Introduction to Statistics

1

INTRODUCTION

An educated citizen needs an understanding of basic statistical tools to function in a world that is becoming increasingly dependent on quantitative information. Statistics means numerical description to most people. In fact, the term statistics is generally used to mean numerical facts and figures such as per capita income, agricultural production during a year, rate of inflation and so on. However, as a subject of study, statistics refers to the body of principles and procedures developed for the collection, classification, summarization and interpretation of numerical data and for the use of such data. Without the assistance of statistical methods an organisation would find it impossible to make sense of the huge data. The purpose of statistics is to manipulate, summarize, and investigate data so that useful decision-making information results. In fact, every business manager needs a sound background of statistics. Statistics is a set of decision-making techniques which aids businessmen in drawing inferences from the available data.

ORIGIN OF STATISTICS

The term statistics has its origin in Latin word *Status*, Italian word *Statista* or German term *statistik*. All the three terms mean "Political State". In fact, the beginning of statistics was made to meet the administrative needs of the state. In ancient periods, the states were required to collect statistical data mainly for two purposes. One, concerning population so that state may come to know of the number of youngmen in the country that can be recruited in the army. Two, concerning land holdings so that state may calculate the total amount of land revenue that can be collected. It is because of this reason that statistics is called **"Science of Statecraft"** or **"Political Arithmetic"**. In modern times, statistics is not related to the administration of the state alone, but it has close relation with almost all those activities of our lives which can be expressed in **quantitative terms**.

MEANING OF STATISTICS

Broadly speaking, the term statistics has been generally used in two senses—**(1) Plural sense**, and **(2) Singular sense**. In the plural sense, the term statistics refers to numerical statements of facts relating to any field of enquiry such as data relating to production, income, expenditure, population, prices, etc. In other words, the term statistics in its plural sense refers to **numerical data or statistical data**. In its singular sense, the term statistics refers to a science in which we deal with the techniques or methods for collecting, classifying, presenting, analysing and interpreting the data. In other words, the concept in its singular sense, refers to **statistical methods**. Thus, the word 'statistics' refers either to the data themselves or to the methods dealing with numerical data.

◘ DEFINITIONS OF STATISTICS

Different writers have defined statistics differently. These are broadly divided into t categories:

(I) Definitions in the Plural Sense.

(II) Definitions in the Singular Sense.

○ (I) Definitions of Statistics in the Plural Sense or as Numerical Data

In the plural sense, the term Statistics is used for numerical data. Some of its importa definitions are:

"Statistics are numerical statements of facts in any department of enquiry placed in relation each other."
—Bowl

"By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes."
—**Yule and Kendal**

"Statistics are numerical descriptions of quantitative aspects of things and they take the form counts or measurements."
—**Wallis and Robert**

The most popular definition of statistics in terms of numerical data has been given by **Horac Secrist** which is given below:

"By statistics we mean aggregate of facts affected to a marked extent by multiplicity of cause numerically expressed, enumerated or estimated according to reasonable standards of accuracy collected in a systematic manner for a pre-determined purpose and placed in relation to eac other."

It is a comprehensive definition that covers all aspects of any meaningful quantitative information.

▶ Features or Characteristics of Statistics in terms of Numerical Data

Some of the important characteristics of statistics in terms of numerical data are as follows:

(1) Aggregate of Facts: A single number does not constitute statistics. No conclusion can be drawn from it. It is only the aggregate of facts capable of offering some meaningful conclusion that constitute statistics. Likewise, the ratio of radius and circumference of a circle cannot be called statistics. For example, if it is stated that there are 1000 students in our college, then it has no statistical significance. But if it is stated that there are 300 students in arts faculty, 400 in commerce faculty and 300 in science faculty in our college, it makes statistical sense as this data conveys statistical information. Similarly, if it is stated that population of India is 91.5 crore or that the value of total exports from India is Rs. 1,06,353 crore, then these aggregate of facts will be termed as statistics. It can, therefore, be said: *'All statistics are expressed in numbers but all numbers are not 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, the fact Kapil Dev is tall and Gavaskar is short, has no statistical sense. However, if the height of Kapil Dev is 6′ 2" and that of Gavaskar is 5′4", it would be taken as statistical information.

(3) Affected by Multiplicity of Causes: Statistics are not affected by any single factor. These are influenced by many factors simultaneously. For instance, 30 per cent 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. For example, difference of one kg. of weight in five kg. of sweetmeat is a height of inaccuracy but if against the weight of one quintal of wheat there is difference of one kg. of wheat, the inaccuracy will be treated as negligence and insignificant.

(5) Placed in Relation to each other: Such numericals alone will be called statistics as are mutually related and comparable. Unless they have the quality of comparison, they cannot be called statistics. For example, if it is stated "Ram is 40 years old, Mohan is 5 ft. tall, Sohan has 60 kg. of weight", then these numbers will not be called statistics, as they are neither mutually related nor comparable. However, if the age, height and weight of all the three are inter-related, then the same will be considered as statistics.

(6) Pre-determined Purpose: Statistics are collected with some pre-determined objective. Any information gathered without any definite purpose will only be a numerical value and not statistics. If data pertaining to the farmers of a village are being collected, there must be some pre-determined objective. It may be to know the economic status of the farmers or distribution of land among them or to know their population or for any other purpose.

(7) Enumerated or Estimated: Statistics may be collected by enumeration or these may be estimated. If the field of investigation is vast, the procedure of estimation may be helpful. For example, one lakh people attended the rally addressed by the Prime Minister in Delhi and two lakh in Bombay. These statistics are based on estimation. As against it, if the field of enquiry is limited, the enumeration method is appropriate. For example, it can be verified by enumeration whether 20 students are present in the class or 10 workers are working in the factory.

(8) Collected in Systematic Manner: Statistics should have been collected in a systematic manner. Before collecting them, a plan must be prepared. No conclusion can be drawn from statistics 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 only lead to confusion, not any meaningful conclusion.

In short, numerical data, in the absence of the above characteristics, would not be called statistics.

(II) Definitions of Statistics in Singular Sense or Statistics as a Subject

In the singular sense, Statistics means science of statistics or statistical methods. It refers to the study of the techniques relating to the collection, classification, presentation, analysis and interpretation of quantitative data. Some of the important definitions of the science of statistics are given below:

Dr. Bowley has given the following three definitions of statistics:

(1) Statistics is the Science of Counting: According to this definition, statistics is only a science of counting. But this is an incomplete definition of statistics mainly due to two reasons. (*i*) Statistics is not related only with the collection of data. It is also concerned with the presentation,

tabulation, analysis and interpretation of data. This definition is incomplete since it emphasizes only one aspect of statistical methods, *i.e.*, counting. It ignores other methods like analysis and interpretation which are equally important. (*ii*) Counting is used in the collection of data, where field of enquiry is limited. But where the field is big or data are large, counting becomes impossible. Under such conditions estimates are made. For example, the production of wheat in India cannot be counted, it can only be estimated. **Bowley** himself observes, *"Great numbers are not counted, they are estimated"*.

(2) Statistics may rightly be called the Science of Averages: This is also an incomplete definition of statistics. This definition throws light only on an average which is one of the methods like dispersion, skewness, correlation, index numbers, etc. Hence, it limits the scope of science of statistics.

(3) Statistics is the science of measurement of social organism regarded as a whole in all its manifestations: This is another incomplete definition of statistics. Two reasons lead to the above conclusion (a) This definition limits the application of statistical methods to man's social activities or to social sciences only. (b) This definition mentions only one statistical method, *i.e.*, measurement and does not tell about other statistical methods. Hence, it cannot be regarded as a proper definition of statistics.

"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**

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

▶ Features of Statistics as Science or Statistical Methods

According to all the above definitions of statistics as a science, it is clear that there are five stages of statistics which are given below:

(1) Collection of Data: Collection of relevant data concerning a problem is the first step in statistical method. Depending upon the problem under study, it is decided as to how, when and where and what kind of data are to be collected.

(2) Organisation of Data: The second step in statistical methods is to organize the collected data. With a view to rendering the collected data more comparable and simple, it is classified on the basis of time, place and quality, etc.

