

UNIT – 1

BUSINESS STATISTICS

What is Statistics?

The word “Statistics” has been derive from the Latin word “*Status*” or Italian word “*Statista*” or German word “*Statistika*”. Each of these words means Political State. Initially, Statistics was used to collect the information of the people of the state about their income, health, illiteracy and wealth etc.

But now a day, Statistics has become an important subject having useful application in various fields in day to day life.

Statistics in Plural Sense:--

In the plural sense, Statistics refers to information in terms of numbers or numerical data such as Population Statistics, Employment Statistics etc. However any numerical information is not statistics.

Example: Ram gets ₹100 per month as pocket allowance is not Statistics. It is neither an aggregate nor an average. Whereas average pocket allowance of the students of Class X is ₹100 per month and there are 80 students in class XI & 8 students in Class XII are Statistics.

The following table shows a set of data that which is Statistics and which is not Statistics.

Data which are not Statistics	Data which are Statistics
A cow has 4 legs.	Average height of the 26 plus male people in India is 6 feet compare to 5 feet in Nepal.
Ram has 200 rupees in his pocket.	Birth rate in India is 18 per thousand compare to 8 per thousand in USA.
A young lady was run over by a speeding truck at 100 km per hour.	Over the past 10 years, India has won 60 test matches in cricket and lost 50.

From above information we can say that “**All Statistics are data, but all data are not Statistics**”

Definition:-

According to **Bowley** - “**Statistics are numerical statements of facts in any department of enquiry placed in relation to each other.**”

According to **Yule and Kendall** ----- “**By Statistics we mean quantitative data affected to marked extent by multiplicity of causes.**”

Characteristics of Statistics in Plural Sense

Main characteristics of Statistics in terms of numerical data are as follows:

- (1) **Aggregate of Facts** – A single number does not constitute Statistics. We can not draw any conclusion from single number. We can draw any conclusion by the aggregate number of facts.
For example, if it is stated that there are 1,000 students in our college then it has no significance. But if it is stated that there are 300 students in arts, 400 students in commerce and 300 in science in our college. It makes statistical sense as this data convey statistical information. Similarly if it is stated that population of India is 130 crore or the value of total exports from India is ₹11, 66,439 crore then these aggregate of facts will be termed as Statistics.
- (2) **Numerically Expressed** - Statistics are expressed in terms of numbers. Qualitative aspects like small or big, rich or poor etc. are not statistics. For instance if we say that Irfan Pathan is tall Sachin is short then this statement has no statistical sense. However if it is stated that height of Irfan Pathan is 6 ft and 2 inch and the height of Sachin is 5 ft and 4 inch then these numerical will be called Statistics.
- (3) **Affected by Multiplicity of Causes** – Statistics are not affected by any single factor but it is affected by many factors. For instance 30% rise in prices may have been due to several causes like reduction in supply, increase in demand, shortage of power, rise in wages, rise in taxes, etc.
- (4) **Reasonable Accuracy** - A reasonable degree of accuracy must be kept in view while collecting statistical data. This accuracy depends on the purpose of investigation, its nature, size and available resources.
- (5) **Pre-determined Purpose** - Statistics are collected with some pre-determined objective. Any information collected without any definite purpose will only be a numerical value and not Statistics. If data pertaining to the farmers of a village is collected, there must be some pre-determined objective. Whether the statistics are collected for the purpose of knowing their economic position or distribution of land among them or their total population. All these objectives must be pre – determined.
- (6) **Collected in a Systematic Manner** – Statistics should be collected in a systematic manner. Before collecting the data, a plan must be prepared. No conclusion can be drawn from data collected in haphazard manner. For instance, data regarding the marks secured by the students of a college without any reference to the class, subject, examination, or maximum marks, etc will lead no conclusion.

Statistics in Singular Sense

In a singular sense, statistics means science of statistics or statistical methods. **It refers to techniques or methods relating to collection, classification, presentation, analysis and interpretation of quantitative data.**

Definition

----- Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data. ----- **CROXTON AND COWDEN**

----- Statistics is the science which deals with the collection, classification and tabulation of numerical facts as a basis for the explanation, description and comparison of phenomena. ----- **LOVITT**

Subject Matter of Statistics

Subject matter of statistics includes two components:

1. Descriptive Statistics

2. Inferential Statistics

1. Descriptive Statistics: Descriptive Statistics refers to those methods which are used for the collection, presentation as well as analysis of data. These methods relate measurement of central tendencies, measurement of dispersion, measurement of correlation etc. **For Example:** Descriptive statistics is used when you estimate average height of the secondary students in your school. Descriptive statistics is also used when you find the marks in science and mathematics of the students in all classes are intimately related to each other.

2. Inferential Statistics: Inferential Statistics refers to all such methods by which conclusion are drawn related to the universe or population on the basis of a given sample. **For example:** If your class teacher estimate average weight of the entire class on the basis of average weight of only a sample of sample of students of the class then we use the inferential statistics.

Important terminology in statistics

- 1. Population:** By population we mean a well defined set or group of all the objects for a particular study. The objects may be persons, plants, books, fishes in ponds, shops etc. the population will consist of certain elements like the plants of a certain kind in a specified field, the fishes in a pond, the unemployed person in India, books in library and so on. For instance, if we want to study the properties of students in a school then the population consists of all the students of school. For instance if we want to study about the books in a library then the population includes all the books of the library etc. if the number of elements are limited then the population is finite. On the other hand if the

number of elements is not limited then the population is infinite. Mostly we deal with finite population.

2. **Sample:** It is a part of the population selected by some sampling procedure. The process of selection of sample is known as sampling. The number of objects in the sample is called the size of the sample. It is believed that a sample is best representative of the population.

For instance, suppose a research worker is required to study the weight of fishes in a pond after a particular period of growth. For this purpose suppose that there are 3,000 fishes in the pond, he may either measure the weight of all the fishes in the pond or he may decide to select a small group of fishes and measure their weights. The first approach of measuring the weight of all fishes is called **complete enumeration or census**. Another approach in which only a small group of fishes is considered is called **sample survey**. In brief we can say that in complete enumeration, information is collected on all the units of the universe and in sample survey, only a part of the universe is considered.

3. **Variable:** A property of objects is known as variable which differ from object to object and is expressible numerically, in terms of numbers.

For instance: the marks in Mathematics of students in a class can be expressed in the term of marks obtained by the students. So it is a variable property which is expressible quantitatively.

4. **Attribute:** A property and characteristic of objects is known as attribute which are not expressible quantitatively in number. We can express the data qualitatively. For example, smoking, color, honesty etc.

CHARACTERISTICS OF STATISTICS

1. Statistics are aggregate of facts.
2. Statistics are numerically expressed.
3. Statistics are affected to a marked extent by multiplicity of causes.
4. Statistics are either enumerated or estimated with reasonable standard of accuracy
5. Statistics are collected in a systematic manner.
6. Statistics are collected for a pre-determined purpose.
7. Statistics should be placed in relation to each other.

In the absence of the above characteristics numerical data can't be called Statistics and hence "all statistics are numerical statements of facts but all numerical statements of facts are not statistics."

According to above Definitions, Statistics is both a science and an art. It is related to the study and application of the principles and methods applicable in the **collection**,

presentation, analysis, interpretation and forecasting of data. Or statistical facts influenced by several factors and related to any area of knowledge or research so that concrete and intelligent decisions may be taken in the phase of uncertainty

NATURE OF STATISTICS

Statistics as a science: science refers to a systematized body of knowledge. It studies cause and effect relationship and attempts to make generalizations in the form of scientific principles or laws. “Science, in short, is like a light house that gives light to the ships to find out their own way but doesn’t indicate the direction in which they should go.” Like other sciences, Statistical Methods are also used to answer the questions like, how an investigation should be conducted. In what way the valid and reliable conclusions can be drawn? Statistics is called the science of scientific methods.

In words of **Croxton and Cowden**, “Statistics is not a science, it is scientific methods.” According to **Tippet**, “as science, the statistical method is a part of the general scientific method and is based on the same fundamental ideas and processes.”

Statistics as an art: we know that science is a body of systematized knowledge. How this knowledge is to be used for solving a problem is work of an art. An art is an applied knowledge. It refers the skill of handling facts so as to achieve a given objective. It is concerned with ways and means of presenting and handling data, making inferences logically and drawing relevant conclusion. Art aspects of statistics tell, ‘how to use statistical rules and principles to study the problems and finding their solutions. ‘Collections of statistics (data) its use and utility are itself an art.

Statistics is both science and art: After studying science and art aspects of statistics, it is used not only to gain knowledge but also to understand the facts and draw important conclusions from it. If science is knowledge, then art is action. Looking from this angle statistics may also be regarded as an art. It involves the application of given methods to obtain facts, derive results and finally to use them for devising action.

STAGES IN A STATISTICAL INVESTIGATION 5 stages -



- 1. Collection:** This is the primary step in a statistical study and data should be collected with care by the investigator. If data are faulty, the conclusions drawn can never be reliable. The data may be available from existing published or unpublished sources or

else may be collected by the investigator himself. The first hand collection of data is one of the most difficult and important tasks faced by a statistician.

2. **Organization:** Data collected from published sources are generally in organized form. However, a large mass of figures that are collected from a survey frequently needs organization. In organizing, there are 3 steps as

(A) Editing (B) Classify (C) Tabulation.

- (A) Editing:** The collected data must be editing very carefully so that the omissions, inconsistencies irrelevant answers and wrong computation in the returns from a survey may be corrected or adjusted.

- (B) Classify:** Classification is the process of arranging the data according to some common characteristics possessed by the items constituting the data.

- (C) Tabulation:** To arrange the data in columns and rows.

Hence collected data is organized properly so that the desire information may be highlighted and undesirable information avoided.

3. **Presentation:** Arranged data is not capable to influence a layman. Thus, it is necessary that data may be presented with the help of tables, diagrams and graphs. By these devices facts can be understood easily.
4. **Analysis:** A major part of it is developed to the methods used in analyzing the presented data, mostly in a tabular form. For this analysis, a number of statistical tools are available, such as averages, correlation, regression etc.
5. **Interpretation:** the interpretation of a data is a difficult task and necessitates a high degree of skills and experience in the statistical investigation because certain decisions made on the basis of conclusions drawn.

SCOPE OF STATISTICS

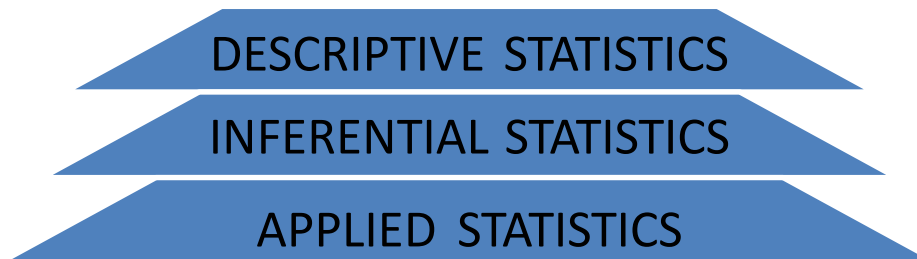
In early stages, the scope of statistics was very limited. It was confined mainly to the administration of government and was, therefore, called the 'Science of Kings'. But in modern time, the scope of statistics has widened usually all those facts come in the purview of statistics, which are expressed in quantitative terms directly or indirectly. That is why **Croxton & Cowden observed**, "Today there is hardly a phase of endeavor which does not find statistical devices at least occasionally useful." It is not unfair to say, **science without statistics bears no fruit and statistics without science have no root.**" The applications of statistics are so numerous that it is often remarked, "Statistics is what statisticians do." Now let us examine a few fields or areas in which statistics is applied.

- 1. Statistics and the State:** in recent years the functions of the state have increased tremendously. The concept of the state has changed from that of simply maintaining law and order to that of a welfare state. Statistical data and statistical methods are of great help in promoting human welfare. The government in most countries is the biggest collector and user of statistical data. These statistics help in framing suitable policies.
- 2. Statistics in Business and Management:** with growing size and increasing competition, the problems of business enterprises have become complex. Statistics is now considered as an indispensable tool in the analysis of activities in the field of business, commerce and industry. The object can be achieved by properly conducted market survey and research which greatly depends on statistical methods. The trends in sales and production can be determined by statistical methods like time-series analysis which are essential for future planning of the phenomena. Statistical concepts and methods are also used in controlling the quality of products to satisfaction of consumer and the producer. The bankers use the objective analysis furnished by statistics and then temper their decisions on the basis of qualitative information.
- 3. Statistics and Economics:** R.A.Fisher complained of “the painful misapprehension that statistics is a branch of economics.” Statistical Data and methods are of immense help in the proper understanding of the economic problems and in the information of economic policies. In the field of exchange, we study markets, law of prices based on supply and demand, cost of production, banking and credit instruments etc. The development of various economic theories owe greatly to statistical methods, e.g., ‘Engel’s law of family expenditure’, ‘Malthusian theory of population’. The impact of mathematics and statistics has led to the development of new disciplines like ‘Econometrics’ and ‘Economic Statistics’. In fact, the concept of planning so vital for growth of nations would not have been possible in the absence of data and proper statistical analysis.
- 4. Statistics and Psychology and Education:** Statistics has found wide application in psychology and education. Statistical methods are used to measure human ability such as; intelligence, aptitude, personality, interest etc. by tests. Theory of learning is also based on Statistical Principles. Applications of statistics in psychology and education have led to the development of new discipline called ‘Psychometric’.
- 5. Statistics and Natural science;** Statistical techniques have proved to be extremely useful in the study of all natural sciences like biology, medicine, meteorology, botany etc. for example- in diagnosing the correct disease the doctor has to rely heavily on factual data like temperature of the body, pulse rate, B.P. etc. In botany- the study of plant life, one has to rely heavily on statistics in conducting experiments about the plants, effect of temperature, type of soil etc. In agriculture- statistical techniques like ‘analysis of variance’ and ‘design of experiments’ are useful for isolating the role of manure, rainfall, watering process, seed quality etc. In fact it is difficult to find any scientific activity where statistical data and methods are not used.

6. **Statistics and Physical Science:** The physical sciences in which statistical methods were first developed and applied. It seems to be making increasing use of statistics, especially in astronomy, chemistry, engineering, geology, meteorology and certain branches of physics.
7. **Statistics and Research;** statistics is indispensable in research work. Most of the advancement in knowledge has taken place because of experiments conducted with the help of statistical methods. Statistical methods also affect research in medicine and public health. In fact, there is hardly any research work today that one can find complete without statistical methods.
8. **Statistics and Computer:** The development of statistics has been closely related to the evolution of electronic computing machinery. Statistics is a form of data processing a way of converting data into information useful for decision-making. The computers can process large amounts of data quickly and accurately. This is a great benefit to business and other organizations that must maintain records of their operations. Processing of raw data is extensively required in the application of many statistical techniques.

CLASSIFICATION OF STATISTICS

Statistics can be divided into 3 parts;



1. **Descriptive Statistics:** Descriptive statistics is related to numerical data or facts. Such data are collected either by counting or by some other process of measurement. It is also related to those methods, includes editing of data, classification, tabulation, diagrammatic or graphical presentation, measures of central tendency, measures of dispersion, correlation etc., help to make the description of numerical facts simple, systematic, synoptic understandable and meaningful.
2. **Inferential Statistics:** Inferential statistics help in making generalizations about the population or universe on the basis of study of samples. It includes the process of drawing proper and rational conclusion about the universe. Among these methods, probability theory and different techniques of sampling test are important.
3. **Applied Statistics;** It involves application of statistical methods and techniques to the problems and actual facts. For example-statistics related to national income, industrial and agricultural production, population, price etc. are called applied statistics. It can be divided into 2 parts-(1) Descriptive Applied Statistics- it deals with the study of the dat

which are known and which naturally relate. Its main object is to provide descriptive information either to the past or to the present for any area. For example- price index number and vital statistics comes under the category of descriptive applied statistics. (2) Scientific Applied Statistics- under this branch of statistical science, statistical methods are used to formulate and verify scientific laws. For example-an effort is made by an economist to establish the law of demand, quantitative theory of money, trade circle etc. These are established and verify by the help of scientific applied statistics.

IMPORTANCE OF STATISTICS

In recent days, we hear talking about statistics from a common person to highly qualified person. It only show that how statistics has been intimately connected with wide range of activities in daily life. They realize that work in their fields require some understanding of statistics. It indicates the importance of the statistics. **A.L.Bowley** says, “Knowledge of statistics is like knowledge of foreign language or of algebra. It may prove of use at any time under any circumstances”.

- 1. Importance to the State or Government;** In modern era, the role of state has increased and various governments of the world also take care of the welfare of its people. Therefore, these governments require much greater information in the form of numerical figures. Statistics are extensively used as a basis for government plans and policies. For example-5-years plans are framed by using reliable statistical data of different segments of life.
- 2. Importance in Human Behavior;** Statistical methods viz., average, correlation etc. are closely related with human activities and behavior. For example-when a layman wishes to purchase some article, he first enquiries about its price at different shops in the market. In other words, he collects data about the price of a particular article and aims at getting idea about the average of the prices and the range within which the price vary. Thus, it can be concluded that statistics play an important role in every aspect of human activities and behavior.
- 3. Importance in Economics;** Statistics is gaining an ever increasing importance in the field of economics. That is why Tugwell said, “The science of economics is becoming statistical in its method.” Statistics and economics are so interrelated to each other that the new disciplines like econometrics and economic statistics have been developed. Inductive method of generalization used in economics, is also based on statistical principle. There are different segments of economics where statistics are used-

(A) Consumption- By the statistics of consumption we can find the way in which people in different group spend their income. The law of demand and elasticity of demand in the field of consumption are based on inductive or inferential statistics.

(B) Production- By the statistics of production supply is adjusted according to demand. We can find out the capital invested in different productive units and its output. The decision

about what to produce, how much to produce, when to produce is based on facts analyzed statistically.

(C) Distribution- Statistics play a vital role in the field of distribution. We calculate the national income of a country by statistical methods and compare it with other countries. At every step we require the help of figures without them. It is difficult to move and draw inferences.

4. Importance in Planning; for the proper utilization of natural and manual resources, statistics play a vital role. Planning is indispensable for achieving faster rate of growth through the best use of a nation's resources. Sometimes said that, "Planning without statistics is a ship without rudder and compass." For example- In India, a number of organizations like national sample survey organization(N.S.S.O.), central statistical organization (C.S.O.) are established to provide all types of information.

5. Importance in Business: The use of statistical methods in the solution of business problems dates almost exclusively to the 20th century. Or now days no business, large or small, public or private, can prosper without the help of statistics. Statistics provides necessary techniques to a businessman for the formulation of various policies and planning with regard to his business. Such as-

(A) Marketing- In the field of marketing, it is necessary first to find out what can be sold and then to evolve a suitable strategy so that goods reach the ultimate consumer. A skillful analysis of data on population, purchasing power, habits of people, competition, transportation cost etc. should precede any attempt to establish a new market.

(B) Quality Control- To earn the better price in a competitive market, it is necessary to watch the quality of the product. Statistical techniques can also be used to control the quality of the product manufactured by a firm. Such as - Showing the control chart.

(C) Banking and Insurance Companies- banks use statistical techniques to take decisions regarding the average amount of cash needed each day to meet the requirements of day to day transactions. Various policies of investment and sanction of loans are also based on the analysis provided by statistics.

(D) Accounts writing and Auditing- Every business firm keeps accounts of its revenue and expenditure. Statistical methods are also employed in accounting. In particular, the auditing function makes frequent application of statistical sampling and estimation procedures and the cost account uses regression analysis.

(E) Research and Development- Many business organizations have their own research and development department which are responsible for collection of such data. These departments also prepare charts groups and other statistical analysis for the purpose.

FUNCTIONS OF STATISTICS

Statistics performs the functions of making the numerical aspects of facts simple, precise, comparable and reliable. In fact, the various functions performed by statistics are the basis of its

utility. **R.W. Burgess** says, “The fundamental gospel of statistics is to push back the domain of ignorance, prejudice, rule of thumb, arbitrary and premature decisions, tradition & dogmatism and to increase the domain in which decisions are made. Principles are formulated on the basis of analyzed quantitative facts.”

- 1. Numerical and definite expression of facts:** The first function of the statistics is the collection and presentation of facts in numerical form. We know that the numerical presentation helps in having a better understanding of the nature of a problem. One of the most important functions of statistics is to present general statements in a precise and definite form. Statements and facts conveyed in exact quantitative terms are always more convincing than vague utterances.
- 2. Simplifies the data (condensation):** Not only does statistics present facts in a definite form but it also helps in condensing mass of data into a few significant figures. According to A.E.Waugh, “the purpose of a statistical method is to simplify great bodies of numerical data.” In fact, human mind cannot follow the huge, complex and scattered numerical facts. So these facts are made simple and precise with the help of various statistical methods like averages, dispersion, graphic or diagrammatic, presentation, classification, tabulation etc. so that a common man also understand them easily.
- 3. Comparison of facts: Baddington states,** “The essence of the statistics is not only counting but also comparison.” The function of comparison does help in showing the relative importance of data. For example- the pass % of examination result of a college may be appreciated better when it is compared with the result of other college or the results of previous years of the same college.
- 4. Establishment of relationship b/w two or more phenomena;** to investigate the relationship b/w two or more facts is the main function of statistics. For example- demand and supply of a certain commodity, prices and wages, temperature and germination time of seeds are interrelated.
- 5. Enlarges individual experiences: In word of Bowley,** “the proper function of statistics indeed is to enlarge individual experience.” Statistics is like a master key that is used to solve problems of mankind in every field. It would not be exaggeration to say that many fields of knowledge would have remained closed to the mankind forever but for the efficient and useful techniques and methodology of the science of statistics.
- 6. Helps in the formulation of policies:** statistics helps in formulating policies in different fields, especially in economic, social and political fields. The government policies like industrial policy, export-import policies, taxation policy and monetary policy are determined on the basis of statistical data and their movements, plan targets are also fixed with the help of data.
- 7. Helps in forecasting:** statistical methods provide helpful means in estimating the available facts and forecasting for future. Here *Bowley's statement* is relevant that, “a statistical estimate may be good or bad, accurate or the reverse; but in almost all cases it is likely to be more accurate than a casual observer's impression.”

- 8. Testing of hypothesis:** statistical methods are also employed to test the hypothesis in theory and discover newer theory. For example-the statement that average height of students of college is 66 inches is a hypothesis. Here students of college constitute the population. It is possible to test the validity of this statement by the use of statistical techniques.

LIMITATIONS OF STATISTICS

Newsholme states, “Statistics must be regarded as an instrument of research of great value but having several limitations which are not possible to overcome and as such they need out careful attention.”

- 1. Statistics does not study qualitative facts:** Statistics means aggregate of numerical facts. It means that in statistics only those phenomena are studied which can be expressed in numerical terms directly or indirectly. Such as- (1) directly in numerical terms like age, weight and income of individual (2) no directly but indirectly like intelligent of students and achievements of students (3) neither directly nor directly like morality, affection etc. such type of facts don't come under the scope of statistics.
- 2. Statistics doesn't study individual:** According to **W.I.King**, “Statistics from their very nature of subject cannot and will never be able to take into account individual causes. When these are important, other means must be used for their study.” These studied are done to compare the general behavior of the group at different points of time or the behavior of different groups at a particular point of time.
- 3. Statistical results are true only on the average:** The statistical laws are not completely true and accurate like the law of physics. For example – law of gravitational forces is perfectly true & universal but statistical conclusions are not perfectly true. Such as the average age of a person in India is 62 years. It does not mean that every person will attain this age. On the basis of statistical methods we can say only in terms of probability and not certainty.
- 4. Statistics as lack of complete accuracy:** According to Conner, “Statistical data must always be treated as approximations or estimates and not as precise measurements.” Statistical result are based on sample or census data, are bound to be true only approximately. For example – according to population census 2001, country's population is 1,02,70,15,247 but can real population may not be more or less by hundred, two hundred and so on.
- 5. Statistics is liable to be misused:** Statistical deals with figures and it can be easily manipulated, distorted by the inexpert and unskilled persons it is very much likely to be misused in most of the cases. In other words, the data should be handled by experts. Thus it must be used by technically sound persons.

6. **Statistics is only one of the methods of studying a phenomenon;** According to Croxton & Cowden, “It must not be assumed that the statistical method is the only method to be used in research; neither this method be considered the best attack for every problem.” The conclusions arrived at with the help of statistics must be supplemented with other evidences.
7. **Statistical results may be misleading;** Without any reference, statistical results may provide doubtful conclusions. For example – on the basis of increasing no. of prisoners in the prison, it may be conclude that crime is increasing. But it may be possible that due to rude behavior of police administration the number of prisoners is increasing but crime is decreasing.

Therefore, it is worth-mentioning that every science based on certain assumption and limitations. This does not reduce the importance of the subject but lays emphasis on the fact that precautions should be taken while dealing with statistical analysis and interpretations.

DISTRUST OF STATISTICS

For practical view point statistics is very useful and important science. We know that utility of statistics lies not merely in data but in correct analysis and proper interpretation of data. Several times due to ignorance and bias, people misuse this delicate tool of knowledge and it creates distrust about data.

Opinion about distrust of statistics;

1stOpinion–Statisticians fully trust on the statistical conclusions because data is collected, edited, analyzed and interpreted on the basis of statistical methods. Thus there is no reason to doubt on it and said that “Figures don’t lie” or “Figures can prove nothing.”

2nd Opinion –The statistics is looked upon with a suspicious eye and is quite often condemned as “Figures are tissue of flesh hood. **Discardi remarks** that there are three kinds of lies- lies, damned lies and statistics or “There are black lies, white lies, multi-chromatic lies and statistics is rainbow of lies.”

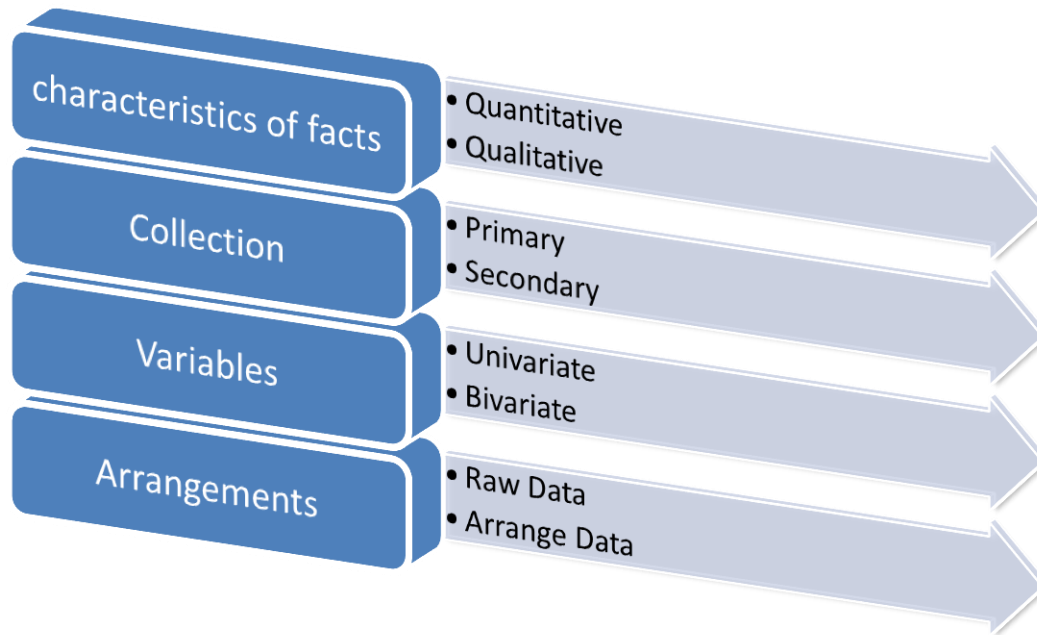
Many persons feel that data are false, confusing and incorrect and with their help truth can be proved wrong and lies can be put as truth. Hence it is said that “Statistics can prove anything” or “Statistics are like clay of which you can make God or Devil, as you please.”

In this context, the observation is worth quoting that “Statistician is the person who is deeply involved in statistical data. He can freely play with them, misuse them and can cheat common people. So he is just magician who shows the games of tricks of hand through statistical data. His result can be surprising but not trustworthy.

TYPES OF DATA

Data are the foundation stones and basic raw material in relation to any statistical investigation that can be counted, classified, measured or quantified.

Types of Data are following;



ON THE BASIS OF CHARACTERISTICS OF FACTS

Data may be divided into two types;

- 1. Quantitative Data or Numerical Data:** These types of data can be measured directly such as age, income, production, marks etc. those facts are called variables and variables may be discrete or continuous.

Discrete variable– Those variables whose values are individually distinct and discontinuous. There is a definite difference between two variables. According to Boddington, “Discrete variables is one where the variables (Individual values) differ from each other by definite

amounts.” For example – number of students of a class, number of children in a family, number of cattle’s etc. It takes integral values such as 0, 1, 2, 3, 4 ...etc.

Continuous variable – A continuous variable is one which assumes all values within an interval. That is no definite breaks are visible in this type of series. For example – age, weight, height.....

Questions; State which of the following represents Discrete data or Continuous data?

- I. No. of accidents on each day in a month
- II. Lengths of 1,000 bolts produced in a factory
- III. Speed of an automobile in kilometer per hour
- IV. No. of books on a library shelf

2. Qualitative Data or Categorical Data: They include data relating to such facts which can’t be measured directly but are counted or categorized to the basis of attributes such as literates, illiterates, unemployed, honest etc. are called attributes. For example- population can be classified on the basis of males and females or males may be classified on the basis of marital status, i.e. married or unmarried. Qualitative Data may further be classified into two categories

ON THE BASIS OF VARIABLES

On the basis of variables, also data may be of two types;

- (1) **Univariate Data:** When the frequencies are determined on the basis of one variable. For example – no. of workers on the basis of wages, no. of persons on the basis of age etc.
- (2) **Bivariate Data:** When the data are edited or presented on the basis of two variables simultaneously. For this two-way frequency table is constructed, one variable is placed horizontally and the second one vertically. For example – to present the number of students in one table on the basis of marks obtained in two subjects, to tabulate the no. of persons in one table on the basis of two variables i.e. height and weight.

ON THE BASIS OF ARRANGEMENTS

Data may be categorized into two types;

- (1) **Raw Data:** When the data is arranged and analyzed. It is called ‘Raw’ because it is unprocessed by statistical methods.
- (2) **Arrange Data:** When the data is processed and is arranged, summarized, classified and tabulated in proper way.

Terms like 'Data Point' and 'Data Set' are also used in order to distinguish between the numbers relating to individual or single facts and the aggregate of facts. For example– the data of production of sugar for ten years will be termed as 'Data Set' and the figures for production of one year will be as 'Data Point'.

CLASSIFICATION

After collection and editing of data the first step towards further processing the same is classification. Classification is a process in which the collected data are arranged in separate classes, groups or subgroups according to their characteristics. According to Secrist, "Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different classes."

It concludes that classification means the arrangements and systematization of data into different classes and these classes are determined on the basis of nature, objectives and scope of the enquiry.

OBJECTIVES OF CLASSIFICATION

Classification is a method or technique for extracting the essential information supplied by the raw data.

- (1) **To condense the data:** the main objective of classification is to condense and simplify the statistical material, so that the same may be easily understandable.
- (2) **To bring out points of similarities and dissimilarities of data:** classification brings out clearly the points of similarity and dissimilarities of statistical facts because data of similar characteristics are placed in one class i.e., males and females, literates and illiterates, married and unmarried etc.
- (3) **To make facts comparable:** by arranging the data according to the points of similarity and dissimilarities, it helps in comparison.
- (4) **To bring out relationship:** classification helps in finding cause and effect relationship in the data. For example- based on literacy and criminal tendency of a group peoples, it can be established whether literacy has any impact on criminal tendency or not.
- (5) **To prepare ground for tabulation:** tabulation is the basis of statistical analysis and classification is the basis for tabulation.

It concludes that classification occupies an important place in the process of statistical investigation. The fact is that the process of tabulation, presentation and analysis can't even be shorted without classification.

CHARACTERISTICS OR RULES FOR A GOOD CLASSIFICATION

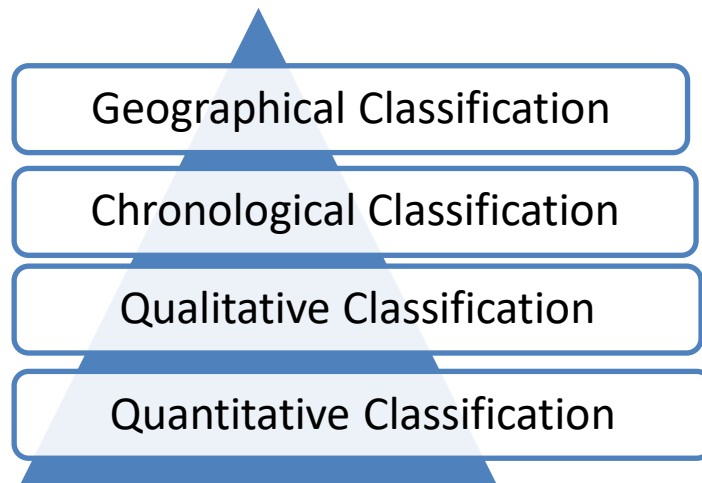
- (1) **Unambiguity:** the various classes should be so defined that there is no room for doubt and confusion. For example–population is classified as literates or illiterates.
- (2) **Exhaustive and mutually exclusive:** classification should be so exhaustive (clear all aspects) and one item may not be find place in more than one class. For example – students of a

college are classified into three groups – urban, rural and hostlers. This classification is not mutually exclusive because among hostlers some may be urban and some other rural.

- (3) **Stability:** the classification of data into various classes must be stable over a period of time of investigation.
- (4) **Suitability:** the classification should conform to the objectives of enquiry. For example–to study the relationship between sex and university education, there is no need to classify on the basis of age and religion.
- (5) **Flexibility:** a good classification should be flexible so that adjustments may be easily made in classes according to changed situations. An ideal classification is one that can adjust itself to these changes and yet retains its stability.

METHODS OF CLASSIFICATION

There are 4 methods of classification;



TABULATION

Tabulation is the next step of classification of the data and is designed to summarize lots of information in a simple manner. In common language tabulation is the process of arranging data in a systematic manner in the form of rows and columns. **According to Blair**, “Tabulation in its broadest sense is any orderly arrangement of data in columns and rows.”

OBJECTIVES OF TABULATION

1. To simplify complex data
2. To facilitate comparison
3. To economise Space
4. To facilitate presentation
5. Help in analysis of data
6. To help in reference

DIFFERENCE BETWEEN CLASSIFICATION & TABULATION

Basis of Difference	Classification	Tabulation
Presentation	Classifies into different classes.	Classifies into row and columns.
Sequence	First step	Second step
Methods	Method of statistical Analysis.	Method of data Presentation
Use of data	Original data are used.	Derivatives like percentages, coefficients Proportion, etc. may also be used.

FREQUENCY DISTRIBUTION

The tabular arrangement of data showing the frequency of each item is called a **frequency distribution**. According to Croxton and Cowden, "Frequency distribution is a statistical table in which different values of variable are shown in the sequence of magnitude along with corresponding frequencies."

TYPES OF FREQUENCY DISTRIBUTION

- (1) **Discrete frequency distribution:** It is a discontinuous frequency distribution, where observations are independent to each other. Each observation is different and separates from others. Example – the no. of two children in 20 families;

1,1,2,3,4,3,2,1,1,4,5,2,4,2,2,1,3,3,2,5

Construction of discrete frequency distribution

1. A table with three columns is prepared.
2. In 1st column of this table all possible values of the variable are placed.
3. In the 2nd column, tally marks or tally bars are put, keeping in view the repetition of each value.
4. After putting tally marks, they are counted and this counting is shown in the 3rd column, entitled frequency.

No. of children	Tally Bars	Frequency
1		5
2		6
3		4
4		3
5		2

(2) **Continuous frequency distribution:** A continuous frequency distribution is such a distribution in which data are arranged in classes or groups which are not exactly measureable. Groups or class-intervals are always in a continuous form from the beginning of the frequency distribution, till the end, within a given range of the data. For example-

**39,25,5,33,19,21,12,48,13,21,9,1,10,
9,8,12,17,40,12,46,37,17,27,30,6,2,23**

Construction of continuous frequency distribution

- (1) Find out the range = maximum marks – minimum marks
- (2) Class size = decide the class size
- (3) Width of class interval = $\frac{\text{range}}{\text{class size}}$
- (4) Mid value = $\frac{\text{upper class limit} + \text{lower class limit}}{2}$
- (5) Class frequency = the no. of observations corresponding to a particular class is known as frequency of the class.

Marks	Tally Bars	No. of Students
0 – 10		7
10 – 20		8
20 – 30		5
30 – 40		4
40 – 50		3

There are two types of series according to class interval;

- (1) **Inclusive form;** A frequency distribution in which each upper limit of each class is also included. Such as; 0-9, 10-19, 20-29.....
- (2) **Exclusive form;** In which the upper limit of the next class-interval. Such as; 0-10, 10-20, 20-30.....

MEASURES OF CENTRAL TENDENCY

OBJECTIVES

After going through this unit, you will learn:

- the concept and significance of measures of central tendency
- to compute various measures of central tendency, such as arithmetic mean, median, mode and quartiles
- the relationship among various averages.

INTRODUCTION

The objective here is to find one representative value which can be used to locate and summarise the entire set of varying values. This one value can be used to make many decisions concerning the entire set. We can define measures of central tendency (or location) to find some central value around which the data tend to cluster.

SIGNIFICANCE OF MEASURES OF CENTRAL TENDENCY

Measures of central tendency i.e. condensing the mass of data in one single value, enable us to get an idea of the entire data. For example, it is impossible to remember the individual incomes of millions of earning people of India. But if the average income is obtained, we get one single value that represents the entire population.

Measures of central tendency also enable us to compare two or more sets of data to facilitate comparison. For example, the average sales figures of April may be compared with the sales figures of previous months.

PROPERTIES OF A GOOD MEASURE OF CENTRAL TENDENCY

- [1] It should be easy to understand and calculate.
- [2] It should be rigidly defined.
- [3] It should be based on all observations.
- [4] It should be least affected by sampling fluctuation.
- [5] It should be capable of further algebraic treatment.
- [6] It should be least affected by extreme values.
- [7] It should be calculated in case of open end interval.

Following are some of the important measures of central tendency which are commonly used in business and industry.

- Arithmetic Mean
- Weighted Arithmetic Mean
- Median
- Quantiles (quartiles, deciles and percentiles)
- Mode
- Geometric Mean
- Harmonic Mean

ARITHMETIC MEAN

The arithmetic mean (or mean or average) is the most commonly used and readily understood measure of central tendency. In statistics, the term average refers to any of the measures of central tendency.

Ungrouped data/Raw data

The arithmetic mean is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations. Symbolically, it can be represented as:

$$\mu = \frac{\sum Xi}{N}$$

where,

$\sum X$ indicates the sum of the values of all the observations, and N is the total number of observations.

For example, let us consider the monthly salary (Rs.) of 10 employees of a firm x

2500, 2700, 2400, 2300, 2550, 2650, 2750, 2450, 2600, 2400

If we compute the arithmetic mean, then $2500+2700+2400+2300+2550+2650+2750+2450+2600+2400 = 25300$

Mean=25300 /10= Rs. 2530.

Therefore, the average monthly salary is Rs. 2530.

Discrete data

When the observations are classified into a frequency distribution, Therefore, for discrete data; the arithmetic mean is defined as

$$\mu = \frac{\sum fXi}{N}$$

Where, f is the frequency for corresponding variable x and N is the total frequency, i.e. $N = \sum f$.

X	f	fx
10	12	120
20	23	460
30	35	1050
40	47	1880
50	38	1900
60	29	1740
70	16	1120
Sum	200	8270

Mean=8270/200= 41.35

Continuous Data

When the observations are classified into a frequency distribution, Therefore, for grouped data; the arithmetic mean is defined as

$$\mu = \frac{\sum fXi}{N}$$

Where X is midpoint of various classes, f is the frequency for corresponding class and N is the total frequency, i.e. $N = \sum f$.

This method is illustrated for the following data which relate to the monthly sales of 200 firms. the midpoint of the class interval would be treated as the representative average value of that class.

Monthly Sales (Rs. Thousand)	Mid point X	No. of firms f	fX
300-350	325	5	1625
350-400	375	14	5250
400-450	425	23	9775
450-500	475	50	23750
500-550	525	52	27300
550-600	575	25	14375
600-650	625	22	13750
650-700	675	7	4725
700-750	725	2	1450
Sum		200	10200

Mean=102000/200=510

MERITS OF MEAN

- [1] It is easy to understand and calculate.
- [2] It is rigidly defined.
- [3] It is based on all observations.
- [4] It is least affected by sampling fluctuation.
- [5] It is capable of further algebraic treatment.

DEMERITS OF MEAN

- 1) It is highly affected by extreme values.
- 2) It is not calculated in case of open end interval.

MEDIAN

A second measure of central tendency is the median. Median is that value which divides the distribution into two equal parts. Fifty per cent of the observations in the distribution are above the value of median and other fifty per cent of the observations are below this value of median. The median is the value of the middle observation when the series is arranged in order of size or magnitude (Ascending order).

UNGROUPEd DATA

If the number of observations is **odd**, then the median is equal to one of the original observations (Middle).

$$\text{Median} = \left(\frac{N+1}{2}\right) \text{ th value}$$

For example, if the income of seven persons in rupees is 1100, 1200, 1350, 1500, 1550, 1600, 1800, then

$$\text{Median} = ((7 + 1)/2) = 4^{\text{th}} \text{ value}$$

Median =1500

If the number of observations is even, then the median is the arithmetic mean of the two middle observations.

$$\text{Median} = \frac{\left\{\left(\frac{N}{2}\right) + \left(\frac{N}{2} + 1\right)\right\}}{2} \text{ th value}$$

For example, if the income of eight persons in rupees is 1100, 1200, 1350, 1500, 1550, 1600, 1800,1850, then the median income of eight persons would be $1500+1550/2= 1525$

DISCRETE SERIES

First we find cumulative frequency.then locate $(N+1/2)$ the value in cumulative frequency.corresponding that value of x is median.

X	f	cf
10	12	12
20	23	35
30	35	70
40	47	117
50	38	155
60	29	184
70	16	200
Sum	200	

$$N=200$$

$$N+1/2=100.5$$

Median= 40

CONTINUOUS DATA

For continuous data, First we find cumulative frequency. then locate $(N+1/2)$ the value in cumulative frequency. corresponding class interval is median class.the following formula may be used to locate the value of median.

$$\text{Median} = l_1 + \frac{\{(N/2 - cf) * h\}}{f}$$

where l_1 is the lower limit of the median class, cf is the preceding cumulative frequency to the median class, f is the frequency of the median class and h is the width of the median class.

Consider the following data which relate to the age distribution of 1000 workers in an industrial establishment.

The location of median value is facilitated by the use of a cumulative frequency distribution as shown below in the table.

Age (Years)	No. of workers f	Cumulative frequency c.f
Below 25	120	120
25-30	125	245
30-35	180	425(cf)
35-40	160(f)	585
40-45	150	735
45-50	140	875
50-55	100	975
55 and Above	25	1000

$N=1000$

Median Class= $(1000+1)/2=500.5^{\text{th}}=(35-40)$

$l_1 = 35, cf = 425, f = 160, N/2 = 500, h = 5$

Median = $35 + \{(500-425)*5\}/160 = 35 + 375/160 = 35 + 2.34 = \mathbf{37.34}$

MERITS OF MEDIAN

- [1] It is easy to understand and calculate.
- [2] It is rigidly defined.
- [3] It is not affected by extreme values.
- [4] It is calculated in case of open end interval.
- [5] It is located by graphically also.

DEMERITS OF MEDIAN

- [1] It is not based on all observations.
- [2] It is affected by sampling fluctuation.
- [3] It is not capable of further algebraic treatment.

MODE

The mode is the typical or commonly observed value in a set of data. It is defined as the value which occurs most often or with the greatest frequency. The dictionary meaning of the term mode is most usual'.

For example, in the series of numbers 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, the mode is 8 because it occurs the maximum number of times. That means in **ungrouped data** mode can find by inspection only.

DISCRETE DATA

Mode is the value of X which has highest frequency.

For example,

X	f
10	12
20	23
30	35
40	47
50	38
60	29
70	16
Sum	200

Mode=40

CONTINUOUS DATA

First, we find Modal class = corresponding to highest frequency.

$$\text{MODE} = l_1 + \left\{ \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right\} * h$$

where l_1 is lower limit of the modal class, f_1 is the frequency of the modal class, f_0 the frequency of the preceding class, f_2 is the frequency of the succeeding class, h is the size of the modal class.

To illustrate the computation of mode, let us consider the following data.

Age (Years)	No. of workers f
Below 25	120
25-30	125(f_0)
30-35	180(f_1)
35-40	160(f_2)
40-45	150
45-50	140
50-55	100
55 and Above	25

Modal class=(30-35)

$l_1 = 30, f_1 = 180, f_0 = 125, f_2 = 160, h = 5$

$$\text{Mode} = 30 + \left\{ \frac{(180-125)}{(2 \times 180 - 125 - 160)} \right\} \times 5 = 30 + \left(\frac{55}{75} \right) \times 5 = 30 + 3.667 = \mathbf{33.67}$$

MERITS OF MODE

- [1] It is easy to understand and calculate.
- [2] It is not affected by extreme values.
- [3] It is calculated in case of open end interval.
- [4] It is located by graphically also.

DEMERITS OF MODE

- [1] It is not based on all observations.
- [2] It is highly affected by sampling fluctuation.
- [3] It is not capable of further algebraic treatment
- [4] It is not rigidly defined.

RELATIONSHIP AMONG MEAN, MEDIAN AND MODE

$$\text{MODE} = 3\text{MEDIAN} - 2\text{MEAN}$$

QUARTILES

Quartiles are those values which divide the total data into four equal parts. Since three points divide the distribution into four equal parts, we shall have three quartiles. Let us call them Q1, Q2, and Q3. The first quartile, Q1, is the value such that 25% of the observations are smaller and 75% of the observations are larger. The second quartile, Q2, is the median, i.e., 50% of the observations are smaller and 50% are larger. The third quartile, Q3, is the value such that 75% of the observations are smaller and 25% of the observations are larger.

UNGROUPED DATA

If the number of observations is **odd**, then the median is equal to one of the original observations (Middle).

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

$$Q3 = 3 \left(\frac{N+1}{4} \right) \text{ th value}$$

For example, if the income of seven persons in rupees is 1100, 1200, 1350, 1500, 1550, 1600, 1800, 1850 then

$$Q1 = \left(\frac{8+1}{4} \right) = 2.25^{\text{th}} \text{ value}$$

$$Q1 = 1200 + 0.25(1350 - 1200) = 1200 + 37.5 = 1237.5$$

$$Q3 = 3 \left(\frac{8+1}{4} \right) = 6.75^{\text{th}} \text{ value}$$

$$Q3 = 1600 + 0.75(1800 - 1600) = 1600 + 150 = 1750$$

DISCRETE SERIES

First we find cumulative frequency. then locate $(N+1/4)$ and $3(N+1/4)$ the value in cumulative frequency .corresponding that value of x is $Q1$ and $Q2$ respectively.

First quartile

X	f	cf
10	12	12
20	23	35
30	35	70
40	47	117
50	38	155
60	29	184
70	16	200
Sum	200	

Third quartile

$N=200$

$Q1=N+1/4=50.25^{\text{th}}$ value

$Q1 = 30$

$Q3 = 3*(N+1/4)= 150.75^{\text{th}}$ value

$Q3=50$

CONTINUOUS DATA

For continuous data, First we find cumulative frequency. then locate $(N+1/4)$ and $3(N+1/4)$ the value in cumulative frequency. corresponding class interval is first quartile class and third quartile class respectively.the following formula may be used to locate the value of quartiles.

$$Q1=l_1+ \{(N/4 - cf)*h\} /f$$

where l_1 is the lower limit of the first quartile class, cf is the preceding cumulative frequency to the first quartile class, f is the frequency of first quartile class and h is the width of the first quartile class .

$$Q3=l_1+ \{(3N/4 - cf)*h\} /f$$

where l_1 is the lower limit of the third quartile class, cf is the preceding cumulative frequency to the third quartile class, f is the frequency of third quartile class and h is the width of the third quartile class .

Consider the following data which relate to the age distribution of 1000 workers in an industrial establishment.

The location of quartile value is facilitated by the use of a cumulative frequency distribution as shown below in the table.

Age (Years)	No. of workers f	Cumulative frequency c.f
Below 25	120	120
25-30	125	245
30-35	180	425(cf)
35-40	160(f)	585
40-45	150	735
45-50	140	875
50-55	100	975
55 and Above	25	1000

First quartile class

Third quartile class

$$N=1000$$

$$Q1=(1000+1)/4=250.25^{\text{th}}=(25-30)$$

$$l_1=25, cf=120, f=125, N/4=250, h=5$$

$$Q1=25+\{(250-120)*5\}/125=35+390/125=35+3.12=38.12$$

$$Q3=3(1000+1)/4=750.75^{\text{th}}=(45-50)$$

$$l_1=45, cf=735, f=140, 3N/4=750, h=5$$

$$Q3=45+\{(750-735)*5\}/140=45+75/140=45+0.53=45.53$$

1] Following is the cumulative frequency distribution of preferred length of study-table obtained from the preference study of 50 students.

Length	No. of students	Length	No. of students
more than 50 cms	50	more than 90 cms	25
more than 60 cms	46	more than 100 cms	18
more than 70 cms	40	more than 110 cms	7
more than 80 cms	32		

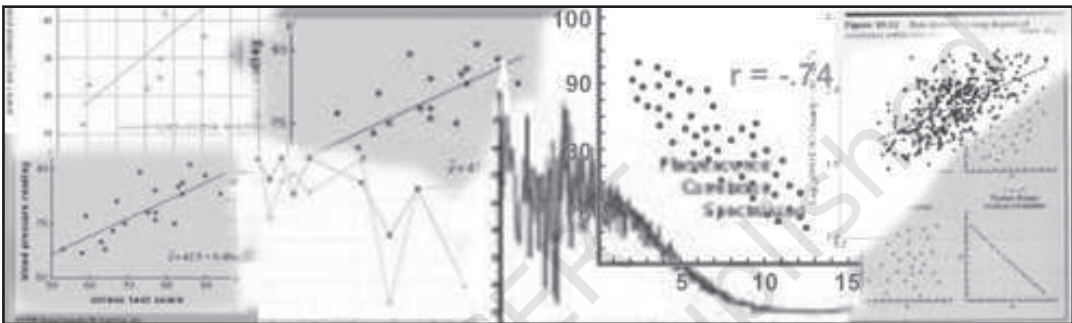
A manufacturer has to take decision on the length of study-table to manufacture. What length would you recommend and why?

2] An incomplete distribution of daily sales (Rs. thousand) is given below. The data relate to 229 days.

Daily sales (Rs. thousand)	No. of days	Daily sales (Rs. thousand)	No. of days
10-20	12	50-60	?
20-30	30	60-70	25
30-40	?	70-80	18

You are told that the median value is 46. Using the median formula, fill up the missing frequencies and calculate the arithmetic mean of the completed data.

Correlation



Studying this chapter should enable you to:

- understand the meaning of the term correlation;
- understand the nature of relationship between two variables;
- calculate the different measures of correlation;
- analyse the degree and direction of the relationships.

1. INTRODUCTION

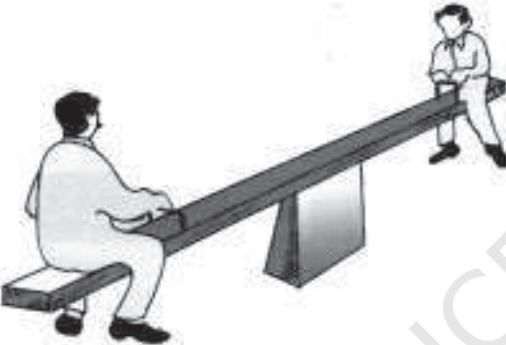
In previous chapters you have learnt how to construct summary measures out of a mass of data and changes among similar variables. Now you will learn how to examine the relationship between two variables.

As the summer heat rises, hill stations, are crowded with more and more visitors. Ice-cream sales become more brisk. Thus, the temperature is related to number of visitors and sale of ice-creams. Similarly, as the supply of tomatoes increases in your local *mandi*, its price drops. When the local harvest starts reaching the market, the price of tomatoes drops from a princely Rs 40 per kg to Rs 4 per kg or even less. Thus supply is related to price. Correlation analysis is a means for examining such relationships systematically. It deals with questions such as:

- Is there any relationship between two variables?



- If the value of one variable changes, does the value of the other also change?



- Do both the variables move in the same direction?



- How strong is the relationship?

2. TYPES OF RELATIONSHIP

Let us look at various types of relationship. The relation between movements in quantity demanded and the price of a commodity is an

integral part of the theory of demand, which you will study in Class XII. Low agricultural productivity is related to low rainfall. Such examples of relationship may be given a cause and effect interpretation. Others may be just coincidence. The relation between the arrival of migratory birds in a sanctuary and the birth rates in the locality cannot be given any cause and effect interpretation. The relationships are simple coincidence. The relationship between size of the shoes and money in your pocket is another such example. Even if relationship exist, they are difficult to explain it.

In another instance a third variable's impact on two variables may give rise to a relation between the two variables. Brisk sale of ice-creams may be related to higher number of deaths due to drowning. The victims are not drowned due to eating of ice-creams. Rising temperature leads to brisk sale of ice-creams. Moreover, large number of people start going to swimming pools to beat the heat. This might have raised the number of deaths by drowning. Thus temperature is behind the high correlation between the sale of ice-creams and deaths due to drowning.

What Does Correlation Measure?

Correlation studies and measures the direction and intensity of relationship among variables. Correlation measures covariation, not causation. Correlation should never be

interpreted as implying cause and effect relation. The presence of correlation between two variables X and Y simply means that when the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change), but in a definite way. For simplicity we assume here that the correlation, if it exists, is linear, i.e. the relative movement of the two variables can be represented by drawing a straight line on graph paper.

Types of Correlation

Correlation is commonly classified into negative and positive correlation. The correlation is said to be positive when the variables move together in the same direction. When the income rises, consumption also rises. When income falls, consumption also falls. Sale of ice-cream and temperature move in the same direction. The correlation is negative when they move in opposite directions. When the price of apples falls its demand increases. When the prices rise its demand decreases. When you spend more time in studying, chances of your failing decline. When you spend less hours in your studies, chances of scoring low marks/grades increase. These are instances of negative correlation. The variables move in opposite direction.

3. TECHNIQUES FOR MEASURING CORRELATION

Three important statistical tools used to measure correlation are scatter diagrams, Karl Pearson's coefficient of correlation and Spearman's rank correlation.

A scatter diagram visually presents the nature of association without giving any specific numerical value. A numerical measure of linear relationship between two variables is given by Karl Pearson's coefficient of correlation. A relationship is said to be linear if it can be represented by a straight line. Spearman's coefficient of correlation measures the linear association between ranks assigned to individual items according to their attributes. Attributes are those variables which cannot be numerically measured such as intelligence of people, physical appearance, honesty, etc.

Scatter Diagram

A scatter diagram is a useful technique for visually examining the form of relationship, without calculating any numerical value. In this technique, the values of the two variables are plotted as points on a graph paper. From a scatter diagram, one can get a fairly good idea of the nature of relationship. In a scatter diagram the degree of closeness of the scatter points and their overall direction enable us to examine the relation-

ship. If all the points lie on a line, the correlation is perfect and is said to be unity. If the scatter points are widely dispersed around the line, the correlation is low. The correlation is said to be linear if the scatter points lie near a line or on a line.

Scatter diagrams spanning over Fig. 7.1 to Fig. 7.5 give us an idea of the relationship between two variables. Fig. 7.1 shows a scatter around an upward rising line indicating the movement of the variables in the same direction. When X rises Y will also rise. This is positive correlation. In Fig. 7.2 the points are found to be scattered around a downward sloping line. This time the variables move in opposite directions. When X rises Y falls and vice versa. This is negative correlation. In Fig. 7.3 there is no upward rising or downward sloping line around which the points are scattered. This is an example of no correlation. In Fig. 7.4 and Fig. 7.5 the points are no longer scattered around an upward rising or downward falling line. The points themselves are on the lines. This is referred to as perfect positive correlation and perfect negative correlation respectively.

Activity

- Collect data on height, weight and marks scored by students in your class in any two subjects in class X. Draw the scatter diagram of these variables taking two at a time. What type of relationship do you find?

A careful observation of the scatter diagram gives an idea of the nature and intensity of the relationship.

Karl Pearson's Coefficient of Correlation

This is also known as product moment correlation and simple correlation coefficient. It gives a precise numerical value of the degree of linear relationship between two variables X and Y. The linear relationship may be given by

$$Y = a + bX$$

This type of relation may be described by a straight line. The intercept that the line makes on the Y-axis is given by a and the slope of the line is given by b . It gives the change in the value of Y for very small change in the value of X. On the other hand, if the relation cannot be represented by a straight line as in

$$Y = X^2$$

the value of the coefficient will be zero. It clearly shows that zero correlation need not mean absence of any type of relation between the two variables.

Let X_1, X_2, \dots, X_N be N values of X and Y_1, Y_2, \dots, Y_N be the corresponding values of Y. In the subsequent presentations the subscripts indicating the unit are dropped for the sake of simplicity. The arithmetic means of X and Y are defined as

$$\bar{X} = \frac{\sum X}{N}; \quad \bar{Y} = \frac{\sum Y}{N}$$

and their variances are as follows

$$\sigma_x^2 = \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum X^2}{N} - \bar{X}^2$$

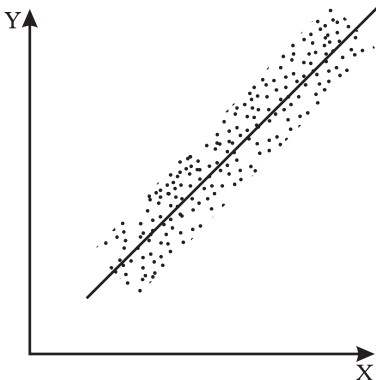


Fig. 7.1: Positive Correlation

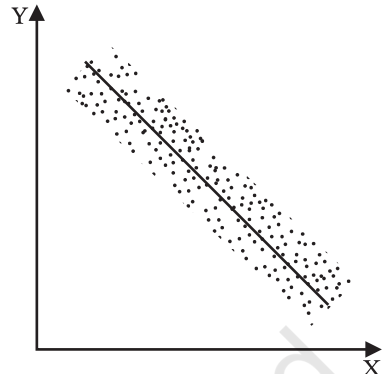


Fig. 7.2: Negative Correlation

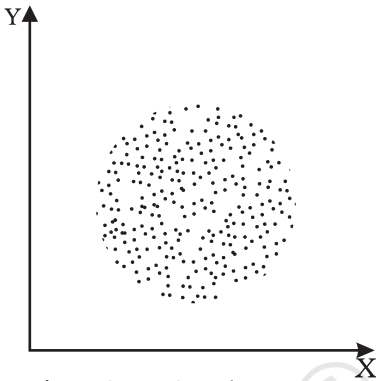


Fig. 7.3: No Correlation

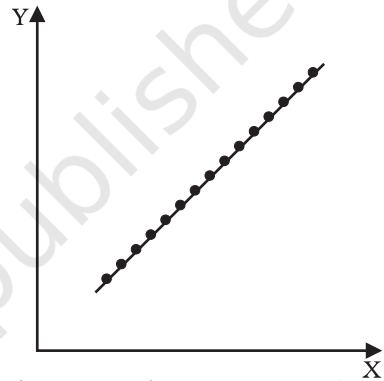


Fig. 7.4: Perfect Positive Correlation

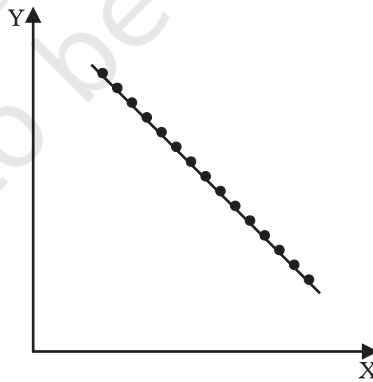


Fig. 7.5: Perfect Negative Correlation

and $\sigma^2_y = \frac{\Sigma(Y - \bar{Y})^2}{N} = \frac{\Sigma Y^2}{N} - \bar{Y}^2$

The standard deviations of X and Y respectively are the positive square roots of their variances. Covariance of X and Y is defined as

$$\text{Cov}(X,Y) = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N} = \frac{\Sigma xy}{N}$$

Where $x = X - \bar{X}$ and $y = Y - \bar{Y}$ are the deviations of the i^{th} value of X and Y from their mean values respectively.

The sign of covariance between X and Y determines the sign of the correlation coefficient. The standard deviations are always positive. If the covariance is zero, the correlation coefficient is always zero. The product moment correlation or the Karl Pearson's measure of correlation is given by

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} \dots(1)$$

or

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2} \sqrt{\Sigma(Y - \bar{Y})^2}} \dots(2)$$

or

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}} \dots(3)$$

or

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \dots(4)$$

Properties of Correlation Coefficient

Let us now discuss the properties of the correlation coefficient

- r has no unit. It is a pure number. It means units of measurement are not part of r . r between height in feet and weight in kilograms, for instance, could be say 0.7.
- A negative value of r indicates an inverse relation. A change in one variable is associated with change in the other variable in the opposite direction. When price of a commodity rises, its demand falls. When the rate of interest rises the demand for funds also falls. It is because now funds have become costlier.



- If r is positive the two variables move in the same direction. When the price of coffee, a substitute of tea, rises the demand for tea also rises. Improvement in irrigation facilities is associated with higher yield. When temperature rises the sale of ice-creams becomes brisk.

- The value of the correlation coefficient lies between minus one and plus one, $-1 \leq r \leq 1$. If, in any exercise, the value of r is outside this range it indicates error in calculation.
- If $r = 0$ the two variables are uncorrelated. There is no linear relation between them. However other types of relation may be there.
- If $r = 1$ or $r = -1$ the correlation is perfect. The relation between them is exact.
- A high value of r indicates strong linear relationship. Its value is said to be high when it is close to $+1$ or -1 .
- A low value of r indicates a weak linear relation. Its value is said to be low when it is close to zero.
- The magnitude of r is unaffected by the change of origin and change of scale. Given two variables X and Y let us define two new variables.

$$U = \frac{X - A}{B}; V = \frac{Y - C}{D}$$

where A and C are assumed means of X and Y respectively. B and D are common factors and of same sign. Then

$$r_{xy} = r_{uv}$$

This property is used to calculate correlation coefficient in a highly simplified manner, as in the step deviation method.

As you have read in chapter 1, the statistical methods are no substitute for common sense. Here, is another example, which highlights the need for understanding the data properly

before correlation is calculated and interpreted. An epidemic spreads in some villages and the government sends a team of doctors to the affected villages. The correlation between the number of deaths and the number of doctors sent to the villages is found to be positive. Normally the health care facilities provided by the doctors are expected to reduce the number of deaths showing a negative correlation. This happened due to other reasons. The data relate to a specific time period. Many of the reported deaths could be terminal cases where the doctors could do little. Moreover, the benefit of the presence of doctors becomes visible only after some time. It is also possible that the reported deaths are not due to the epidemic. A tsunami suddenly hits the state and death toll rises.

Let us illustrate the calculation of r by examining the relationship between years of schooling of farmers and the annual yield per acre.

Example 1

No. of years of schooling of farmers	Annual yield per acre in '000 (Rs)
0	4
2	4
4	6
6	10
8	10
10	8
12	7

Formula 1 needs the value of $\sum xy$, σ_x , σ_y

From Table 7.1 we get,

$$\Sigma xy = 42,$$

$$\sigma_x = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{112}{7}},$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{N}} = \sqrt{\frac{38}{7}}$$

Substituting these values in formula (1)

$$r = \frac{42}{7 \sqrt{\frac{112}{7}} \sqrt{\frac{38}{7}}} = 0.644$$

The same value can be obtained from formula (2) also.

$$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2} \sqrt{\Sigma (Y - \bar{Y})^2}} \dots(2)$$

$$r = \frac{42}{\sqrt{112} \sqrt{38}} = 0.644$$

Thus years of education of farmers and annual yield per acre are positively correlated. The value of *r* is also large. It implies that more the number of years farmers invest in

education, higher will be the yield per acre. It underlines the importance of farmers' education.

To use formula (3)

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}} \dots(3)$$

the value of the following expressions have to be calculated i.e. $\Sigma XY, \Sigma X^2, \Sigma Y^2$.

Now apply formula (3) to get the value of *r*.

Let us know the interpretation of different values of *r*. The correlation coefficient between marks secured in English and Statistics is, say, 0.1. It means that though the marks secured in the two subjects are positively correlated, the strength of the relationship is weak. Students with high marks in English may be getting relatively low marks in statistics. Had the value of *r* been, say, 0.9, students with high marks in English will invariably get high marks in Statistics.

TABLE 7.1
Calculation of *r* between years of schooling of farmers and annual yield

Years of Education (X)	(X - \bar{X})	(X - \bar{X}) ²	Annual yield per acre in '000 Rs (Y)	(Y - \bar{Y})	(Y - \bar{Y}) ²	(X - \bar{X})(Y - \bar{Y})
0	-6	36	4	-3	9	18
2	-4	16	4	-3	9	12
4	-2	4	6	-1	1	2
6	0	0	10	3	9	0
8	2	4	10	3	9	6
10	4	16	8	1	1	4
12	6	36	7	0	0	0
$\Sigma X=42$		$\Sigma (X - \bar{X})^2=112$	$\Sigma Y=49$		$\Sigma (Y - \bar{Y})^2=38$	$\Sigma (X - \bar{X})(Y - \bar{Y})=42$

An example of negative correlation is the relation between arrival of vegetables in the local mandi and price of vegetables. If r is -0.9 , vegetable supply in the local mandi will be accompanied by lower price of vegetables. Had it been -0.1 large vegetable supply will be accompanied by lower price, not as low as the price, when r is -0.9 . The extent of price fall depends on the absolute value of r . Had it been zero there would have been no fall in price, even after large supplies in the market. This is also a possibility if the increase in supply is taken care of by a good transport network transferring it to other markets.

Activity

- Look at the following table. Calculate r between annual growth of national income at current price and the Gross Domestic Saving as percentage of GDP.

Step deviation method to calculate correlation coefficient.

When the values of the variables are large, the burden of calculation can be considerably reduced by using a property of r . It is that r is independent of change in origin and scale. It is also known as step deviation method. It involves the

TABLE 7.2

Year	Annual growth of National Income	Gross Domestic Saving as percentage of GDP
1992-93	14	24
1993-94	17	23
1994-95	18	26
1995-96	17	27
1996-97	16	25
1997-98	12	25
1998-99	16	23
1999-00	11	25
2000-01	8	24
2001-02	10	23

Source: *Economic Survey, (2004-05) Pg. 8,9*

transformation of the variables X and Y as follows:

$$U = \frac{X - A}{h}; V = \frac{Y - B}{k}$$

where A and B are assumed means, h and k are common factors and have same signs.

Then $r_{UV} = r_{XY}$

This can be illustrated with the exercise of analysing the correlation between price index and money supply.

Example 2

Price index (X)	120	150	190	220	230
Money supply in Rs crores (Y)	1800	2000	2500	2700	3000

The simplification, using step deviation method is illustrated below. Let $A = 100$; $h = 10$; $B = 1700$ and $k = 100$

The table of transformed variables is as follows:

Calculation of r between price index and money supply using step deviation method

TABLE 7.3

U	V	U^2	V^2	UV
$\left(\frac{X-100}{10}\right)$	$\left(\frac{Y-1700}{100}\right)$			
2	1	4	1	2
5	3	25	9	15
9	8	81	64	72
12	10	144	100	120
13	13	169	169	169

$$\Sigma U = 41; \Sigma V = 35; \Sigma U^2 = 423;$$

$$\Sigma V^2 = 343; \Sigma UV = 378$$

Substituting these values in formula (3)

$$r = \frac{\Sigma UV - \frac{(\Sigma U)(\Sigma V)}{N}}{\sqrt{\Sigma U^2 - \frac{(\Sigma U)^2}{N}} \sqrt{\Sigma V^2 - \frac{(\Sigma V)^2}{N}}} \quad (3)$$

$$= \frac{378 - \frac{41 \times 35}{5}}{\sqrt{423 - \frac{(41)^2}{5}} \sqrt{343 - \frac{(35)^2}{5}}}$$

$$= 0.98$$

The strong positive correlation between price index and money supply is an important premise of monetary policy. When the money supply grows the price index also rises.

Activity

- Using data related to India's population and national income, calculate the correlation between them using step deviation method.

Spearman's rank correlation

Spearman's rank correlation was developed by the British psychologist C.E. Spearman. It is used when the variables cannot be measured meaningfully as in the case of price, income, weight etc. Ranking may be more meaningful when the measurements of variables are suspect. Consider the situation where we are required to calculate the correlation between height and weight of students in a remote village. Neither measuring rods nor weighing scales are available. The students can be easily ranked in terms of height and weight without using measuring rods and weighing scales.

There are also situations when you are required to quantify qualities such as fairness, honesty etc. Ranking may be a better alternative to quantification of qualities. Moreover, sometimes the correlation coefficient between two variables with extreme values may be quite different from the coefficient without the extreme values. Under these circumstances rank correlation provides a better alternative to simple correlation.

Rank correlation coefficient and simple correlation coefficient have the same interpretation. Its formula has

been derived from simple correlation coefficient where individual values have been replaced by ranks. These ranks are used for the calculation of correlation. This coefficient provides a measure of linear association between ranks assigned to these units, not their values. It is the Product Moment Correlation between the ranks. Its formula is

$$r_s = 1 - \frac{6\sum D^2}{n^3 - n} \quad \dots(4)$$

where n is the number of observations and D the deviation of ranks assigned to a variable from those assigned to the other variable. When the ranks are repeated the formula is

$$r_s = 1 - \frac{6 \left[\sum D^2 + \frac{(m_1^3 - m_1)}{12} + \frac{(m_2^3 - m_2)}{12} + \dots \right]}{n(n^2 - 1)}$$

where m_1, m_2, \dots , are the number of repetitions of ranks and $\frac{m_1^3 - m_1}{12}, \dots$ their corresponding correction factors. This correction is needed for every repeated value of both variables. If three values are repeated, there will be a correction for each value. Every time m_1 indicates the number of times a value is repeated.

All the properties of the simple correlation coefficient are applicable here. Like the Pearsonian Coefficient of correlation it lies between 1 and -1. However, generally it is not as accurate as the ordinary method. This is due the fact that all the information concerning the data is

not utilised. The first differences of the values of items in the series, arranged in order of magnitude, are almost never constant. Usually the data cluster around the central values with smaller differences in the middle of the array. If the first differences were constant, then r and r_k would give identical results. The first difference is the difference of consecutive values. Rank correlation is preferred to Pearsonian coefficient when extreme values are present. In general r_k is less than or equal to r .

The calculation of rank correlation will be illustrated under three situations.

1. The ranks are given.
2. The ranks are not given. They have to be worked out from the data.
3. Ranks are repeated.

Case 1: When the ranks are given

Example 3

Five persons are assessed by three judges in a beauty contest. We have to find out which pair of judges has the nearest approach to common perception of beauty.

Judge	Competitors				
	1	2	3	4	5
A	1	2	3	4	5
B	2	4	1	5	3
C	1	3	5	2	4

There are 3 pairs of judges necessitating calculation of rank correlation thrice. Formula (4) will be used.

$$r_s = 1 - \frac{6\sum D^2}{n^3 - n} \quad \dots(4)$$

The rank correlation between A and B is calculated as follows:

A	B	D	D ²
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
Total			14

Substituting these values in formula (4)

$$r_s = 1 - \frac{6\sum D^2}{n^3 - n} \quad \dots(4)$$

$$= 1 - \frac{6 \times 14}{5^3 - 5} = 1 - \frac{84}{120} = 1 - 0.7 = 0.3$$

The rank correlation between A and C is calculated as follows:

A	C	D	D ²
1	1	0	0
2	3	-1	1
3	5	-2	4
4	2	2	4
5	4	1	1
Total			10

Substituting these values in formula (4) the rank correlation is 0.5. Similarly, the rank correlation between the rankings of judges B and C is 0.9. Thus, the perceptions of judges A and C are the closest. Judges B and C have very different tastes.

Case 2: When the ranks are not given

Example 4

We are given the percentage of marks, secured by 5 students in Economics and Statistics. Then the ranking has to be worked out and the rank correlation is to be calculated.

Student	Marks in Statistics (X)	Marks in Economics (Y)
A	85	60
B	60	48
C	55	49
D	65	50
E	75	55

Student	Ranking in Statistics (R _x)	Ranking in Economics (R _y)
A	1	1
B	4	5
C	5	4
D	3	3
E	2	2

Once the ranking is complete formula (4) is used to calculate rank correlation.

Case 3: When the ranks are repeated

Example 5

The values of X and Y are given as

X	25	45	35	40	15	19	35	42
Y	55	60	30	35	40	42	36	48

In order to work out the rank correlation, the ranks of the values are worked out. Common ranks are given to the repeated items. The

common rank is the mean of the ranks which those items would have assumed if they were slightly different from each other. The next item will be assigned the rank next to the rank already assumed. The formula of Spearman's rank correlation coefficient when the ranks are repeated is as follows

$$r_s = 1 - \frac{6 \left[\Sigma D^2 + \frac{(m^3_1 - m_1)}{12} + \frac{(m^3_2 - m_2)}{12} + \dots \right]}{n(n^2 - 1)}$$

where m_1, m_2, \dots are the number of repetitions of ranks and $\frac{m^3_1 - m_1}{12}, \dots$, their corresponding correction factors.

X has the value 35 both at the 4th and 5th rank. Hence both are given the average rank i.e.,

$$\frac{4 + 5}{2} \text{th} = 4.5^{\text{th}} \text{ rank}$$

X	Y	Rank of XR'	Rank of YR'	Deviation in Ranking D=R'-R''	D ²
25	55	6	2	4	16
45	80	1	1	0	0
35	30	4.5	8	3.5	12.25
40	35	3	7	-4	16
15	40	8	5	3	9
19	42	7	4	3	9
35	36	4.5	6	-1.5	2.25
42	48	2	3	-1	1
Total		$\Sigma D = 65.5$			

The necessary correction thus is

$$\frac{m^3 - m}{12} = \frac{2^3 - 2}{12} = \frac{1}{2}$$

Using this equation

$$r_s = 1 - \frac{6 \left[\Sigma D^2 + \frac{(m^3 - m)}{12} \right]}{n^3 - n} \quad \dots(5)$$

Substituting the values of these expressions

$$r_s = 1 - \frac{6(65.5 + 0.5)}{8^3 - 8} = 1 - \frac{396}{504} = 1 - 0.786 = 0.214$$

Thus there is positive rank correlation between X and Y. Both X and Y move in the same direction. However, the relationship cannot be described as strong.

Activity

- Collect data on marks scored by 10 of your classmates in class IX and X examinations. Calculate the rank correlation coefficient between them. If your data do not have any repetition, repeat the exercise by taking a data set having repeated ranks. What are the circumstances in which rank correlation coefficient is preferred to simple correlation coefficient? If data are precisely measured will you still prefer rank correlation coefficient to simple correlation? When can you be indifferent to the choice? Discuss in class.

4. CONCLUSION

We have discussed some techniques for studying the relationship between

two variables, particularly the linear relationship. The scatter diagram gives a visual presentation of the relationship and is not confined to linear relations. Karl Pearson's coefficient of correlation and Spearman's rank correlation measure linear relationship among variables. When the variables cannot

be measured precisely, rank correlation can be used. These measures however do not imply causation. The knowledge of correlation gives us an idea of the direction and intensity of change in a variable when the correlated variable changes.

Recap

- Correlation analysis studies the relation between two variables.
- Scatter diagrams give a visual presentation of the nature of relationship between two variables.
- Karl Pearson's coefficient of correlation r measures numerically only linear relationship between two variables. r lies between -1 and 1 .
- When the variables cannot be measured precisely Spearman's rank correlation can be used to measure the linear relationship numerically.
- Repeated ranks need correction factors.
- Correlation does not mean causation. It only means covariation.

EXERCISES

1. The unit of correlation coefficient between height in feet and weight in kgs is
 - (i) kg/feet
 - (ii) percentage
 - (iii) non-existent
2. The range of simple correlation coefficient is
 - (i) 0 to infinity
 - (ii) minus one to plus one
 - (iii) minus infinity to infinity
3. If r_{xy} is positive the relation between X and Y is of the type
 - (i) When Y increases X increases
 - (ii) When Y decreases X increases
 - (iii) When Y increases X does not change

4. If $r_{xy} = 0$ the variable X and Y are
 (i) linearly related
 (ii) not linearly related
 (iii) independent
5. Of the following three measures which can measure any type of relationship
 (i) Karl Pearson's coefficient of correlation
 (ii) Spearman's rank correlation
 (iii) Scatter diagram
6. If precisely measured data are available the simple correlation coefficient is
 (i) more accurate than rank correlation coefficient
 (ii) less accurate than rank correlation coefficient
 (iii) as accurate as the rank correlation coefficient
7. Why is r preferred to covariance as a measure of association?
8. Can r lie outside the -1 and 1 range depending on the type of data?
9. Does correlation imply causation?
10. When is rank correlation more precise than simple correlation coefficient?
11. Does zero correlation mean independence?
12. Can simple correlation coefficient measure any type of relationship?
13. Collect the price of five vegetables from your local market every day for a week. Calculate their correlation coefficients. Interpret the result.
14. Measure the height of your classmates. Ask them the height of their benchmate. Calculate the correlation coefficient of these two variables. Interpret the result.
15. List some variables where accurate measurement is difficult.
16. Interpret the values of r as 1 , -1 and 0 .
17. Why does rank correlation coefficient differ from Pearsonian correlation coefficient?
18. Calculate the correlation coefficient between the heights of fathers in inches (X) and their sons (Y)
- | | | | | | | | | |
|---|----|----|----|----|----|----|----|----|
| X | 65 | 66 | 57 | 67 | 68 | 69 | 70 | 72 |
| Y | 67 | 56 | 65 | 68 | 72 | 72 | 69 | 71 |
- (Ans. $r = 0.603$)
19. Calculate the correlation coefficient between X and Y and comment on their relationship:
- | | | | | | | |
|---|----|----|----|---|---|---|
| X | -3 | -2 | -1 | 1 | 2 | 3 |
| Y | 9 | 4 | 1 | 1 | 4 | 9 |
- (Ans. $r = 0$)

20. Calculate the correlation coefficient between X and Y and comment on their relationship

X	1	3	4	5	7	8
Y	2	6	8	10	14	16

(Ans. $r = 1$)

Activity

- Use all the formulae discussed here to calculate r between India's national income and exports taking at least ten observations.

© NCERT
not to be republished



Regression Analysis

**By: Dr.
Yamini**

Regression - Introduction

The statistical technique of estimating the unknown value of one variable (called **dependent variable**) from the known value of other variable (called **independent variable**) is called **Regression Analysis**.

Utility of Regression Analysis

- It provides a functional form of relationship between two related variables with the help of which the **unknown value of one variable can be estimated from the other known variable.**
- It provides a measure of **coefficient of correlation (denoted by 'r')** between the two variables, which is the square root of the product of **two regression coefficients,**

i.e.,

$$R = \sqrt{b_{xy} \cdot b_{yx}}$$

Utility of Regression Analysis

- It establishes the rate of change in one variable in terms of the change in another variable.
- In regression analysis, one variable is assumed to be **cause** and the other is the **effect**.
- The coefficient of determination "**r²**" can be obtained by multiplying the two regression coefficients. i.e

$$r^2 = b_{xy} \cdot b_{yx}$$

Correlation &

CORRELATION	REGRESSION
Correlation explains the degree & direction of relationship between two variables.	Regression explains the cause and effect relationship between two variables.
In correlation, there is no dependent and independent variable.	In regression, one variable is dependent and other variable is independent.
Correlation coefficient value lies between ± 1 .	It shows if there is change of 1 unit in the value of one variable , how much average change will be there in another variable.
Its use is limited upto linear relationship of two variables.	It studies the linear or non-linear relationship between two or more variables.

Types of Regression

Simple & Multiple Regression.

- ❖ In simple regression, **two variables** are involved (one dependent and other independent).
- ❖ In multiple regression, more than two variables are involved (one dependent and more than two independent).

Types of Regression

- Linear and Non-linear Regression:

- ❖ After plotting the data on graph paper, scatter diagram is obtained. These dots follow some path, which may be described by a line or a curve.
- ❖ If dots gives no message, then variables are not related.

Regression lines

- In bivariate data, there are always two lines of regression: Regression line of Y on X and Regression line of X on Y .
- In regression line of X on Y , Y is independent variable and X is dependent variable.
- In regression line of Y on X , X is independent variable and Y is dependent variable.

Properties of Regression Lines

- Both the regression lines intersect each other at mean point (mean of X and Y).
- If both the regression lines are perpendicular to each other, then there is no correlation between X and Y . ($r = 0$)
- If both the regression lines coincide each other, then there is a perfect correlation between the variables X and Y . ($r=1, -1$)
- There are always two regression lines.

Methods of studying Regression

- Graphic Method:

- Under this method, the pair of points are plotted on a graph paper and a scatter diagram is obtained.

- Thereafter, a regression line is drawn in between these points.

Methods of studying Regression

- Mathematical Method:

There are two methods used for obtaining regression lines:

- (a) Regression equations through **Normal Equations**.
- (b) Regression equations through **Regression Coefficients**.

Regression

Equations through

- Regression Line of Y on X:

Normal Equations

In this, X is the independent variable and Y is the dependent variable. The form of equation is given by:

$$Y \text{ (dependent)} = a + bX \text{ (independent)}$$

Above equation shows that Y is a linear function of X.

'a' is called the intercept

Regression

Equations through Normal Equations

The values of 'a' and 'b' are determined by the method of least squares.
The following are the two normal equations:

$$\Sigma Y = Na + b\Sigma X \dots\dots\dots(1)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \dots\dots\dots(2)$$

Solving these equations, the value of 'a' and 'b' are obtained and regression line is obtained.

Regression Equations through Normal Equations

- Regression Line of X on Y:

In this, X is the dependent variable and Y is the independent variable. The form of equation is given by:

$$X \text{ (dependent)} = a + bY \text{ (independent)}$$

Above equation shows that X is a linear function of Y.

'a' is called the intercept

Regression

Equations through Normal Equations

The values of 'a' and 'b' are determined by the method of least squares.
The following are the two normal equations:

$$\sum X = Na + b\sum Y \dots\dots\dots(1)$$

$$\sum XY = a\sum Y + b\sum Y^2 \dots\dots\dots(2)$$

Solving these equations, the value of 'a' and 'b' are obtained and regression line is obtained.

Regression Equations through Regression Coefficients


(a) Regression Equation of Y on X:

$$Y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (X - \bar{X})$$

The above equation can also be rewritten as

$$(Y - \bar{Y}) = b_{yx} (X - \bar{X})$$

(*) Here, σ_x = Standard deviation of X , σ_y = Standard deviation of Y



Calculation of Regression Coefficients byx

Regression Equations through Regression Coefficients


(a) Regression Equation of X on Y:

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

The above equation can also be rewritten as

$$(X - \bar{X}) = b_{xy} (Y - \bar{Y})$$

(*) Here, σ_x = Standard deviation of X , σ_y = Standard deviation of Y



Calculation of Regression Coefficients b_{xy}

Kurtosis

Dr. Yamini Negi

Kurtosis - Meaning

- Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution.
- Kurtosis is a measure of the shape of a distribution.
- It is a measure of the relative peakedness of its frequency curve in relation to the normal curve.

Kurtosis – Meaning (Contd.)

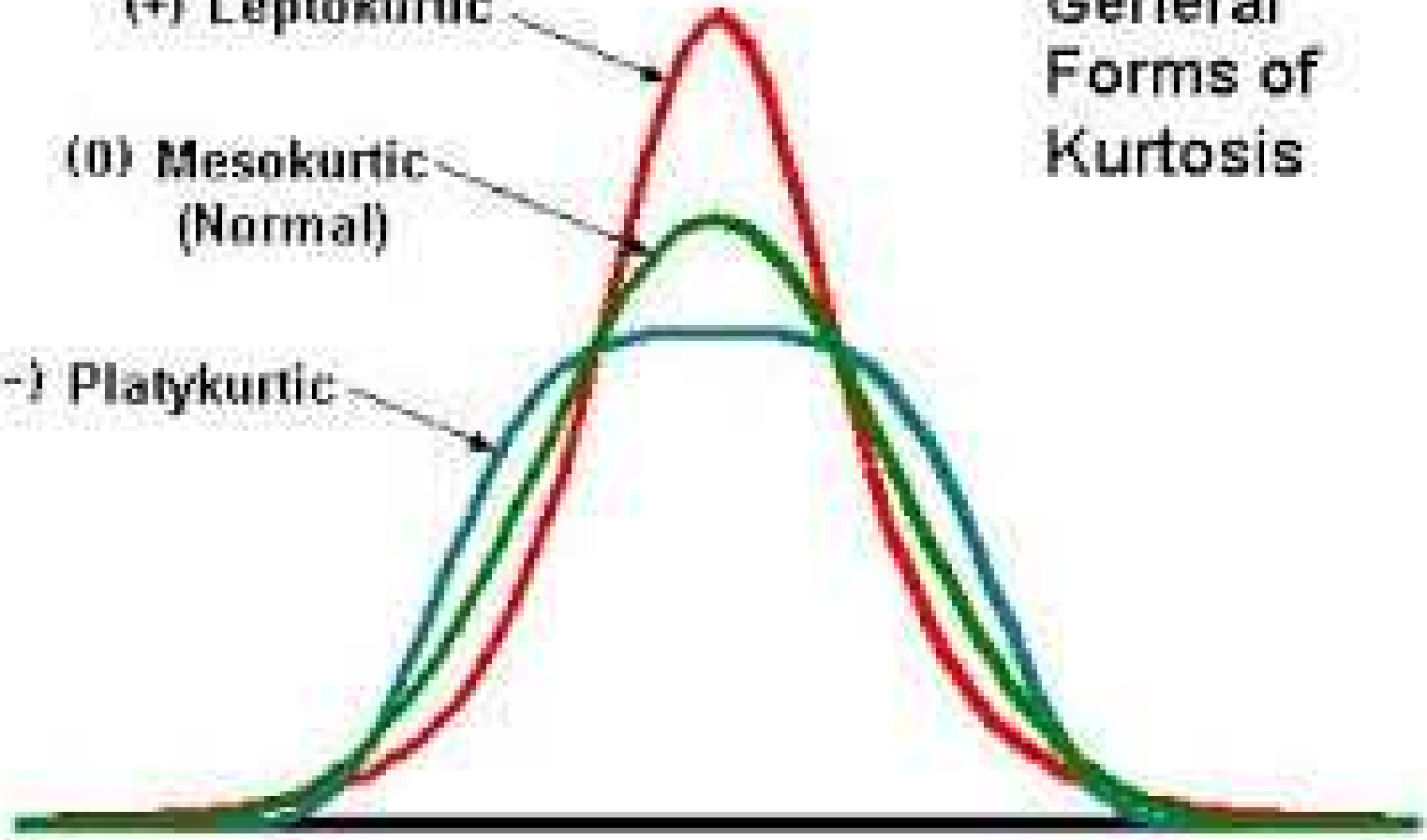
- It indicates the extent to which a distribution is **more peaked or flat topped** than the normal curve.
- Frequency curves can be divided into three categories depending upon **the shape of their peak**.
- The three shapes are termed as **Leptokurtic, Mesokurtic and Platykurtic.** (Theory*)

General Forms of Kurtosis

(+) Leptokurtic

(0) Mesokurtic
(Normal)

(-) Platykurtic



- Skewness measures the **symmetry** of the distribution, while kurtosis determines the **heaviness** of the **distribution tails**. **(difference *)**
- In finance research, kurtosis is used as a measure of **financial risk**.
- High value of kurtosis signifies **high level of risk** of an investment. (i.e. **high probability of extremely large or extremely small returns**).
- Small kurtosis value shows moderate level of risk as the probability of extreme returns are relatively low.

Excess Kurtosis

- The kurtosis of a normal distribution equals to 3. Thus, the excess kurtosis is found using the formula below:

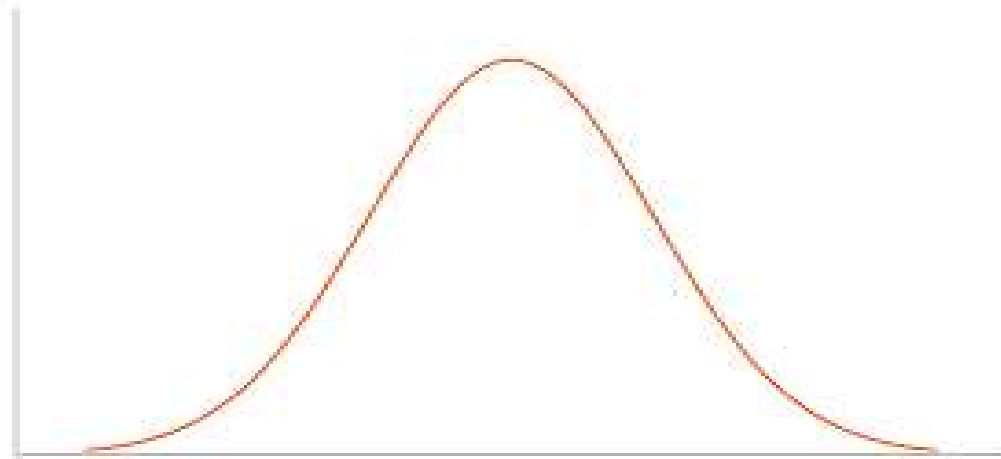
$$\text{Excess Kurtosis} = \text{Kurtosis value} - 3$$

Types of Kurtosis

- The types of kurtosis are determined by the **excess kurtosis** of a particular distribution.
- The excess kurtosis can take **positive (4-3)** or **negative (2-3)** values as well as values close to **zero (3-3)**

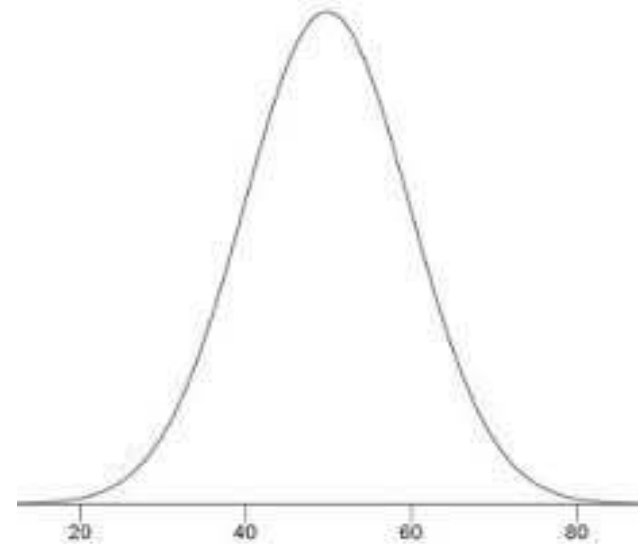
Mesokurtic

- Data that follows a mesokurtic distribution shows an excess kurtosis or **zero** or close to zero.
- It means that if the data follows a normal distribution, it follows a mesokurtic distribution.



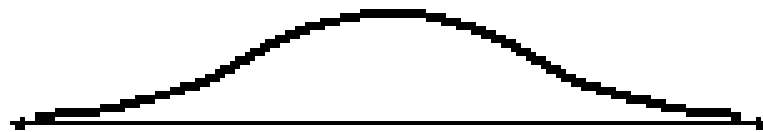
Leptokurtic

- Leptokurtic indicates a **positive excess kurtosis**. Leptokurtic distribution shows heavy tails on either side, indicating the large outliers.
- In finance, a leptokurtic distribution shows that investment returns may be prone to extreme values
(investments are considered to be risky).



Platykurtic

- A platykurtic distribution shows a **negative excess kurtosis**.
- In finance, platykurtic distribution shows that there is a small probability that the investment would experience extreme returns.



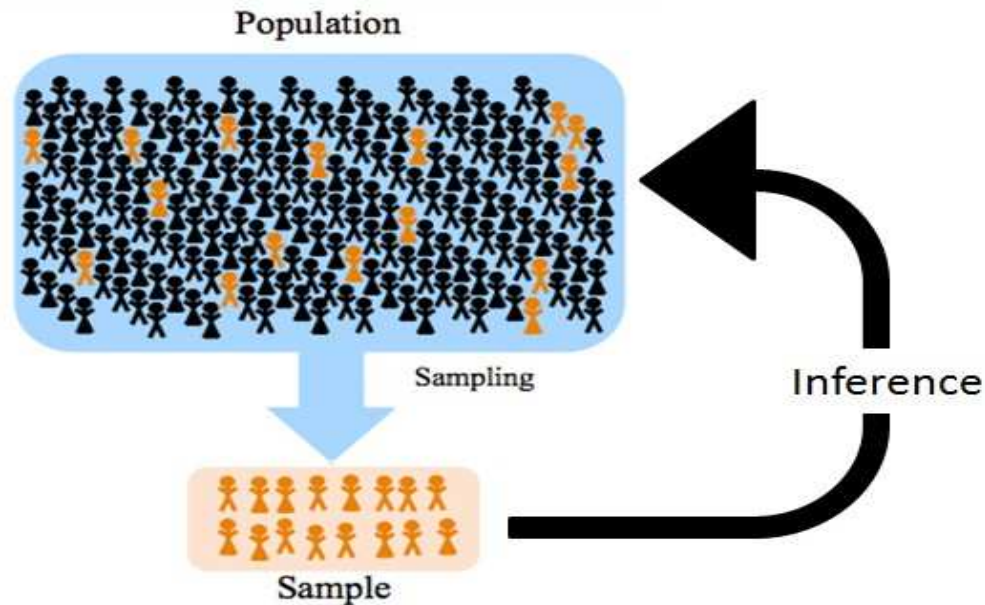
Platykurtic
distribution

- Mesokurtic = Normal Curve = 3, excess kurtosis = 0
- Leptokurtic = More peaked than Normal curve (positive excess kurtosis)
- Platykurtic = Less peaked than Normal curve (negative excess kurtosis)

SAMPLING: TYPES & ERRORS

Dr. Yamini Negi

- A population can be defined as including all people or items with the characteristic one wishes to understand.
- As there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample of that population.



Approaches of Investigation:

- **Census Investigation:**
 - ❖ When information on each and every unit of the population is collected, it is called census investigation.
 - ❖ The results in this approach are expected to be accurate, representative and reliable.
 - ❖ **Example:** Knowing the average life time of a batch of 5000 bulbs produced by a certain machine on a given day.

Approaches of Investigation (contd):

- **Sample Investigation:**
 - ❖ When each and every unit of the population is not investigated but only a part of the population called sample is investigated, it is called sample investigation.
 - ❖ **Example:** Choosing 100 bulbs and determining their average life and inferences can be applied to all the bulbs.

Assumptions underlying in sample investigation:

- **Homogeneity Amidst Complexity:** If two units were alike in any respect, the sampling would have been impossible.
- **Possibility of Representative Selection:** Each unit of the population should get equal chance of being selected.
- **Absolute Accuracy not essential:** 95% accuracy is sufficient for large scale studies.
- **Independency:** All items in a sample should be independent of each other.

Sample & Census Investigation: Comparison

- **Cheaper**
- **Time saving**
- **Errors**
- **Preference**
- **Obtaining data**

Steps in Statistical Investigation

- **Objective of the survey**
- **Population to be sampled**
- **Data to be collected**
- **Methods of measurement**
- **Selection of the sample**
- **The Pretest**
- **Summary & Analysis of the data**
- **Information gained for future surveys**

Terminology:

- **Sampling Fraction:** The ratio of sample size 'n' to the population size 'N'.
- **Sampling Unit:** An identifiable part of the population on which some information has to be collected.
- **Sampling list or Source list:** Ordered representation of the sampling units of a population from which the sample is selected.
- **Population Parameter:** Mean of population
- **Estimator:** Mean of sample.

Methods of Sampling

- **Random Sampling or Probability Sampling**
 - ✓ Simple Random Sampling
 - ✓ Stratified Random Sampling
 - ✓ Systematic Sampling
 - ✓ Multi-stage sampling
 - ✓ Cluster Sampling
- **Non-Random Sampling or non-Probability sampling**
 - ✓ Judgement Sampling
 - ✓ Quota Sampling
 - ✓ Convenience Sampling

• Simple Random Sampling:

- ❖ Each unit of the universe has an equal chance of its selection.
- ❖ No item is given preference.
- ❖ With replacement sampling/without replacement sampling.
- ❖ **Methods:** By lottery, by random numbers.

• Stratified Random Sampling:

- ❖ Population under consideration is divided into number of sub-populations called strata.
- ❖ Samples are selected independently from each stratum.
- ❖ **Types:** Equal allocation, Proportional allocation.

• Systematic Sampling:

- ❖ involves selecting only first unit at random, the rest being selected according to some predetermined pattern.
- ❖ This sampling is also called “**Quasi-random Sampling**”.
- ❖ Here, $k = N/n$

where k = sampling interval,

N = Population

n = Sample

• Cluster Sampling:

- ❖ It involves grouping the population units and selected some of these groups by random sampling technique.
- ❖ In cluster sampling, within cluster variation is more and between cluster variation is less.

• Multi-stage sampling:

- ❖ This sampling is used to reduce the cost of survey.
- ❖ It involves selection of sample at various levels: from large-sized sample to selection of smaller number of units.

• Non-Random Sampling:

- ❖ Sample is selected with a definite purpose and choice of the sampling units depends entirely on the personal judgement of the investigator.
- ❖ The sample size selected may not be the true representativeness of the population.

- **Judgement Sampling:**

- ❖ This sampling is also called **purposive sampling**.
- ❖ In this, the sample is selected by the **personal judgement** of the investigator.
- ❖ sample selected by this method is **not a random one**.

- **Convenience Sampling:**

- ❖ The sample is taken from a group of people who are easy to contact or to reach. **For example**, standing at a mall or a grocery store and asking people to answer questions would be an example of a convenience sample.
- ❖ This sampling is also called **accidental sampling or grab sampling**.

- **Snowball Sampling:**

- ❖ The research starts with a key person and introduce the next one to become a chain.
- ❖ Make contact with one or two cases in the population.
- ❖ Ask these cases to identify further cases.

- **Self Selection Sampling:**

- ❖ It occurs when each case is allowed to **identify their desire to take part in the research.**
- ❖ Need for the cases is publicized, either by advertising or by asking them to take part.
- ❖ Collect data from those who respond.

• Sampling Errors:

- ❖ The difference between a **sample estimate (mean of the sample)** and the **population parameter (mean of the population)** obtained is called the sampling error.
- ❖ Sampling error = **Sample Estimate – Population parameter**
- ❖ These errors can be reduced **by increasing sample size.**
- ❖ A sample with minimum fluctuations is taken as the best representative of the population.

• Non - Sampling Errors:

- ❖ The errors mainly arising at the stages of **ascertainment and processing of data is termed as non-sampling error.**
- ❖ This error is common both to **complete enumeration** (census investigation and sample surveys).
- ❖ Non sampling errors can be of '**Non Response Error**' and '**Response Error**'.

- **Non - Response Errors:**

- ❖ A non-response error occurs when units selected as part of the sampling procedure do not respond in whole or in part.

• Response Errors:

- ❖ A response or data error is any systematic bias that occurs during data collection, analysis or interpretation.
- ❖ It can be any types like Respondent error, Interviewer error, Recording errors, Poorly designed questionnaires, Measurement error.

- **Respondent Error:**

- ❖ A respondent **gives an incorrect answer**, e.g. due to prestige, or due to social undesirability of question.
- ❖ Respondent misunderstands the requirements.
- ❖ Question requires memory/recall.
- ❖ Proxy respondents are used, i.e. taking answers from someone other than the respondent.

- **Interviewer Bias:**

- ❖ Different interviewers administer a survey in different ways.
- ❖ Differences occur in reactions of respondents to different interviewers.
- ❖ Inadequate attention to the selection of interviewers.
- ❖ Too much workload on the interviewer.

- **Measurement Error:**

- ❖ The question is unclear, ambiguous or difficult to answer.
- ❖ Requested information assumes a framework unfamiliar to the respondent.



Test of Significance:

Hypothesis Testing

By:

Hypothesis

- A hypothesis is a tentative generalisation, the validity of which remains to be tested.
- In its most elementary stage, the hypothesis may be a any guess, imaginative idea, which become the basis

Basic Concepts:

Null Hypothesis:

- A null hypothesis is a statement about population parameter (such as μ), and the test is used to decide whether or not accept the hypothesis.

- The null hypothesis is always expressed in the form of mathematical statement which includes:

$$H_0 : \mu(\leq, \square, \geq) \mu_0$$

- **Alternate Hypothesis:**

The alternate hypothesis states that specific population parameter value is not equal to the value stated in the null hypothesis and can be written as:

$$H_1 : \mu \neq \mu_0$$

Procedure in Hypothesis Testing:

- (1) Formulate a Hypothesis
- (2) Set up a suitable **significance level**.
- (3) Select Test Criterion (z test, T test, F test, Chi-square test)
- (4) Compute the value.

Types of Errors in Hypothesis Testing:

When a hypothesis is tested, there are four possibilities:

- (1) Hypothesis is true but our test leads to its rejection.**



Type I Error:

Hypothesis is true but our test leads to its rejection.

Type II Error: