured in ratio scale, psychophysics etc.

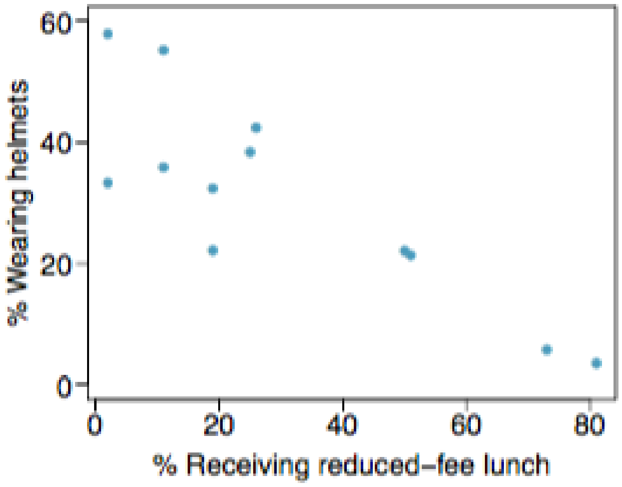


5.5 Regression

The term regression was first used by Francis Galton with reference to the inheritance of stature. He found that the children of tall parents tended to be less tall while children of short parents tended to be shorter. Thus, the heights of the off-springs tended to move towards the mean height of the general population. The tendency of maintaining the mean value was called the principle of regression by Galton and the line describing the relationship of height in parent and offspring was called the regression line. Thus the predictor variable here becomes the parent height and the outcome variable is the child's height. The prediction of one variable from the other is the concept of regression.

Similar to correlation, regression is used to analyze the relationship between two continuous variables. It is also better suited for studying functional dependencies between factors that is where X partially determines the level of Y. For instance, as age increases, blood pressure increases. But the length of arm does not have any effect on length of the leg. Also, it becomes better suited than correlation for studying samples in which the investigator fixes the distribution of X or the predictor variable.

For example, if independent variable be the percentage of children receiving reduced fee school lunches in a particular neighborhood as a substitute for neighborhood socio economic status and the dependent variable be the percentage of bicycle riders wearing helmets. The researcher finds a strong negative correlation of –0.85. the data obtained is useful if the researcher wants to know about the helmet wearing behavior on the basis of data obtained on socio- economic status. A straight line of best fit can be fitted to the data using the least square criterion (readers may refer to the module on correlation to read more about this). The line of best fit enables the statistician to develop the regression model.



5.5.1 Regression model

A line of best fit is described as a straight line that runs through the data in such a manner that the sum of square of the deviances from the mean is minimum. Let us understand this concept.

A line is recognised by its slope or the angle of the line describing the change in Y per unit X and intercept is where the line crosses the Y axis. Regression describes the relation between X and Y with just such a line.

$\hat{y}$ = predicted value of Y

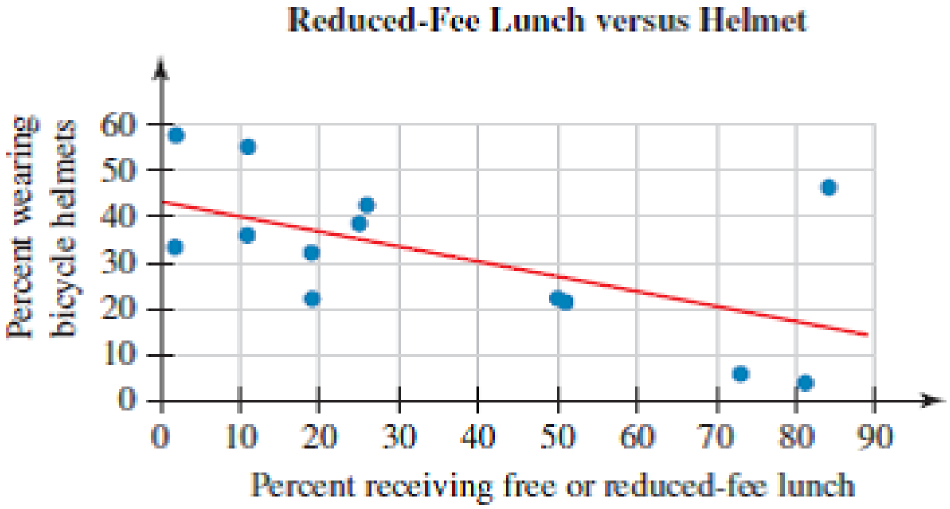
a = intercept of the best fitting line

b = slope of the line

Hence, the regression model is represented by: $\hat{y} = a + bx$

Now identifying the best line for the data becomes a question. Had all the data points fell on a single line, identification of the slop and intercept would have been easier. But since the statistical data has random scatter, identifying a good line becomes a process requiring effort.

The random scatter around the line is recognised as the distance of each point from the predicted line and these distances are referred to as residuals shown below.



The slope in the regression model is the average change in Y per unit X. thus, the slope of -0.54 predicts 0.54 fewer helmet users per 100 bicycle riders for each additional percentage of children receiving reduced fee meals.

Since the regression model can be used to predict the value of Y at a given level of X, like the neighborhood in which half the children receive reduced fee lunch (X=50) has an expected helmet use rate (per 100 riders) as 47.49+(-0.54)(50) = 20.5.

Factors effecting regression

Number of cases: when doing regression, the cases-to-independent variable ratio should be 20:1 that is 20 cases for every independent variable in the model. The lowest ratio could be 5:1 that is 5 cases for every independent variable in the model.

Accuracy of data: one should check the accuracy of data entry to ensure that all the values for each variable are valid.

Missing data: one should look for missing data and if there are a lot of missing values, one should not include those variables in analyses or, delete the cases if there are few cases with missing values or if its important, place the mean values of that variable in the missing places.

Outliers: once should check data for outliers that is an extreme value on a particular item which is at least 3 standard deviation above or below the mean. One may delete these cases if they are not a part of the same population or retain it but reduce how extreme it is that is recode the value.

Assumptions of Regression

Normality: one should check for normal distribution of the data by constructing a histogram or construct a normal probability plot.

Linearity: one assumes linearity or a single straight line relationship between the independent and dependent variables as regression analysis tests for only linear relationships. Any non linear relationship gets ignored.

Homoscedasticity: one assumes homoscedasticity that is the residuals are approximately equal for all predicted dependent variable scores or variability of scores for the independent variables is same at all values of the dependent variable.

Multicollinearity and Singularity: multicollinearity is when the independent variables are very highly correlated (0.90 or greater).singularity is when independent variables are perfectly correlated and one independent variable is a combination of one or more of the other independent variables.

Applications of regression

Regression method can be employed in

- To seek some sortof descriptive relationshipbetween a set of measured variables like a sociologist wanting to establish a relationship between the occupational status of an individual and the educational level of the person as well as his/her parents’educational level
- To provide evidence for a theory for instance in estimating coefficients for an established model employed in relating a particular plant’sweight with the amount of water it receives, available nutrients in soil and sunlight it is exposed to.

- To predict some response variable at a certain level of other input variables which play a role in planning, monitoring, altering or evaluating a process.
  - Areas of demand management and data analysis in the field of business
  - Social, psychological or any research of any field like agriculture to predict the yield of a crop for instance by studying the role of quality of seed, soil fertility, temperature, rainfall etc.

Focus of analysis: the purpose of carrying out the multiple regression analysis in quantitative psychological research is to analyze the extent to which two or more independent variables relate to a dependent variable.

Variables involved: there may two or more than two independent variables which are continuously scaled. The dependent variables are also continuously scaled i.e. either interval or ratio scale of measurement.

Realtioship of the participants' scores across the groups being compared: to be suitable for multiple regression analysis, the participants should have scored on all the variables, or in other words the scores are dependent upon each other.

Assumptions underlying regression analysis:

The assumptions of normality, homoscedasticity, linearity, independence of errors and multicollinearity are applicable to multiple regression analysis.