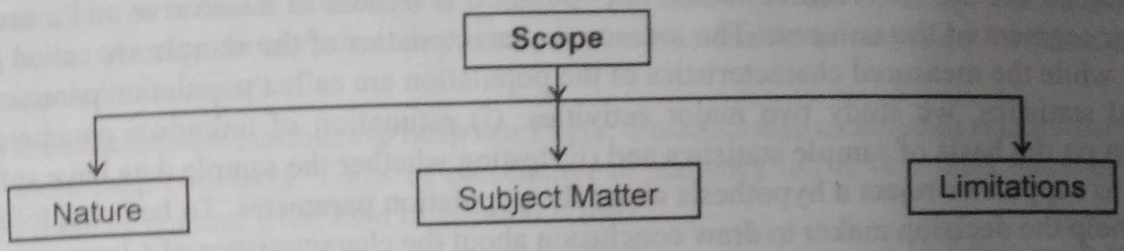
(3) Presentation of Data: To make the data intelligible, brief and attractive, it is presented in the form of tables, diagrams and graphs.

(4) Analysis of Data: The fourth step is the analysis of data. To draw conclusions, it is necessary to analyse the data. There are different methods of analysing the data, *e.g.*, measures of central tendency, measures of variation, correlation, etc.

(5) Interpretation of Data: Under this method, conclusions are drawn after analysing the data. Two or more kinds of data are compared and conclusions drawn. Even a layman may understand them. The conclusions are expressed in simple and easy language. On the basis of such conclusions future estimates are made.

SCOPE OF STATISTICS

The scope of statistics may be classified in the following parts:



(I) Nature of Statistics

The study of nature of statistics is to find out whether it is a Science or Art. As a science, statistics studies numerical data in a systematic manner and as an art, it makes use of the data to solve the problems of real life. Some scholars call it a study of Statistical Methods in preference to Statistics science, because its methods are used in all sciences.

Tippet says, *"Statistics is both a science and an art."* It is a science as its methods are basically systematic and have general applications. It is an art as its successful application depends to a considerable degree on the skill and special experience of a statistician.

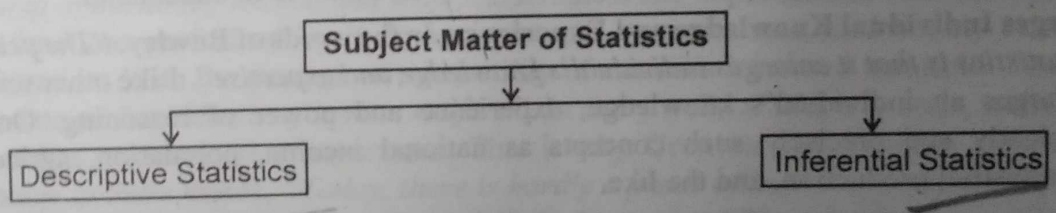
(II) Subject Matter of Statistics

In order to facilitate its study, subject matter of statistics is divided into two parts namely:

(1) Descriptive Statistics

(2) Inferential Statistics

A brief description of above branches of statistics is made as under:



(1) Descriptive Statistics: As the name suggests, the descriptive statistics merely describe the data and consists of the methods and techniques used in the collection, organisation, presentation and analysis of data in order to describe the various features and characteristics of such data. These methods can either be graphical or computational. These data can be presented in the form of chart or table in order to show trends, proportions, maximum and minimum values, etc. In addition to the organisation of data, descriptive statistics is concerned with the analysis of data so that the data can be easily understood. Measures of central tendency, dispersion, skewness and Kurtosis summarises and describes the univariate data and correlation and regression help in the establishing of the relationship in bivariate data. In descriptive statistics, nothing can be inferred from the data nor can decision be made or conclusion drawn.

(2) Inferential Statistics: It deals with methods which help in estimating the characteristics of a population or making decisions concerning a population on the basis of the sample results. Sample and population are the two relative terms. A population is treated as a universe and a sample as a fraction or segment of the universe. The measured characteristics of the sample are called sample statistics, while the measured characteristics of the population are called population parameters. In inferential statistics, we study two major activities: (i) estimation of unknown parameter of a population on the basis of sample statistics and (ii) testing whether the sample data have sufficient evidence to support or reject a hypothesis about the population parameter. To be brief, inferential statistics help the decision maker to draw conclusion about the characteristics of a large population on the basis of sample results.

◘ FUNCTIONS OF STATISTICS

Main functions of statistics are as follows:

(1) Expression of Facts in Numbers: One of the principal function of statistics is to express facts relating to different phenomena in numbers. Statement is vested with certainty when facts are expressed in numbers. For example, the statement that per capita income of India is increasing is not so precise. But if this statement is expressed in numbers as: India's per capita income which was Rs. 245 in 1950-51 rose to Rs. 9000 in 1995-96, then it becomes easy to understand and interpret with certainty.

(2) Simple Presentation: Another function of statistics is to present complex data in a simple form, so that it becomes easy to comprehend. Statistics renders complex data very simple by expressing it in terms of aggregate, average, percentage, graphs and diagrams. For example, data concerning changes in prices of main commodities between 1951 and 1995 may be so voluminous and cumbersome that it would be difficult to understand them or to draw any conclusion about them. But when presented in the form of index numbers these become simple to understand.

(3) Enlarges Individual Knowledge and Experience: In the words of **Bowley**, *"The principal function of statistics is that it enlarges individual's knowledge and expertise"*. Like other sciences, statistics enlarges an individual's knowledge, experience and power of reasoning. One can understand clearly and precisely such concepts as national income, population, agricultural production, industrial production, and the like.

(4) It Compares Facts: Another function of statistics is to compare the data relating to facts. Data have no meaning unless these are compared and inter-related. For example, if it is stated that the per capita consumption of sugar in India is 8 kg per annum, then some people may conclude that it is very low rate of consumption while the others may conclude that it is very high. But when it is compared with the consumption of sugar in other countries, say America where it is 38 kgs and Russia where it is 46 kgs, then one can draw a more meaningful conclusion.

(5) It Facilitates Policy Formulation: To facilitate formulation of policy is another function of statistics. Precise nature of each problem can be ascertained from the analysis and interpretation of data. As a result of it, some policy may be formulated. For example, it was with the help of data that **Engel** formulated a law concerning family budget. Monetary and Fiscal policies of country are formulated on the basis of relevant data.

(6) It Helps Other Sciences in Testing their Laws: Statistics also helps in testing the laws of other sciences. Many laws of economics, namely, law of demand, law of supply, Keynes' theory of employment have been verified with the help of statistics. On the contrary, classical theory of employment, quantity theory of money, etc. were not amenable to statistical verification and so were subjected to criticism.

(7) It Establishes Relationship between Facts: Statistics also establishes relationship between two or more than two facts. Tools of correlation of statistics tell if two facts have any relation between them or not and what kind of relation it has?

(8) It Helps in Forecasting: Statistics helps in forecasting changes in future with regard to a problem. For example, forecast can be made with regard to changes, in future, in food production, supply of power, growth of population, etc. as a result of five year plans in India, with the help of statistics. Extrapolation technique of statistics helps in making forecast on the basis of present facts.

(9) It Enables Realisation of Magnitude: Statistics enables realisation of magnitude of a problem. For example, from the statement that India's population has been increasing rapidly, one cannot fully realize the gravity of the situation. But, if it is stated numerically that population of India increases at the rate of 1.30 crore annually which is equal to the entire population of Australia, the magnitude of the population problem can be realized properly and easily.

(10) Presentation of Data in Condensed Form: Primary data are very much complex and haphazard. Such complex data make it impossible to draw any conclusion. Thus, it becomes imperative to present them in a concise form so that conclusions could be drawn. Statistics present complex data into a condensed form.

◘ USES AND IMPORTANCE OF STATISTICS

In the words of **H.G. Wells** *"Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write"*. Usefulness and importance of statistics can be measured by the functions performed by it. In ancient times, statistics was used mainly as an aid to run administration. With the passage of time, utility of statistics increased manifold. Today, statistics has become an *"arithmetic of human welfare"*. In every walk of life, economic, social or political, statistics has assumed great importance. That is why **Croxton and Cowden** have expressed its importance in these words, *"Today, there is hardly a phase of human activity which does not find statistical devices at least occasionally useful"*. In the words of **Tippet**, *"Statistics affects everybody and touches life at many points"*. Usefulness of statistics becomes evident from the following:

(1) Importance for Administrators or Importance for Administration: Statistics has been in use in running the administration since ancient times. In modern times, usefulness of statistics has increased all the more due to increased welfare and other activities of the state. Every administrator has to depend on statistics for the sake of efficient administration. Statistics are the eyes of government administration. (*i*) Finance Minister makes use of data relating to revenue and expenditure while preparing budget. (*ii*) Finance Minister takes decision regarding imposition of new taxes or enhancement or curtailment of public expenditure on the basis of the data relating to production, import, export, national income, etc. (*iii*) Each government formulates its policies concerning family planning, nationalisation, establishment of new industries, twenty-point programmes, etc. on the basis of the data. (*iv*) Government makes assessment and measures efficiency of its different departments like health, education, industry, agriculture, etc. on the basis

of numerical data. (*v*) Statistics prove helpful in taking decisions with regard to the defence of the country, internal law and order and crime situation, police and armed forces, etc. (*vi*) Reports of different commissions and committees appointed by the government are substantiated by the statistical data. (*vii*) All the policies and programmes chalked out by the governments of under-developed countries with the objective of economic development, economic planning, removal of income of inequalities, generation of employment opportunities and price stability are invariably based on statistical data.

(2) Importance for Businessman or Industrialist or Agriculturist or in the Field of Business, Industry or Agriculture: Knowledge of statistics is of great importance to every businessman, industrialist and agriculturist. In the words of **A.L. Boddington**, *"The successful businessman is one whose estimate most closely approaches accuracy"*. A successful businessman or agriculturist estimates demand for and supply of the commodity on the basis of relevant data. While making his estimates regarding demand and supply, he has to take into consideration data relating to seasonal changes, trade cycles, tastes of the consumers, customs, etc. It is necessary for the businessmen to know the nature and place of demand of the goods that they are dealing in and what is the future policy of the government and the possibility of changes in the price. All this can be known with the help of statistics alone. With a view to increasing his sales, a businessman conducts market survey and makes plans for advertisement on the basis of statistics. Thus, states **Ya-Lun-Chou**, *"It is not an exaggeration to say that today nearly every decision in business is made with the aid of statistical data and statistical method"*. Statistics is of great significance to every industrialist and farmer. They have to take decision about the volume of production by anticipating demand on the basis of past data and present trend. Besides, policies regarding purchase of raw materials, sale of finished products, publicity, transport, labour, mobilization of financial resources, price determination, etc. are checked out on the basis of appropriate data. All important facts discussed above are taken into account at the time of setting up of new industries.

(3) Importance in Economics: Statistics is the basis of Economics. In the words of **Ya-Lun-Chou** *"Economists depend upon statistics to measure economic aggregates such as gross national output, consumption, saving, investment, expenditure and changes in the value of money. They also use statistical methods to verify economic theory and to test hypothesis"*.

The importance of statistics in Economics is clear from the following:

(*i*) **Consumption:** How a consumer can get maximum satisfaction is determined on the basis of data pertaining to income and expenditure. Law of demand depends on data concerning price and quantity.

(*ii*) **Production:** A producer aims at earning maximum profit on the basis of data relating to cost and revenue. Efficiency of land, labour, capital and enterprise is studied with the help of statistical data. Government of a country estimates national income on the basis of data relating to production. Likewise, relative contribution of each sector to national income is also studied on the basis of statistical data.

(*iii*) **Exchange:** Price of a commodity is determined on the basis of data relating to its buyers, sellers, their demand and supply, cost and profit, etc.

(*iv*) **Distribution:** In the determination of factor-income, *viz*. rent, wage, interest and profit, statistical data play a significant role.

Study of modern economic problems like revenue, economic planning, national income, employment, inflation, etc. is very much dependent on statistical data. Indeed, economics as a science could develop so much only because of the increasing use of data.

Highlighting the significance of statistics, **Marshall** had to concede, *"Statistics are the straw out of which I, like every other economist, have to make bricks"*. It is because of the blend of Statistics that Economics is generating its new off- shoots such as Econometrics. In the words of **Tippet**, *"A day might come when the departments of economics in the universities will go out of the control of economic theoreticians and come under the control of statistical workshops, in the same manner as the department of physics and chemistry have come under the control of experimental laboratories"*. Statistics prove very useful in the formulation of economic policies. Statistical data is required to formulate and evaluate economic policies like monetary policy, fiscal policy, price control policy, export-import policy, etc. Analysis and interpretation of data enable us to know the precise nature of every problem. Indeed, in almost every field of economics, statistics plays an important role.

(4) Importance for Politicians or in political fields: Statistics has its importance for the politicians as well. In the modern democratic era, politicians play an important role in designing the economic, social and educational policies of the country. It is very essential that politicians be fully aware of the statistical data pertaining to per capita income, unemployment, import and export, black money, public debt, etc., of the country. The opposition parties, on the basis of these statistics, can make constructive criticism of the government and compel it for the revision of its programmes and policies. Policies of the party in power can give wide publicity to the achievements of their government on the basis of statistical data. A popular politician is one who admires or criticises the government on the strength of statistical data.

(5) Importance for Social Reformer or in the Social Field: Statistics is used in solving the social problems also. To a social reformer, statistics relating to such social evils as dowry, alcoholism, gambling, divorce, etc., are of great significance. With the help of the concerned data, they come to know of the gravity of these evils. They can suggest remedies against these evils only when they are equipped with relevant data. Civic problems like shortage of power and water supply can be solved if the seriousness of these problems is underlined with the help of numerical data. Whether the citizens of a country are getting adequate supply of necessaries of life like, food, cloth and shelter, etc., or not, is also known with the help of statistics. If not, then reasons for the same must be traced. It must be ascertained if the income of the people is very low or that they are spending it on liquor or other harmful drugs. One can know answer to all these questions from the data concerned. It is from the statistics relating to standard of living and level of education of the people belonging to scheduled castes and backward classes that the social reformers come to know of their difficulties. In short, no society can formulate plans and projects of its development in the absence of statistics.

(6) Importance in the Field of Science and Research: Statistics has great significance in the field of physical and natural sciences as well as research. Statistics is used both in propounding and verifying scientific laws. Thus, statistics is often used to formulate standards of body temperature, pulse rate, weight, blood pressure, etc., of an healthy person. In every science, research is conducted with the help of statistics and results of the research are also expressed in terms of statistics. Success of modern computers depends on the conclusions drawn on the basis of statistics.

(7) Importance for Banking: To every banker and banking industry, statistics relating demand deposits, time deposits, credit, etc., are of great significance. It is on the basis of d relating to demand and time deposits that the banking system in a country determines its cre policy. Progress of banking industry in a country is measured in terms of statistics concerning to deposits, loans to primary sector, branches of banks, cash reserves, etc. Banks determine their cre policy on the basis of Theory of Probability.

(8) Importance for Insurance Companies: Insurance companies also use statistic information. These companies determine the rate of insurance premium on the basis of statisti relating to average expectancy of life in the country. Expectancy of life is calculated on the basis Life Tables. These tables depend on the Theory of Probability. These tables show that chances remaining alive at younger age are more and so the rate of premium is also low. Life-Expectanc reduces with the age; accordingly premium increases with the age. All this is the play of statistic methods and statistical facts and figures.

(9) Importance in the Field of Education: Progress in the field of education is measured terms of the literacy rate of population, number of schools, colleges and universities in the count and the number of students studying therein. Shortcomings of education system are known from th data relating to examination results of the students. Data concerning male and female educatio adult education, etc. is necessary to formulate any suitable education policy. Statistics regardin teacher-pupil ratio, number of students in each class, number of books issued by the library, etc. ar of great significance for introducing education reforms.

(10) Importance for Economic Planning: According to **Tippet**, *"Planning is the order of th day and without statistics planning is inconceivable"*. Statistics is of prime importance in economi planning. Priorities of planning are determined on the basis of the statistics relating to resource bas of the country and the short-term and long-term needs of the country. Again, success or failure o planning is measured in terms of statistical facts and figures. According to **Planning Commissio** *"Planning for the economic development of the country depends on the maximum use of statistics"*.

◘ LIMITATIONS OF STATISTICS

In modern times, Statistics has come to occupy an important place. However, it has certai limitations. While making use of statistical methods, these limitations are kept in view. In the word of **Newshome**, *"Statistics must be regarded as an instrument of research of great value but barrin severe limitations which are not possible to overcome"*. The main limitations of statistics are a follows:

(1) Study of Numerical Facts only: Statistics studies only such facts as can be expressed in numerical terms. It does not study qualitative phenomena, like honesty, friendship, wisdom, health patriotism, justice, etc.

(2) Study of Aggregates only: Statistics studies only the aggregates of quantitative facts. It does not study any particular unit. For example, if the income of Ram is Rs. 2000 per month, it has no relevance in statistics. But if the income of Ram is Rs. 2000 p.m., that of Sohan is Rs. 3000 p.m., and that of Shyam is Rs. 4000 p.m., then these aggregates will form part of study of statistics. Their average income will work out to be Rs. 3000. This average income will lead to the conclusion that all the three persons belong to middle class. Such a conclusion would not have been possible from the study of Ram's income alone.

(3) Not the only Method: Statistical method is not the only method to study. Many a time this method does not suggest the best solution of each problem. The conclusions drawn on the basis of statistics should be verified with the help of the conclusions drawn with the help of qualitative methods.

(4) Homogeneity of Data: Quantitative data must be uniform and homogeneous. To compare the data, it is essential that whatever statistics are collected, the same must be uniform in quality. Data of different qualities and kinds cannot be compared. For example, production of foodgrains cannot be compared with the production of cloth. It is because cloth is measured in metres and foodgrains in tonnes. However, it is possible to compare their value instead of comparing the volume of their production.

(5) Results are true only on an Average: Laws of statistics are true only on an average. They express tendencies. Unlike the laws of physical science or chemistry, they are not absolutely true. They are not valid always and under all conditions. For instance, if it is said that per capita income in India is Rs. 6000 per annum, it does not mean that the income of each and every Indian is Rs. 6000 per annum. Some may have more and some may have less than it. It is true only on an average.

(6) Without Reference Results may Prove Wrong: In order to understand the conclusions very well, it is necessary that the circumstances and conditions under which these conclusions have been drawn are also studied, otherwise they may prove wrong. For example, in the business of cloth and paper, profits earned during three years may be Rs. 1000, Rs. 2000 and Rs. 3000 respectively. Thus the average profit in both the businesses comes to Rs. 2000 per annum. It may lead to the conclusion that both the businesses have similar economic position, but it is not true. If studied in proper perspective, we will find that whereas cloth-business is making progress, paper-business is on the decline.

(7) Can be used only by Experts: Statistics can be used only by those persons who have special knowledge of statistical methods. Those who are ignorant about these methods cannot make use of it. It can, therefore, be said that data in the hands of an unqualified person is like a medicine in the hands of a quack who may *abuse* it out of ignorance leading to dangerous results. In the words of **Yule and Kendall**, *"Statistical Methods are most dangerous tools in the hands of an inexpert"*.

(8) Misuse of Statistics is Possible: Misuse of statistics is possible. It may prove true what actually is not true. It is usually said, *"Statistics are like clay of which you can make a god or devil, as you please"*. Misuse of statistics is, therefore, its greatest misuse.

(9) Only Means and not a Solution: Some scholars are of the opinion that statistics are only a means in the solution of any problem. It is not a solution to the problem. To check the misuse of statistics, conclusions should be drawn impartially and without any selfish interest. Otherwise, statistics may not become a proper means for the solution of any problem.

In short, while making use of Statistics, its limitations as discussed above, must always be kept in mind.

DISTRUST OF STATISTICS

Inspite of the usefulness of statistics and the confidence of the people in its efficacy, so people have misgivings about it and they distrust it. Those who distrust statistics make the follow observations about it:

(1) In the words of **Disraeli**, *"There are three kinds of lies- lies, damned lies and statistic*

(2) Statistics is a rainbow of lies.

(3) Statistics are tissues of falsehood.

(4) Statistics can prove anything.

(5) Statistics cannot prove anything.

(6) Statistics are like clay of which you can make god or devil, as you please.

It is evident from the above observations that statistics are nothing but bundle of lies and so not trustworthy. The main cause of mistrust is that most of the people believe statistics readi Thus, to take undue advantage of their *credulity*, some selfish people make misuse of the statisti data. They can present the statistics in such a distorted way as to prove right what is wrong a wrong what is right. For instance, the government claimed that in 1994-95, per capita income India increased by about 5 percent and as such economic planning was a success. On the other han the opposition party claimed that is 1994-95, per capita income increased by 1.5 per cent only and such economic planning was a failure. Statistics presented by the government as well as oppositi party are correct, the only snag is that government statistics are calculated at current prices while t statistics presented by the opposition party are calculated at 1980-81 constant prices. Main causes the mistrust of statistics are as under:

(1) Different kinds of statistics are obtained in respect of a given problem.

(2) Statistics can be altered to prove predetermined conclusions.

(3) Authentic statistics can also be presented in such a manner as to confuse the reader.

(4) When statistics are collected in a partial manner, the results are mostly wrong. Consequentl people lose faith in them.

It may be noted that if statistics are presented in a wrongful manner, the fault does not lie wit the statistics. The fault lies with those people who collect wrong statistics or those who draw wron conclusions. Statistics, as such, do not prove anything. They are simply tools in the hands of th statisticians. If the statistician misuses the data, then the blame lies squarely on the statistician an not on the data. A competent doctor can cure the malady by making good use of the medicine but th same medicine in the hands of an incompetent doctor becomes poison. The fault in this case is not the medicine but of the unqualified doctor. In the same way, statistics are never faulty. It pertinently said, *"Figures would not lie, but liar figure"*.

In fact, statistics should not be relied upon blindly nor distrusted outright. *"Statistics should no be used as blind man uses a lamp post for support rather than for illumination, whereas its rea purpose is to serve as illumination and not as a support"*.

In making use of statistics one should be cautious and vigilant. In the words of **King**, *"Th science of statistics is the most useful servant, but only of great value to those who understand it proper use"*.

In short, it is the duty of the students of economics to make use of the knowledge of statistics to seek the truth.

Remedies to Remove Distrust

Following measures may be taken to remove distrust of statistics:

(*i*) **Consideration of Statistical Limitations:** While making use of statistics, limitations of statistics must be taken care of, for instance, statistics should be homogeneous.

(*ii*) **No Bias:** Researcher should be impartial. He should make use only of proper data and draw conclusions without any bias or prejudice.

(*iii*) **Application by Experts:** Statistics should be used only by the experts. If they use it carefully and scientifically, the possibilities of errors will be little.

QUESTIONS

1. Define statistics and discuss its functions and limitations.
2. What is statistics? Explain the importance of statisticsin business world with suitable examples.
3. Explain the functions, importance and limitations of statistics.
4. Discuss the distrust of statistics.
5. Explain the utility of statistics as a managerial tool. Also discuss its limitations.
6. Define Statistics as a subject. Also bring out its scope.
7. Differentiate between descriptive statistics and inferential statistics.
8. "Statistics are numerical statements of facts in any department of inquiry and placed in relation to each other." Comment and discuss the characteristics of Statistics.

2 Collection of Da

INTRODUCTION

Data collection, is in fact, the most important aspect of a statistical survey. The term data as in statistics means quantitative data. In other words, in data, we include that information which capable of numerical expression. Qualitative aspects like intelligence, honesty, good or bad ha significance in statistics until and unless these are assigned some figures. Qualitative aspects w expressed numerically can be studied in statistics.

PRIMARY AND SECONDARY DATA

Statistical data are mainly of two types: (*i*) Primary Data, and (*ii*) Secondary Data.

(1) Primary Data: Data collected by the investigator for his own purpose, for the first ti from beginning to end, is called primary data. It is collected from the source of origin. In the wo of **Wessel** *"Data originally collected in the process of investigation are known as primary dat* Primary data are original. The concerned investigator is the first person to collect this informati The primary data are therefore, a first-hand information. To illustrate, you may be interested studying the socio-economic status of those students studying in BBA-I class who secured fi division in their XII examination. You collect information regarding their pocket allowance, th family income, educational status, their family members and the like. All this information would termed as primary information or primary data, since you happen to be the first person to collect t information from the source of its origin.

(2) Secondary Data: In the words of **M.M. Blair** *"Secondary data are those which are alrea in existence, and which have been collected, for some other purpose than the answering of t question in hand"*. According to **Wessel**, *"Data collected by other persons are called seconda data"*. These data are, therefore, called second-hand data. Obviously, since these data have alrea been collected by somebody else, these are available in the form of published or unpublish reports. For example, data relating to Indian Railways which are annually published by the Railw Board would be secondary data for any researcher.

DISTINCTION BETWEEN PRIMARY AND SECONDARY DATA

The following are the points of difference between primary and secondary data:

(1) Difference in Originality: Primary data are original because these are collected by th investigator from the source of their origin. On the other hand, secondary data are already existence and, therefore, are not original. Primary data are used as raw material while seconda data are finished products.

(2) Difference in the Suitability of Objectives: Primary data are always related to a specific objective of the investigator. These data, therefore, do not need any adjustment for the concerned study. On the other hand, secondary data have already been collected for some other purpose. Therefore, this data need to be adjusted to suit the objective of study in hand.

(3) Difference in Cost of Collection: Primary data are costlier in terms of time, money and efforts involved than the secondary data. This is because primary data are collected for the first time from their source of origin. Secondary data are simply collected from the published or unpublished reports. Accordingly, these are much less expensive.

Of course, it may be noted that there are no fundamental differences between primary data and secondary data. Data are data, whether primary or secondary. These are classified as primary or secondary just on the basis of their collection: first-hand or second-hand. Thus, a particular set of data when collected by the investigator for a specific purpose from the source of origin would be primary data. And the same set of data, when used by some other investigator for his own purpose would be known as secondary data. Thus, **Secrist** has rightly pointed out, *"This distinction between primary and secondary data is one of the degree. Data which are primary in the hands of one party may be secondary in the hands of other".*

◘ METHODS OF COLLECTING PRIMARY DATA

The primary data may be collected by using any of the following methods:

(1) Direct Personal Investigation.

(2) Indirect Oral Investigation.

(3) Information from Correspondents.

(4) Mailed Questionnaire Method.

(5) Schedules sent through Enumerators.

These methods are discussed below:

○ (1) Direct Personal Investigation

In this method, data are collected personally by the investigator. There is a face-to-face contact with the persons from whom the information is to be obtained. Data are collected by asking questions relating to the enquiry to the informants. Suppose, if an investigator wants to collect data about the involvement in politics of the students of Kurukshetra University, Kurukshetra, he would go to the campus and contact each student and obtain the required information.

▶ Suitability

This method is suitable particularly when:

(1) the field of investigation is limited;

(2) a greater degree of originality of the data is required;

(3) information is to be kept secret; and

(4) investigation needs lot of expertise, care and devotion.

▶ **Merits**

(1) Originality: Data have a high degree of originality according to this method.

(2) Accuracy: Data are fairly accurate when personally collected.

(3) Reliable: Because the information is collected by the investigator himself, reliability of data is not doubted.

(4) Other Information: When in direct contact with the informants, the investigator obtain any other related information as well.

(5) Uniformity: There is a fair degree of uniformity in the data collected by the investi himself from the informants. Comparison becomes easy because of uniformity of data.

(6) Flexible: This method is fairly flexible because the investigator can always make neces adjustments in his set of questions.

▶ **Demerits**

(1) Not Proper for Wide Areas: Direct personal investigation becomes very difficult whe area of the study is very wide.

(2) Personal Bias: This method is highly prone to the personal bias of the investigator. result, the data may lose their credibility.

(3) Costly: This method is very expensive in terms of the time, money and efforts evolve

(4) Wrong Conclusions: In this method, area of investigation is generally small. The re are, therefore, less representative. This may lead to wrong conclusions.

○ (2) Indirect Oral Investigation

In the method, the investigator obtains the information not from those persons for whom information is needed. Information is collected orally from other persons who are expecte possess the necessary information. These other persons are known as witnesses. Indirect investigation is usually adopted in those cases where information through direct sources is possible or is less reliable. For example, if a case of murder is to be investigated, it would be impossible to know the facts by contacting the persons directly who are involved in it. In such information is to be obtained from third persons such as friends, neighbours, witnesses, Similarly, if a fire has broken out at a certain place, the cause of the fire may be traced by conta persons living in the neighbourhood of that area.

▶ **Suitability**

This method is suitable particularly when:

(1) the field of investigation is large.

(2) it is not possible to have direct contact with the concerned informants.

(3) the concerned informants are not capable of giving information because of their ignoran

(4) Enquiry committees and commissions appointed by the Government generally adopte method.

▶ **Merits**

(1) Wider Area: This method can be applied even when the field of investigation is very wide.

(2) Less Costly: This is relatively a less costly method.

(3) Expert Opinion: Using this method an investigator can seek opinion of the experts and thereby make his information more reliable.

(4) Free from Bias: This method is relatively free from the personal bias of the investigator.

(5) Simple: This is relatively a simple method of data collection.

▶ **Demerits**

(1) Less Accurate: The data collected by this method are relatively less accurate. This is because the information is obtained from persons other than the concerned informants.

(2) Biased: There is possibility of personal bias of the witness giving information.

(3) Wrong Conclusions: This method may lead to doubtful conclusions due to ignorance and carelessness of the witness.

(3) Information from Local Sources or Correspondents

In this method, the investigator appoints local agents or correspondents in different places to collect information. These correspondents collect the information in their own way and send the same to the central office where the data are processed. Newspaper agencies generally adopt this method. This method is also adopted by various government departments where regular information is to be collected from a wide area. For example, in the construction of wholesale price indices, regular information is obtained from correspondents appointed in different areas.

▶ **Suitability**

This method is suitable particularly when:

(1) accuracy of the data is only modestly needed.

(2) regular and continuous informations are needed.

(3) the area of investigation is large.

(4) the information is to be used by journals, magazines, radio, TV, etc.

▶ **Merits**

(1) Economical: This method is quite economical in terms of time, money or efforts involved.

(2) Wider Coverage: Investigator can cover wider area.

(3) Continuity: The correspondents keep on supplying almost regular information.

(4) Suitable for Special Purpose: This method is particularly advantageous for some special purpose investigations, *e.g.*, price quotations from the different grain markets for the construction of Index Number of Agricultural Prices.

▶ **Demerits**

(1) Less Originality: In this method, there is less originality. Investigation depends more on estimation rather than actual enumeration.

(2) Lack of Uniformity: There is lack of uniformity of data. This is because data is collected a number of correspondents.

(3) Personal Bias: This method suffers from the personal bias of the correspondents.

(4) Less accurate: The data collected by this method are not very accurate.

(5) Delay in Collection: Generally, there is delay in the collection of information through method.

(4) Mailed Questionnaire Method

In this method, a list of questions (known as questionnaire) relating to the survey is prepared sent to the informants by post. The questionnaire contains questions and provides space for answ A covering letter is addressed to the informant explaining the object of survey and making a requ to fill up the questionnaire and send it back within a specified time. It is also assured that information would be kept secret. The informants write the answers against the questions and ret the completed questionnaire to the investigator.

▶ **Suitability**

This method is most suited when:

(1) the area of the study is very wide and

(2) when the informants are educated.

▶ **Merits**

(1) Economical: This method is economical in terms of time, money and efforts involved.

(2) Originality: This method is original and, therefore, fairly reliable. This is because information is supplied by the concerned persons themselves.

(3) Wider area: This method can cover wider areas.

▶ **Demerits**

(1) Lack of Interest: Generally, the informants do not take interest in questionnaires and fail return the questionnaires. Those who return, often send incomplete answers.

(2) Lack of Flexibility: This method lacks flexibility in the sense that when questions are properly replied, these cannot be changed to obtain the required information.

(3) Limited Use: This method has limited use in that questionnaires are answered only by educated informants. Thus, this method cannot be used when the informants are uneducated.

(4) Biased: If the informants are biased, the informations will also be biased.

(5) Less Accuracy: The conclusions based on such investigation have only limited accurac This is because some questions may be difficult and accurate answers may not be possible.

(5) Schedules Filled Through Enumerators

In this method, a questionnaire is prepared as per the purpose of enquiry. The enumerato himself approaches the informant with the questionnaire. **The questionnaires which are filled b the enumerators themselves by putting questions are called schedules.** Thus, under this method the enumerator himself fills the schedules after making enquiries from the informants. Enumerator

are those persons who help the investigators in collecting the data. The enumerators are given training to fill the schedules and put the questions intelligently in the interest of accuracy of information.

▶ **Suitability**

This method is mostly used when

(1) field of investigation is large.

(2) the investigation needs specialised and skilled investigators.

(3) the investigators are well versed in the local language and cultural norms of the informants.

▶ **Merits**

(1) Wide Coverage: This method is capable of wider coverage in terms of the area involved. Even illiterates will also provide information.

(2) Accuracy: There is a fair degree of accuracy in the results. This is because investigations are done by specialized enumerators.

(3) Personal Contact: Unlike in the case of mailing questionnaires, there is personal contact with the informants in this method. Accordingly, accurate and right answers are obtained.

(4) Impartiality: This method is impartial. This is because the enumerators themselves do not need the required information; so they are impartial to the nature of information they obtain.

(5) Complete: Schedules have the merits of completeness, because these are filled in by the enumerators themselves.

▶ **Demerits**

(1) Expensive: This is a very expensive method of investigation because of the involvement of trained investigators.

(2) Difficulties regarding Enumerators: Competent enumerators may not be available. Accuracy of the information accordingly suffers.

(3) Time Consuming: Enumerators may need specialised training for particular investigators. The process of investigation thus becomes time consuming.

(4) Not Suitable for Private Investigation: Since this method is very expensive, it is generally not suitable for private investigations. This method is generally used by the Government institutions.

(5) Inaccurate Data: If the enumerators are biased, the data will not be accurate.

◘ ESSENTIALS/QUALITIES OF A GOOD QUESTIONNAIRE

In the context of collection of primary data, questionnaire has special significance. A questionnaire is a list of questions relating to the field of enquiry and provides space for answers. It may be defined as an instrument of collecting primary data from a large number of persons. The success of the investigation largely depends upon the proper drafting of the questionnaire. Following are some of the notable essentials or qualities of a good questionnaire:

(1) Limited number of Questions: The number of questions should be as limited as possible. Questions should be only relating to the purpose of enquiry.

(2) Simplicity: Language of the questions should be simple and clear. Questions should short and not long or complex. Mathematical questions be avoided.

(3) Proper Order of the Questions: Questions must be placed in a proper order.

(4) No Undesirable Questions: Undesirable questions or personal questions in particular m be avoided. The questions should not offend the informants. Questions likely to offend the perso social and religious feelings of the informants be avoided.

(5) Less Chance of Partially: Questions should be such as can be answered impartially. controversial questions should be asked.

(6) Calculation: Questions involving calculations be avoided. Investigator himself should the calculation job.

(7) Pre-Testing: Before giving the questionnaire a final shape, it should be subjected pre-test. To achieve this objective, some questions be asked from the informants on trial basis their answers involve some difficulty, the same be changed accordingly. Such testing is technica called pilot survey.

(8) Instructions: Clear instructions for filling the questionnaire form be issued.

(9) Cross Verification: Such questions should also be asked as may help cross-verification

(10) Request for Return: Request should be made to return the questionnaire duly filled The informant be assured that the information conveyed by him will be treated as confidential.

◘ METHODS OF COLLECTING SECONDARY DATA/SOURCES OF SECONDARY DATA

The secondary data can be collected from the following two sources:

(I) Published Sources

(II) Unpublished Sources.

○ *(I) Published Sources*

Some of the published sources of secondary data are:

(1) Government Publications: Ministries of the Central and State Governments in Ind publish a variety of statistics as their routine activity. As these are published by the Governme data are fairly reliable. Some of the notable Government publications on Statistics are: *Statistic Abstract of India, Annual Survey of Industries, Agricultural Statistics of India, Report on Curren and Finance, Labour Gazette, Reserve Bank of India Bulletin.*

(2) Semi-Government Publications: Semi-Government Bodies (such as Municipalities a Metropolitan Councils) publish data relating to education, health, births, and deaths. These data a also fairly reliable and useful.

(3) Reports of Committees and Commissions: Committees and Commissions appointed b the Government also furnish lot of statistical information in their reports. Finance Commission Monopolies Commission, Planning Commission are some of the notable commissions in Ind which supply detailed statistical information in their reports.

(4) Publications of Trade Associations: Some of the big trade associations, through their statistical and research divisions, collect and publish data on various aspects of trading activity. For example, Sugar Mills Association published information regarding sugar mills in India.

(5) Publications of Research Institutions: Various universities and research institutions publish information regarding their research activities. In India, for example, **Indian Statistical Institute, National Council of Applied Economic Research** publish a variety of statistical data as a regular feature.

(6) Journals and Papers: Many newspapers such as Economic Times as well as Magazines such as **'Commerce'**, **'Facts for You'** also supply a large variety of statistical information.

(7) Publications of Research Scholars: Individual research scholars also sometimes publish their research work containing some useful statistical information.

(8) International Publications: International organisations such as U.N.O., I.M.F., I.L.O. and foreign governments, etc., also publish lot of statistical information. These are used as secondary data.

(II) Unpublished Sources

There are some unpublished sources as well. These data are collected by the government organisations and others, generally for their self use or office record. These data are not published. These unpublished numerical informations are, however, used as secondary data.

PRECAUTIONS IN THE USE OF SECONDARY DATA

We know, secondary data are collected by others to suit their specific requirements. Therefore, one needs to be very careful while using these data. **Connor** has rightly stated, *"Statistics especially other people's Statistics are full of pitfalls for the users"*. Some of the notable questions to be borne in mind while laying hands at the secondary data are:

(a) Whether the data are reliable?

(b) Whether the data are suitable for the purpose of enquiry?

(c) Whether the data are adequate?

In order to assess the reliability, suitability and adequacy of the data, following points must be kept in mind:

(1) Ability of the Collecting Organisation: One should check the ability of the organisation which initially collected the data. The data should be used only if collected by able, experienced and impartial investigators.

(2) Objective and Scope: One should note the objective of collecting data as well as the scope of investigation. Data should be used only if the objective and scope of the study undertaken earlier match with the objective and scope of the present study.

(3) Method of Collection: The method of collection of data by the original investigator should also be noted. The method adopted must match the nature of investigation.

(4) Time and Conditions of Collection: One should also make sure regarding the period of investigation as well as the conditions of investigations. For example, data collected during war times may not be suitable to generalize certain facts during peace times.

(5) Definition of the Unit: One should also make sure that the units of measurement used in initial collection of data is the same as adopted in the present study. If the unit of measure differs, data must be readusted before use.

(6) Accuracy: Accuracy of the data should also be checked. If the available data do not conform to the high degree of accuracy.

In short, as stated by **Bowley**, *"It is never safe to take published statistics at their face value without knowing their meaning and limitations"*. In using secondary data, the above precautions must be taken.

◘ ROUNDING OFF DATA (OR APPROXIMATION)

Most of the statistical data are approximate figures. The process of approximation (or rounding off data) makes it simple to understand the significance of data. It enables grasp of figures easy, clear and facilitates calculation. If it is stated that Indian population is 68,38,05,051 it will be difficult to remember this figure. On the other hand, if it is said that Indian population is about 68 crores, it will become simple and will easily be remembered. Thereby, there will be no loss of importance. The extent of approximation or rounding off data depends upon the degree of accuracy desired. When approximations are made, the figures should be rounded in such a way to indicate precision about facts.

⊙ Rules of Approximation (or Rounding off Data)

Before rounding off the figures, it should be decided to what extent they are to be approximated *i.e.*, upto three, two or one decimal point or unit, 10, 100, 1,000, 1,00,000 or 1,00,00,000. The various rules of approximation are:

▶ Rule I :

By discarding the digits entirely. According to this method, the digits, after the point to which the figures are to be approximated, are left out entirely. For example, if the figure 43,82,35,282.3759 is to be rounded up according to this method, it will be approximated like this:

upto three decimal points	43,82,35,282.375
" two " "	43,82,35,282.37
" one " "	43,82,35,282.3
" unit	43,82,35,282
" tens	43,82,35,280
" hundreds	43,82,35,200
" thousands	43,82,35,000
" ten thousands	43,82,30,000
" lakhs	43,82,00,000
" ten lakhs	43,80,00,000
" crores	43,00,00,000

Under this method of approximation, the rounded up figures will always be less than the actual figure. The lower is the digit to be removed, the lesser will be the error.

▶ **Rule II :**

By raising the figure to the next higher figure. According to this method, the last digit is raised by one, eliminating the rest of the digits. For example, if the figure 43,82,35,282.3759 is to be rounded up according to this method, the rounded figure will be:

Upto three decimal points	43,82,35,282.376
" two " "	43,82,35,282.38
" one " "	43,82,35,282.4
" unit	43,82,35,283
" tens	43,82,35,290
" hundreds	43,82,35,000
" thousands	43,82,36,000
" ten thousands	43,82,40,000
" lakhs	43,83,00,000
" ten lakhs	43,90,00,000
" crores	44,00,00,000

▶ **Rule III :**

Approximation to the nearest whole number. In this method, the value is unchanged when the remainder to be dropped is less than one-half. It is raised to the next higher digit, if the remainder exceeds one-half. For example, 43,82,35,282.3759 will be rounded as

upto three decimal points	43,82,35,282.376
" two " "	43,82,35,282.38
" one " "	43,82,35,282.4
" unit	43,82,35,282
" tens	43,82,35,280
" hundreds	43,82,35,300
" thousands	43,82,35,000
" ten thousands	43,82,40,000
" lakhs	43,82,00,000
" ten lakhs	43,80,00,000
" crores	44,00,00,000

This method of approximation is regarded as the best, because it is more scientific and reasonable. In such approximation, the approximated value may be more or less than the actual value. The errors being positive as well as negative, they are of compensating nature.

Approximation of percentages. Percentages can also be rounded up like:

	Ist Method	IInd Method	IIIrd Method
45.765%	45.7%	45.8%	45.8%
41.325%	41.3%	41.4%	41.3%

Precautions: The following points should be borne in mind while approximate the figures rounding off data):

(1) As approximation changes the origin figures, they should be rounded up only when importance of the figures does not change.

(2) Figures should be approximated to the lowest point.

(3) The fact of approximation should be mentioned.

(4) Approximation should be done at the last stage.

QUESTIONS

1. Distinguish between primary data and secondary data.
2. Explain the various methods that are used in the collection of primary data pointing out th merits and demerits.
3. What do you mean by secondary data? Mention some of its sources. Explain briefly precautions to be taken while making use of the secondary data.
4. Describe the different methods of collecting primary data indicating the merits and dem of each of them.
5. "It is never to take published statistics at their face value without knowing their meaning limitations"—*Bowley*. Elucidate.
6. What is a questionnaire? What are the essential characteristics of a good questionnair
7. Distinguish between 'primary data' and 'secondary data'. Are the data collected fo census in India and published in the census reports primary or secondary data?
8. Distinguish between primary and secondary data. Explain briefly the various metho collecting primary data.
9. State the basic rules of approximating the data (or rounding off data) and their util statistics.
10. State the advantages of rounding off data in statistics.

Classification of Data: Frequency Distribution

3

INTRODUCTION

After the data have been collected, the next step is to present the data in some orderly and logical form so that their essential features may become explicit. The need for proper presentation of data arises because the mass of collected data in their raw form is often so volumnous, unintelligible and uninteresting that it starts at the face of the reader. The unorganised and shapeless data can neither be easily competent nor interpreted. Therefore, after the collection of data, it is imperative that data are classified and presented in such a way as to bring out points of similarities and dissimilarities in the data.

CLASSIFICATION OF DATA

Classification is the process of arranging the data into different groups or classes according to some common characteristics. In the words of **Connor,** "*Classification is the process of arranging things (either actually or notionally) in groups according to their resemblances and affinities*". According to **Spurr and Smith,** "*Classification is the grouping of related facts into classes*". The process of classification can be compared to the process of sorting out letters and packets in a post office. Just as, in sorting out operation, all collected letters and packets are separated on the basis of a common characteristic, *i.e.*, their destinations, similarly in the process of classification data are classified into various homogenous groups or classes on the basis of similarities and resemblances, *e.g.*, persons living in a certain locality may be classified on the basis of income groups to which they belong, agricultural holdings may be classified on the basis of their sizes and so on. Thus, in the process of classification data are classified into various homogenous groups or classes on the basis of similarities and resemblances.

OBJECTIVES OF CLASSIFICATION

The main objects of classification are as follows:

(1) To condense the mass of data in such a way that their similarities and dissimilarities become very clear.

(2) To facilitate comparisons, *i.e.*, to make the data comparable.

(3) To point out the most important features of the data at a glance.

(4) To present the data in a brief form.

(5) To enable statistical treatment of the data collected.

(6) To make data attractive and effective.

METHODS OF CLASSIFICATION

Broadly, the data can be classified on the following four basis in accordance with characteristics:

(1) Geographical Classification

(2) Chronological Classification

(3) Qualitative Classification

(4) Quantitative Classification.

(1) Geographical Classification: In geographical classification, data are classified on the basis of geographical or locational differences between the various items. For example, we may present the number of firms producing bicycles stateswise as follows.

No. of Firms Producing Bicycles in 2001

State	No. of firms
Punjab	30
Haryana	20
U.P.	25

(2) Chronological Classification: When data are classified on the basis of time, it is known as chronological classification. For example, we may present population figures on the basis of time given in the following manner:

Population of India (1951 to 1991)

Year	Population (in crores)
1951	36.1
1961	43.9
1971	54.8
1981	68.4
1991	84.4

(3) Qualitative Classification: In this type of classification, data are classified on the basis of some attribute or quality such as sex, literacy, religion, etc. This classification may be of two types (*i*) **Simple classification:** When only one attribute is studied, *e.g.*, classification of population according to sex—male or female, this type of classification is called simple classification; (*ii*) **Manifold Classification:** When more than one attribute is studied, it is called manifold classification, e.g. population may be classified as rural and urban. These may be further classified as male or female and still further as educated or uneducated.

(4) Quantitative or Numerical Classification: When data are classified on the basis of some characteristics which is capable of direct quantitative measurement such as height, weight, income, marks, etc., it is called quantitative classification. This type of classification is also called as numerical classification or grouped classification. For instance, students of a college may be classified according to weight as shown in the following table:

Weight (in lbs)	No. of students
70—80	40
80—90	50
90—100	150
100—110	250
110—120	200

In this type of classification, there are two elements: one variable (*i.e.*, weight in our example) and other frequency (*i.e.*, number of students here). **Quantitative classification is the most popular method of classifying the data.** Before we take up a detailed study of quantitative classification, it is necessary to understand the term **'Variable'**.

Variable: The characteristic, which is capable of direct quantitative measurement is called a variable or variate. Height, weight, production, consumption, marks, etc., are all variables. A variable may be either discrete or continuous.

(1) Discrete variable: A discrete variable is that one which takes only isolated or discontinuous values. There are jumps in case of a discrete variable, *e.g.*, number of goals scored in a match is a discrete variable as there can be only 1, 2, 3, etc., goals scored in a match. Intermediate values between 1 and 2 or between 2 and 3 are not possible as there cannot be 1.4 or 2.5 goals in a match. Similarly, the number of children in a family, number of students in a class, fans produced in a factory in a particular year, etc., are examples of discrete variables.

(2) Continuous variable: A continuous variable is one which can take any value in a specified interval. Temperature recorded of patients in a hospital, heights of all BBA students of Kurukshetra University, wages of all workers in a factory are examples of continuous variables.

◘ WAYS TO CLASSIFY NUMERICAL DATA OR RAW DATA

Numerical data or raw data can be classified in any of the two ways:

(I) Ordered Array or Individual Series

(II) Frequency Distribution:

(*a*) Discrete Frequency Distribution or Discrete Series

(*b*) Continuous Frequency Distribution or Continuous Series

○ (I) Ordered Array or Individual Series

An ordered array or individual series is an orderly arrangement of data according to the ascending or descending order of magnitude. So, in order to prepare an array, the only thing to be done is to arrange the data or various values of variable in ascending or descending order of magnitude. An array may be useful if the data are small, but if the variable takes a large number of values, an array becomes unwieldy.

Example 1. Following data relate to the pocket expenses (rupees) of 10 students of B.Com. II class. Array them in ascending and descending order:

50, 20, 30, 15, 45, 40, 35, 25, 20, 43

Solution:

(a)	Pocket Expenses (rupees) of 10 students (In Ascending Order)		(b)	Pocket Expenses (rupees) of 10 students (In Descending Order)	
	15	35		50	30
	20	40		45	25
	20	43		43	20
	25	45		40	20
	30	50		35	15

(II) Frequency Distribution

The frequency distribution is a statistical table which shows the values of the variable arrang in order of magnitude, either individually or in groups, and also the corresponding frequencies s by side. There are two types of frequency distributions:

(*a*) Discrete frequency distribution

(*b*) Grouped frequency distribution.

(*a*) Discrete Frequency Distribution: It is a statistical table which shows the values of variab individually and also the corresponding frequencies side by side. The construction of discr frequency distribution is very simple. In its construction, we count the frequencies of the vario items. To find the frequency of a particular item, we make use of **tally bars**. Each tally bar indica the presence of one value of the item. Tally bars are used in the form of **'Four and Cross Metho** If the value of the item is repeated five times, a cross is put on four lines (𝍸).

Example 2. Twenty students of B.Com. II class secured the following marks in Economics out 50 marks:

11	12	14	11	16	11	17	16	17	14
17	18	20	14	20	17	20	17	14	20

Present the data in a discrete frequency distribution.

Solution:

Formation of a Discrete Frequency Distribution

Marks	Tally bars	Frequency
11	III	3
12	I	1
14	IIII	4
16	II	2
17	𝍸	5
18	I	1
20	IIII	4
	Total	20

(*b*) Grouped Frequency Distribution: It is a statistical table which shows the values of the variable in groups and also the corresponding frequencies side by side. An example of a grouped frequency distribution is given below:

Daily wages (Rs.)	No. of workers
40—50	7
50—60	12
60—70	8
70—80	6
80—90	2
Total	35

Useful terms associated with Grouped Frequency Distribution

For a detailed study of grouped frequency distribution, it is necessary to define and understand the following terms:

(*a*) Class interval, or Class: It is a group of numbers in which items are placed such as 10—20, 20—30, etc.

(*b*) Class frequency: The number of observations falling within a class is called its class frequency. It is denoted by 'f'.

(*c*) Class limits: Each class is located between two numbers. These two numbers constitute class limits. The lowest value of a class is its lower limit and the higher value is termed as upper limit. For example, in the class 10—20, the lower limit is 10 and the upper limit is 20.

(*d*) Class mark (or mid-value): It is the average value of the upper limit (l_2) and the lower limit (l_1). Symbolically,

$$M.V.(m) = \frac{l_1 + l_2}{2}$$

(*e*) Width or Magnitude of the class: The width or size or magnitude of a class is the difference between its lower and upper class limits. Symbolically,

$$i = l_2 - l_1$$

where, i = is the size of the class interval.

GENERAL RULES FOR CONSTRUCTING A GROUPED FREQUENCY DISTRIBUTION OR PROBLEMS IN THE CONSTRUCTION OF FREQUENCY DISTRIBUTION

The following rules are to be observed while forming a grouped distribution or continuous frequency distribution:

(1) Selection of Number of Classes

There are no hard and fast rules about the selection of number of classes. It depends on a number of factors such as (*i*) the number of items to be classified, (*ii*) the magnitude of the class interval

(*iii*) the accuracy desired, (*iv*) the ease of calculation for further processing of data and (
class interval. However, the number of classes should be neither too large nor too small. I
recommended that the number of classes should not be less than 5 or 6 and should not be gre
15 or 20. However there is no rigidity about it.

Prof. **H.A. Sturge** have given a formula by which the number of class interval can be asce

The formula is:

$$k = 1 + 3.322 \log N$$

(Here, k = number of class intervals; N = Total number of observations).

Example 3. If the total number of observations is 1000, the number of class interva
determined by the formula:

$$\begin{aligned} k &= 1 + 3.322 \log N \\ &= 1 + 3.322 \log 1000 \\ &= 1 + 3.322 \times 3 \\ &= 1 + 9.966 \\ &= 10.966 = 11 \end{aligned}$$

Thus, the number of class intervals would be 11, after round up the fractio

(2) Size (or Width) of Class Intervals

The choice of class interval depends on the number of classes for a given distribution a
the data. As far as possible the class intervals should be of equal size. **Prof. Sturge** have
following formula for determining the size of class intervals:

$$\textbf{Size of Class Interval: } i = \frac{\textbf{Largest Value} - \textbf{Smallest Value}}{1 + 3.322 \log N}$$

(Here, N = Total Frequency, i = Size of Class Interval)

Example 4. If the salary of 1,000 employees in a public sector undertaking varied betwee
and 14,000, the size of class intervals according to **Sturge's formula** wou

$$i = \frac{Largest\ Value - Smallest\ Value}{1 + 3.322 \log N}$$

$$i = \frac{\text{Rs.}14{,}000 - \text{Rs.}3{,}000}{1 + 3.322 \log 1000}$$

$$i = \frac{\text{Rs.}11{,}000}{1 + 3.322 \times 3} = \frac{\text{Rs.}11{,}000}{1 + 9.966}$$

$$i = \frac{\text{Rs.}11{,}000}{10.966} = \frac{\text{Rs.}11{,}000}{11}$$

$$i = 1000$$

Therefore, the size of class interval (i) would be Rs. 1000.

The following points must be kept in mind while making a choice of class intervals.

(*i*) As far as possible, class intervals should be such as the class limits are multiples of 5, *e.g.*, 0—5, 5—10, 10—15, 15—20, etc. However, any number can be taken as class interval.

(*ii*) As far as possible, the class interval should be uniform through out the distribution.

(*iii*) As far as possible, every class interval should have a convenient mid point.

(3) Selection of Class Limits

Class limits should be selected in such a way that (a) the mid values of the classes coincide or come very close to the point of concentration in the data (b) the overlapping of classes is avoided (c) the class limits must be stated precisely enough so that there will be no confusion as to what they include.

(4) Kinds of Continuous Series

There is another important problem relevant to constructing a frequency distribution. These relate to the kinds of grouped or continuous series to be formed.

The following are the important kinds of continuous series:

(*a*) Exclusive series

(*b*) Inclusive series

(*c*) Open ended series

(*d*) Mid-value series

(*e*) Cumulative frequency series.

(*a*) Exclusive Series: Exclusive series is that series in which every class interval excludes items corresponding to its upper limit. In this series, the upper limit of one class interval is the lower limit of the next class interval.

For example, in a class interval of 10—15, only such items would be included the value of which is 10 or more than 10 but less than 15. Any item of the value of 15 would be included in the next class interval, *viz.*, 15—20. The following table shows the exclusive series:

Exclusive Series

Marks	Frequency
10—15	4
15—20	5
20—25	8
25—30	5
30—35	4
Total	26

It is clear from this table that the upper limit of a class interval repeats itself as the lower limit of the next class interval. Also, it may be noted that all values corresponding to, say, 15, have been incorporated not in the class interval of 10—15, but in the class interval of 15—20.

(*b*) Inclusive Series: An inclusive series is that series which includes all items upto its up limits, in such series, the upper limit of class interval does not repeat itself as a lower limit of the class interval. Thus, there is a gap between the upper limit of a class interval and the lower limit the next class interval. The gap ranges between 0.1 to 1.0. For example, 10—14, 15—19, 20— etc., represent an inclusive series. Thus, all the items ranging between 10—14 are included in class interval. Likewise, all the items ranging between 15—19 would be included in that cl interval. In short, while in the exclusive series there is an overlapping of the class limits (upper limit of one series being the lower class limit of the next class interval), there is no such overlapp in the inclusive series. Following table shows an inclusive series.

Inclusive Series

Marks	Frequency
10—14	4
15—19	5
20—24	8
25—29	5
30—34	4
Total	26

Conversion of Inclusive Series into Exclusive Series: Inclusive series is used when th some definite difference among the values of various items in the population. In the above tab student has obtained 14.5 or 19.5 marks, these can be expressed only if the inclusive ser converted into an exclusive series. Following steps are involved in the conversion of an inc series into an exclusive series:

(*i*) First, we find the difference between the upper limit of a class interval and the lowe of the next class interval.

(*ii*) In the case of first class interval half of the difference is subtracted from the lower cla half is added to the upper class.

(*iii*) In case of subsequent class intervals half of that difference is added to the upper lir class interval and remaining half to the lower limit of the next class interval.

Using these steps, the above inclusive series can be converted into an exclusive se follows:

Conversion of an Inclusive Series into an Exclusive Series

Marks	Frequency
9.5—14.5	4
14.5—19.5	5
19.5—24.5	8
24.5—29.5	5
29.5—34.5	4
Total	26

(c) Open Ended Series: In some series, the lower class limit of the first class interval and the upper limit of the last class interval are missing. Instead, **less than or below** is specified in place of the lower class limit of the first class interval and **more than or above** is specified in place of the upper class limit of the last class interval. Such series are called 'Open Ended Series'. Thus an open end series is that series in which lower limit of the first class interval and the upper limit of last class interval is missing. The following table shows such a series:

Open Ended Series

Marks	Frequency
Less than 5	1
5—10	3
10—15	4
15—20	6
20 and above	1

In order to determine the limit of the open end class interval, the general practice is to give same magnitude to these class intervals, as is of the other class intervals in the series. However, this practice is adopted when the known magnitudes of different class intervals in the series are equal to each other.

(d) Frequency Series containing Mid-values: Frequency series containing mid-values is that series in which we have only mid-values of the class intervals and the corresponding frequencies. For example:

Mid-value:	5	15	25	35	45
Frequency:	6	5	11	9	8

Such series may be converted into simple frequency series using the following method:

(*i*) First, difference between mid-values is determined; and

(*ii*) Second, the difference so obtained is reduced to half which when deducted from the mid-value gives lower limit of the class interval and when added to the mid-value gives the corresponding upper limit.

Thus,

$$l_1 = m - \frac{i}{2}$$

$$l_2 = m + \frac{i}{2}$$

(Where, m = mid-value; i = difference between mid-values; l_1 = lower limit and l_2 = upper limit)

In the above noted frequency distribution with mid-values, the difference between mid-values $i = 15 - 5$. Half of it is 5. Deducting 5 from each mid-value we get lower limits and adding 5 to each mid-value we get the corresponding upper limits.

The following example shows the frequency distribution with mid-values: