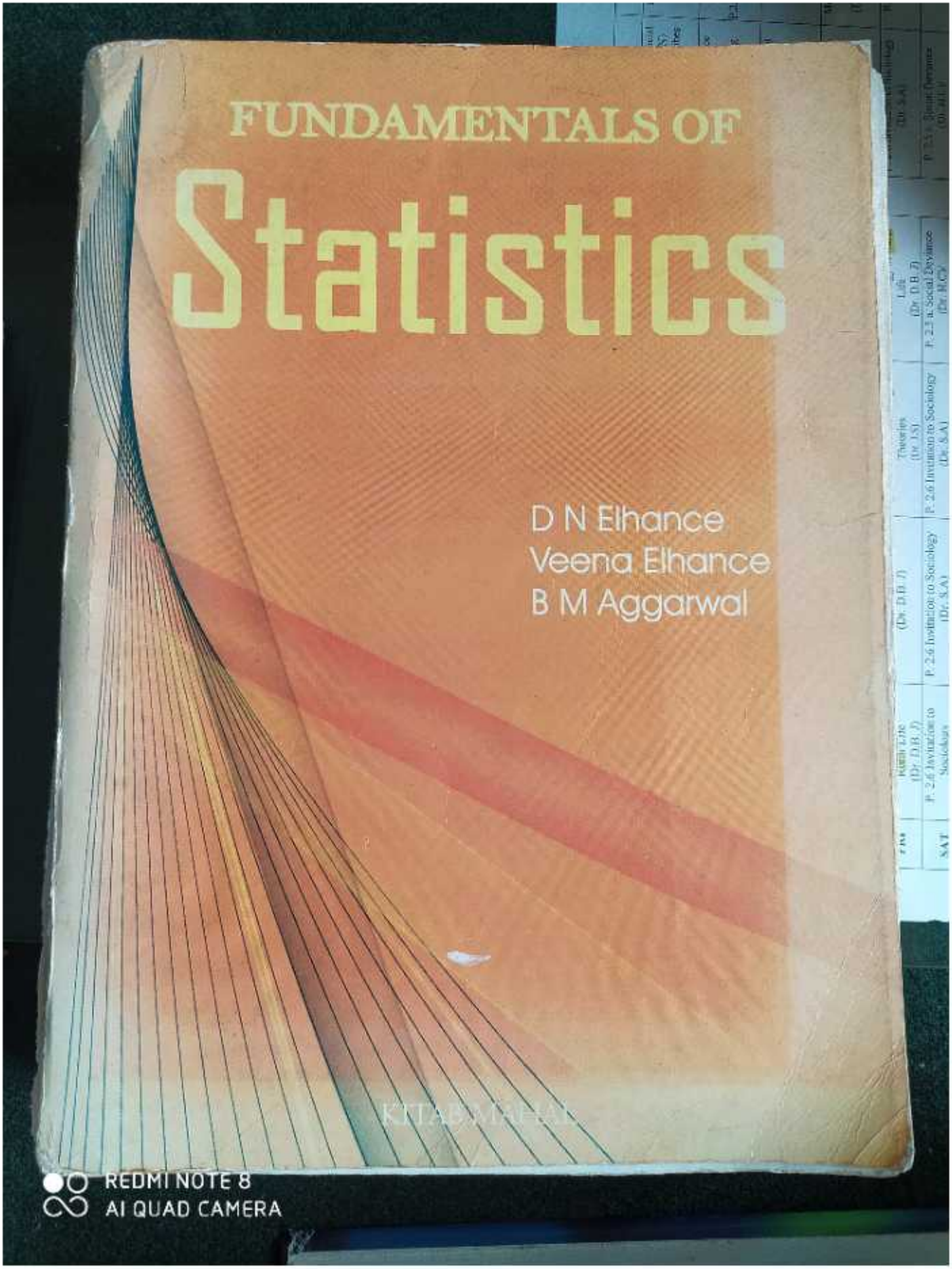
FUNDAMENTALS OF
Statistics
D N Elhance
Veena Elhance
B M Aggarwal
KITAB MAHAL
REDMI NOTE 8
AI QUAD CAMERA

1
STATISTICS – A CONCEPTUAL FRAMEWORK

1.1 INTRODUCTION

Today, Statistics is a very commonly used word but, surprisingly enough, it is understood by different people in different senses. To some, the word statistics connotes just a set of figures like the number of children born in India, in various months of a particular year and their sex. Some people think of statistics as something used to reinforce a qualitative statement. Thus, when someone says that the economic condition of the Indian masses has improved during the last five years and gives figures of rising per capita income during this period, he is using statistics to support a qualitative statement. To many others, statistics means the representation of a phenomenon with the help of figures, charts, diagrams and pictograms, etc. There are many who think of Statistics as a complex set of relationships between a number of variables and also as a technique by which they can reduce the element of uncertainty in their decision process. It is seen by many as a device to achieve a degree of precision in the concept and theories of social sciences which, by nature, are inexact.

However, if we analyse the way the word statistics is looked at, we find that broadly speaking, there are two categories—one, in which the word refers to a set of figures and the other in which it refers to a set of techniques and methods. These categories are elaborated further in the foregoing paragraphs.

Data Set and Data Point

As pointed out above, in common parlance, the word statistics denotes some sort of numerical data. If, for example, somebody says that he has studied the statistics of man-hours lost by the Indian cotton mills due to strikes, or that he has seen the statistics of automobile accidents in the U.S.A., he refers to the numerical figures or data relating to these phenomena. In this sense, statistics are numerical descriptions of the quantitative aspects of things. They take the form of counts or measurements. Statistics about the membership of a certain hostel, for example, include a count of the number of members and separate counts of the number of members of various kinds, as post-graduate and undergraduate or over and under 21 years of age. They might include such measurements as the weights and heights of the members of the numbers computed from such counts or measurements, for example, the proportion of members who are married or the ratios between weights and heights. *The use of the word statistics in this sense is always in plural.* However, any figure or set of figures cannot be called statistics irrespective of any other consideration. Many things are taken into account before using the word statistics for any group of figures. We shall discuss these a little later.

The use of the word statistics in the above sense is, in our opinion, not very correct. A more appropriate word to indicate numerical facts is ***data*** *and, as far as possible, this word should be used in place of statistics in this sense.*

Levin has rightly said that *data are collection of any number of related observations.* We can collect the number of telephones installed in a given day by several workers or the number of telephones installed per day over a period of several days by one worker, and call the results our data. *A collection of data is called a data set and a single observation is data point.*

Statistical Methods

The second sense in which the word statistics is used refers to the techniques and methods used in the collection, analysis and interpretation of data. *In this sense, the word is used in singular.* Statistical methods are applicable to a very large number of fields. They are used in social sciences like economics, anthropology, sociology and psychology. Even professional disciplines, like medicine, engineering and business, rely heavily on statistical methods. Statistical methods are applied, to a limited extent, even in the realm of physical sciences which are by nature exact and experimental in character. *In fact, statistical methods provide a set of tools* which can be profitably used by different sciences in the manner in which they deem fit.

Statistical methods range from a very simple set of devices to highly complicated and complex mathematical procedures which can be used only by those who have received adequate training in these areas.

Statistical methods and experimental methods : Statistical Methods include all those devices which are used in collection and simplification of numerical data so as to render them capable of being analysed, and commonly understood without much difficulty. Statistical methods are different from experimental methods in as much as the latter are more accurate and precise than the former. In experimental methods, it is possible for us to study the effects of any one of the many factors affecting a phenomenon individually by making the other factors inoperative for the time being. Thus, in physics, it is not difficult to study the effects of only heat on the density of air by making other factors inoperative for the duration of study. But, the same thing is not possible in statistical methods. It is not feasible to study the effects of, say, only inflation on prices. The effects of inflation cannot be separately studied, from the effects of many other factors like demand, supply exports and imports, etc. However, by the use of statistical methods, it is possible to have a rough idea of the effects of inflation upon prices. Statistical study cannot be as accurate as the study done by experimental methods. Thus, we see that statistical methods are comparatively less accurate and are usually applied in inexact sciences like sociology, though even in physical sciences (which are classed as exact sciences) the use of these methods is, sometimes necessary. Statistical methods are, thus, of universal application though their primary field is social sciences.

Thus, *"Statistics are numerical facts, but Statistics is a body of methods for making decisions when there is uncertainty arising from the incompleteness or the instability of the information available. The decisions may be made either for the practical purpose of selecting a course of action or for the scientific purpose of gaining general knowledge."*

1.2 DEFINITION

The term Statistics has been defined differently by different authors. Some authors have defined the word as used in the first sense (of numerical data) while others have defined it as used in the second sense (of statistical methods or the science of statistics).

First Type

Of the first type of definitions, the one given by **Horace Secrist** is the most exhaustive. It is as follows :

"By statistics, we mean aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

This definition makes it clear that statistics (as numerical data) should possess the following characteristics :

1. They should be aggregates of facts : Single and unconnected figures are not statistics. A single age of 25 years or 40 years is not statistics but a series relating to the ages of a group of persons would be called statistics. A single figure relating to birth, death, purchase, sale, accident, etc., does not form statistics though aggregates of figures relating to births, deaths, purchases, sales, accidents, etc., would be called statistics because they can be studied in relation to each other and are capable of comparison. It is possible to study them in relation to time, place and frequency of occurrence.

2. They should be affected to a marked extent by multiplicity of causes : Usually, statistical facts are not traceable to a single cause. Since statistics are most commonly used in social sciences, it is only natural that they are affected by a large variety of factors at the same time. It is usually not possible to study the effects of any one of these factors separately as is the case in experimental methods. In statistical methods, the effects of various factors affecting a particular phenomenon are generally studied in a combined form though attempts are also made to study the effects of different sets of factors separately as well. Most of the statistics, however, are affected to a considerable degree by multiple causation. For example, statistics of prices are affected by conditions of supply, demand, exports, imports, currency circulation and a large number of other factors.

3. They should be numerically expressed : Qualitative expressions like good, bad, young, old, etc., do not form part of statistical studies unless a numerical equivalent is assigned to each such expression. If it is said that the production of wheat per acre in 2003 was 20 quintals and in the year 2008 it was 25 quintals, or if it is said that of two persons **A** and **B**, **A** is 20 years old and **B** is 60 years old, we shall be making statistical statements.

4. They should be enumerated or estimated according to reasonable standards of accuracy : Numerical statements can either be enumerated in which case, they are supposed to be accurate and precise or they can be estimated by some expert observers. Where the scope of statistical enquiry is very wide or where the numbers are very large, enumeration is usually out of question and, in such cases, figures can only be estimated. It is obvious that estimated figures cannot be absolutely accurate and precise. The degree of accuracy expected in such figures depends, to a large extent, on the purpose for which statistics are collected and also on the nature of the particular problem about which data are being collected. There cannot be a uniform standard of accuracy for all types of enquiries. For example, if the heights of a group of individuals are being measured, it is all right if the measurements are correct to a centimeter but if we are measuring the distance from Mumbai to Kolkata, a difference of a few kilometers even, can be easily ignored.

5. They should be collected in a systematic manner : If figures are collected in a haphazard fashion, one cannot be sure about the degree of accuracy of such data. It is, therefore, essential that statistics must be collected in a systematic manner so that they may conform to reasonable standards of accuracy.

6. They should be collected for a predetermined purpose : It is obvious that if statistical data are not collected with some predetermined aim, their usefulness would be almost negligible. Figures are usually collected with some end in view, as without it all the efforts made in the collection of figures would be completely wasteful and the figures so collected would not be in any way useful.

7. They should be placed in relation to each other : Statistics are collected mostly for the purpose of comparison. If the collected figures are not capable of being compared with each other, they lose a very large part of their value. It is necessary that the figures which are collected should be a homogeneous lot because it is not possible to compare figures which are of a heterogeneous character and which cannot be placed in relationship to each other. If, for example, the height of a person and the money spent by him in getting his house constructed are placed together, it does not make any sense and the figures cannot be compared to each other. Such figures, naturally, do not come under the category of statistics.

Webster has also defined statistics in the same sense in which **Secrist** has defined it. Webster's definition of statistics is as follows :

"Statistics are the classified facts representing the conditions of the people in a State... specially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement."

This definition is rather narrow. It confines statistics only to those facts which relate to the condition of the people in a State. This was a very old concept of the word statistics and it does not suit modern conditions. At present, statistics relate to all aspects of human activity and, as such, this definition falls short of the modern concept of the term. Moreover, this definition is not as clear and exhaustive as the one given by *Secrist.*

Second Type

Of the second type of definitions of the term statistics (as statistical methods or science of statistics), the one given by **Seligman** is very short and simple and yet quite comprehensive. According to Seligman,

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

According to **King,** *"the science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates."* This definition is not very exhaustive and it limits the scope of the science of statistics. The author himself admits this defect but is of the view that, for practical purposes, it is all right.

A.L. Bowley has given a series of definitions but most of the definitions given by him are not complete and lay emphasis only on some of the aspects of the science. At one place, Bowley says, *"Statistics may be called the science of counting."* At another place, he is of the view that *"Statistics may rightly be called the science of average."* Both these definitions are defective as the science of statistics does not confine itself either to counting or to averaging alone. These are, no doubt, important statistical methods but they do not cover the entire field of the science of statistics. Yet another definition given by the same author characterises statistics as *"the science of measurement of the social organism regarded, as a whole, in all its manifestations."* Obviously, this definition limits the application of the statistical methods to only one field, namely, sociology. Bowley realised this limitation and he himself writes at another place that statistics cannot be confined to any one science.

Boddington has defined statistics as the science of *"estimates and probabilities."* This definition gives expression only to certain methods by which conclusions are derived in this science. No doubt, in most of the cases, statistics are, 'estimates', and 'probabilities' but it should be remembered that the scope of the science is not confined merely to these things.

Lovitt defines the science as *"that which deals with the collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomena"*. This definition is fairly satisfactory and it indicates that the science of statistics is a simple and scientific exposition of statistical methods.

Wallis and Roberts have given a very simple definition. They say that *"Statistics is a body of methods for making decisions in the face of uncertainty"*. **Ya Lun Chu** has also given a similar definition when he says that *'Statistics is a method of decision-making in the face of uncertainty on the basis of numerical data and calculated risk"*. Both these definitions are brief and do not identify the statistical methods and, as such, can embrace any technique used to reduce the risk associated with uncertainty in the process of decision-making. The definitions are too general to give any specific idea about statistical methods.

Having briefly discussed some of the definitions of the term statistics and having seen their drawbacks, we are now in a position to give a simple and complete definition of the term in the following words :

Statistics (as used in the sense of data) are numerical statements of facts capable of analysis and interpretation and the science of statistics is a study of the principles and methods used in the collection, presentation, analysis and interpretation of numerical data in any sphere of enquiry.

1.3 MAIN DIVISIONS AND NATURE OF STATISTICS

Statistics, as a science, can be divided into two main classes, namely, **statistical methods and applied statistics.**

1. Statistical Methods : Under statistical methods are studied all those devices, rules of procedure and general principles which are applicable to all kinds or groups of data. Thus, they include all the general principles and techniques which are commonly used in the collection, analysis and interpretation of data relating to any sphere of enquiry. Statistical methods are the tools in the hands of a statistical investigator. These are devices for achieving the desired ends explained in theory. Since a method is always a means to an end, its accuracy and precision depends on the object which is desired to be achieved and this, in turn, is considerably affected by the peculiar features of the problem to which it is related. This is the reason why different statistical methods are used in different types of enquiries and no uniform standard of accuracy is desired to be achieved in different types of investigations.

2. Applied Statistics : Applied statistics deal with the application of statistical methods to specific problems or concrete forms. If we have to estimate the national income of a country or its industrial or agricultural production, then the special techniques followed to achieve these ends and the results obtained thereof would form part of applied statistics. As is clear from the above explanation, applied statistics can be further divided into two main groups. They may be either **descriptive** or **scientific**.

(*i*) **Descriptive applied statistics** deal with data which are known and which naturally relate either to the present or to the past. For example, business statistics are descriptive applied statistics, as they deal with the analysis, measurement and presentation of business facts relating to past or present. On the basis of these facts, decisions about various business problems are usually taken.

(*ii*) **Scientific applied statistics** deal with the formulation of physical and psychological laws on the basis of quantitative data collected for descriptive purposes by the use of appropriate statistical methods. If, for example, by the use of some business statistics we are in a position to derive certain conclusions, which we use for forecasting the future trend or tendency of that particular phenomenon, we are making use of scientific applied statistics. For purposes of business forecasting, we have to make use of such statistics.

Is statistics a science or an art is a question which is very often asked but not adequately answered. *Science* refers to a systematic study of knowledge and studies cause and effect relationship between different variables and, on the basis of analysis, attempts to draw generalisations which are popularly known as *principles* or *laws* which are supposed to have a very high degree of precision. It is on account of the high degree of precision of scientific laws that man could safely land on moon. *Art*, on the other hand refers to the *skill* of collecting and handling of data to draw *logical inference* and arrive at certain conclusions. The degree of precision in such inferences is not as high as it is in scientific laws.

Judged on the basis of the above definitions of the words *science* and *art*, statistics could lay its claim on both of them. However, statistics is not a science in the same sense in which physics or astronomy are, for the reason that the degree of precision arrived at in statistics is not as high as it is in natural or physical sciences. The reason for this is that statistics deals with such variables whose *individual effects* cannot be studied in isolation as is possible in natural and physical sciences, which are experimental in nature. Under such circumstances, *it would be more appropriate to call statistics not as a science but as a scientific method.* Statistical methods can be and are applied in all sciences, natural or social, and, as such, they should be regarded as indispensable tools for the study of any quantitative phenomenon.

Statistics is an *art* in as much as it is concerned with the skill of handling facts and figures for the purpose of analysis, interpretation and policy formulation. Researches are going on to achieve a higher degree of precision in the collection and analysis of data to arrive at more meaningful conclusions so that future policy formulation may become more rational and dependable.

In the words of **A. L. Boddington**, "*the ultimate end of statistical research is to enable comparison to be made between past and present results, with a view to ascertaining the reasons for changes which have taken place and the effect of such changes in the future.*"

To achieve the above mentioned ends, data relating to past and present, are collected and presented in the shape of time series from which valuable conclusions are drawn and these conclusions are used far the purpose of forecasting the future trend of different problems. Collection, presentation, analysis and interpretation of statistical data are no easy tasks. Latest statistical methods have to be applied for arriving at correct and dependable conclusions. Researches have been going on for improving statistical methods with a view to make them more accurate and precise, so that the laws based on the analysis of the descriptive applied statistics may become comparatively more stable and dependable. It is, thus, very obvious that the science of statistics is very closely associated with the progress of human civilization. It helps in assessing the results of past achievements of human activities and it is also useful for making forecasts about the future courses of events.

1.4 ORIGIN AND GROWTH OF STATISTICS

Early Beginnings

The origin of statistics is suggested by the derivation of this word. It seems to have been derived from the Latin word ***Statis*** which means a political state. In fact, the origin of statistics was due to administrative requirements of the state. Statistics, in the past, were a by-product of administrative activity. Administration of the states required the collection and analysis of data relating to population and material wealth of the country for purposes of war and finance. The earliest form of statistical data, therefore, relate to census of population and property; collection of data for other purposes, however, was not entirely ruled out. Perhaps, one of the earliest census of population and wealth was held in Egypt as early as 3050 B.C. for the erection of pyramids. *Rameses II* conducted a census of all lands of Egypt. During the Middle Ages, such censuses were held in

maintained that police administration has become more rather than less efficient.

2. Performances of the First Five-year Plan : It was claimed by the critics of the First Plan of India that per capita income declined from Rs 266.5 at its beginning to Rs. 255.0 at its end. This led to the erroneous conclusion that the economy suffered deterioration during this period and the Plan was a failure. But, in fact, this decline had occurred due to a fall in the general price level and real per capita income had indeed increased. This could be shown by making comparisons of per capita incomes in both the years at 1948-49 constant prices that it had increased to Rs. 267.8 at the end of the Plan as compared to Rs. 247.5 at its beginning.

Inappropriate Comparison

1. Deaths in Hospitals : The statement that 'the incidence of death among sick persons is higher in hospitals than at home' is lively to lead to the conclusion that more patients die in hospitals than at home due to lack of proper treatment and care. But, this conclusion turns out to be completely erroneous if it is borne in mind that in India only seriously ailing persons are hospitalised.

2. It was claimed by a teacher that his teaching method was superior to that of others. He supported his argument by showing that all the students in his class secured first class. Investigation into the matter revealed that unlike others, all his students had secured first class in previous examination and were merit holders. His success was, therefore, due to better stuff in his class rather than to the superiority of his teaching method.

Defective Method in Selecting Cases

Issue of Abortion : The Morning News reported that 70 per cent people in the country were in favour of legalisation of abortion. It had come to this conclusion by a statistical analysis and interpretation of the replies sent to it by its readers in response to a questionnaire. But, a broad based survey made by a social organisation showed that this was entirely incorrect. In fact, more than 80 per cent people were against it. The newspaper had reached the erroneous conclusion as it was based on the opinion of educated people who constituted only a small minority in the population.

1.8 DISTRUST OF STATISTICS

Figures may be incomplete or manipulated. Despite its importance and usefulness the science of statistics is looked upon with a suspicious eye and is quite often condemned as a tissue of falsehood. It is said that "*an ounce of truth will produce tonnes of statistics*" or that "*Statistics are lies of the first order*". These statements indicate the extent to which the science of statistics has come in disrepute and is not trusted even in modern times when its use has spread over all types of human activities. In our daily life, we tend to accept statistical conclusions and the interpretations placed on them uncritically. But, then we are misled so often by skilful talkers and writers who deceive us with incorrect facts that we come to distrust statistics entirely and assert that "*Statistics can prove anything*"— implying, of course, that "*Statistics can prove nothing.*" Strangely enough, whereas, on the one hand, statistics is condemned in such bitter language, on the other hand, it is also said : "*If figures say so, it cannot be otherwise*" or "*figures don't lie*". The reason for such diversity in views is not far to seek. The reason lies in the innocence of figures. Figures are innocent and easily believable. It is human psychology that when facts which are supported by figures come before a man they are easily believed. Numerical data convey a sense of precision and accuracy and consequently it is only natural that a man believes a statistical statement usually without questioning it. There is a great danger in this type of approach to a numerical statement. Figures which support a particular statement may not be true. They may be incomplete, inaccurate or deliberately manipulated by prejudiced persons who wish to conceal the truth and want to present a false picture to achieve a particular end. Later on when people realise that even a statistical statement has belied their expectation, their faith in the science of statistics is shaken and they begin to condemn it in the strongest possible language. The fault, in such cases does not lie with the science of statistics; it lies with those who use it. If wrong figures have been used they are bound to give wrong conclusions and it is the duty of the persons who use statistics to see that the figures that they use are free from all types of bias and have been properly collected and scientifically analysed.

Can prove anything is not correct : Sometimes, it is remarked that statistics can prove anything. But, people who say so are usually those who do not know the A. B. C. of the subject. Statisticians rarely claim to prove anything. They are taught to examine the reliability of their data and the justification of their conclusion with the utmost suspicion, due care and caution. Statisticians generally take care that the chances of their statement being correct are at least 20 : 1 and as such it is absolutely wrong to say that statistics can prove

anything.

Does not prove anything, is merely a tool : Many people disbelieve statistics because it does not prove a particular thing in a particular manner. It should be clearly understood that statistics does not prove anything. Statistics is only a method of approach; it is a tool in the hands of a statistician to present a phenomenon in a particular manner and nothing beyond it. The science of statistics does not prove or disprove a thing, it merely presents the true facts about a problem and leaves the rest to other people. Different types of conclusions can be arrived at from the same set of figures if there is a difference in the approach of various persons. From one set of figures, a communist can prove that Russia has eliminated unemployment and improved the lot of the working class and from the same set of figures an anti-communist can derive an opposite conclusion. This fundamental difference in approach or we may call it bias in the minds of the investigators, has been responsible for different conclusions being drawn from the same set of figures. For this, the science of statistics cannot be blamed. It is not the fault of the science. It is the mischief of those who use it.

Need of caution : A layman has, therefore, to be very cautious. If figures have been given without the context in which they were collected or if they are not complete, or if they relate to a phenomenon different from the one under investigation, or even if the figures are correct and complete but a faulty or biased logic is applied to them, the conclusions arrived at are bound to be wrong and would strengthen the belief that statistics are lies of the first order. Unfortunately, a set of figures cannot by itself disclose whether it is dependable or not. Figures do not bear a trademark of their accuracy. All figures appear to be correct and innocent. It is this difficulty of separating correct figures from incorrect ones which is responsible for discrediting the science to a considerable extent. It is, therefore, necessary that whenever we use statistics we should, first of all, make sure that they were properly collected and are suitable for the problem under investigation.

Statistical methods are delicate tools : Statistical methods are very delicate and since they are liable to be misused easily they are very dangerous as well. The results of the misuse of statistical methods or statistical data should not be used to discredit the science. If a child cuts his finger by a sharp knife or an insane person hits his own head or that of any one else with a stick, the fault does not lie with the knife or the stick. It lies with the person who uses it. Similarly, if statistical methods are not properly used the fault does not lie with the science of statistics but with the person using it. Statistics are tools and can be used in any way we like and it is in our own interest that we use them in a proper manner.

"He who accepts statistics indiscriminately will often be duped unnecessarily. But, he who distrusts statistics indiscriminately will often be ignorant unnecessarily. There is an accessible alternative between blind gullibility and blind distrust. It is possible to interpret statistics skilfully. The art of interpretation need not be monopolized by statisticians, though, of course, technical statistical knowledge helps. Many important ideas of technical statistics can be conveyed to the non-statistician without distortion or dilution. Statistical interpretation depends not only on statistical ideas but also on ordinary clear thinking. Clear thinking is not only indispensable in interpreting statistics but is often sufficient even in the absence of specific statistical knowledge. For the statistician, not only death and taxes but also statistical fallacies are unavoidable. With skill, common sense, patience and above all objectivity, their frequency can be reduced and their effects minimised. But, eternal vigilance is the price of freedom from serious statistical blunders."

— Wallis and Roberts

1.9 FUNCTIONS OF STATISTICS AND STATISTICIAN

At this stage, it would be worthwhile to examine and identify the functions of the science of statistics and the role which statisticians should play. The functions of the science of statistics are beautifully summed up by **Robert W Burgess** in the following words :

"The fundamental gospel of statistics is to push back the domain of ignorance, rule of thumb, arbitrary or premature decisions, tradition and dogmatism and to increase the domain in which decisions are made and principles are formulated on the basis of analytical quantitative facts."

If we analyse the above statement, we would find that the main functions of the science of statistics are as given below :

1. To collect and present facts in a systematic manner : One of the most important functions of statistics is to collect facts and figures in a systematic manner and to present them in such a form that they are intelligible and readily understood. Collection of data involves the use of many scientifically developed

techniques, as a haphazard collection of data may give us a set of figures whose analysis would lead to erroneous conclusions. Therefore, the data collection has to be done very carefully. The data collected has to be presented in a concise form. The data collected cannot be easily understood if its mass is very great and in such a situation it requires *condensation* for proper comprehension. Human mind is not capable of assimilating huge facts and figures and statistical methods by making those data easily intelligible and readily understandable, render a great service because in its absence the data collected would not have been of any use. Statistical methods describe a phenomenon in a very simple fashion. If the production figures of all textile mills in our country are available for the last 10 years it would be impossible to have a precise idea about the variations in production. However, if the figures of average production per year of all the mills are available the figures would look more meaningful.

The data collected have to be presented in a form in which a comparative study of a problem can be made. For example, it is not enough to say that the man-days lost in the strike in textile mills in India in the year 1982 was, say, 50,000. This figure has to be compared with a similar figure of the year 1981 or 1980 to make it meaningful. Similarly, the production figures of Iron and Steel Industry of India for one year do not convey much sense unless they are compared with figures of production of the previous year or with production figures of the same year of another country – say, Japan.

Thus, one important function of statistics is to *collect figures scientifically*, to *condense them* for purpose of understanding and analysis and to *present* them in a manner *fit for a comparative study*.

2. To help in the formulation and testing of hypothesis : Statistical methods are very useful in formulating and testing various types of hypothesis. In the field of social sciences, the importance of hypothesis formulation and their testing is very important. By the use of statistical methods, we can test the hypothesis whether Indian consumers are really brand loyal, whether the recent credit squeeze has affected the price level, whether a rise in the Railway fares has affected passenger traffic and whether there is delegation of authority in Public Sector Enterprises. Instances of such hypothesis being tested can be multiplied. In fact, in most of the researches in social sciences some hypotheses are formulated and by collecting, analysing and interpreting empirical data the validity of the hypotheses is tested.

3. To help in forecasting of certain types of events and thereby to help in the formulation of suitable strategies and policies : In most organisations, plans of development are prepared much in advance of the time when they have to be implemented. Even at the governmental level developmental plans are prepared in anticipation of what is likely to happen in future. Our Five-year Plans are an example in this context. We think of 5 years hence and try to formulate our policies in anticipation of our expectations. Statistical methods are very helpful in forecasting future trends on the basis of the analysis of past data, as modified in the light of current conditions. Statistical methods help in the formulation of future policies by analysing past and present conditions and making projections for future. If, for example, the government has to decide as to how many motor cars should be manufactured in the country in the year 1986, it would take into account the current demand, the price factor, the position of fuel supply, the road development plans, the buying capacity of people and many other factors which would need statistical analysis before a rational decision is arrived at.

4. To enlarge human experience and to enable man to make rational decisions : The science of statistics enlarges human experience and knowledge by making it easier for man to understand, describe and measure the effects of his action on the action of others. Many fields of knowledge would have ever remained closed to mankind but for the efficient and refined technique and sound methodology provided by the science of statistics. *It has provided such a master key to mankind that we can use it anywhere and can study any problem in its correct perspective and on right lines.*

We have discussed above that the main function of statistics is to collect and present numerical data in a systematic manner so that it may be analysed in a scientific way. Statistics is, as we have seen, not meant to prove anything; it is merely to analyse the phenomena in a scientific fashion. Accordingly, the role of a statistician is to collect the data in a proper fashion, to scientifically analyse it and to set a stage for its correct interpretation. It is futile to expect him to work wonders or to give a particular shape to given material. He has simply to arrange the material in a proper form so that its real worth may be exposed. After doing this, the statistician has finished his job. The task of giving a particular shape to material is beyond the scope of the science of statistics. Statistics are like raw materials and to convert them into finished products is the work of people other than statisticians. The use of economic statistics for the purpose of formulating an economic

policy is the work of an economist, not of a statistician. A statistician would simply collect and analyse the economic statistics. He would not formulate an economic policy on their basis; he would leave this work for the economist.

In general, a successful statistician requires not only a sound knowledge of statistical methods but he has to be a specialist in the branch in which he is carrying on an investigation. If a statistician is asked to find out if fertilizer A is better than fertilizer B when applied to potatoes, he should first know all about fertilizers in general, about their application, growing of potatoes and many other connected things. In this case, he should be an agricultural expert. Practical statistician is, therefore, in the first place, an engineer, economist, a biologist or some other specialist and he has to acquire special knowledge of the field in which he is making use of statistical methods.

A statistician should also not forget the limitations of the science of statistics. He should not forget that laws of statistics are true only on an average, and that he cannot boast of the same precision which is found in experimental methods. He has to work under various handicaps and be should be very cautious and vigilant. Even a slight mistake on his part is liable to render his entire work useless and defective. He should be free from bias, should have profound common sense and should work like a true researcher without any preconceived notions or conclusion about the problem under investigation. It should not be forgotten, as **W.I. King** said, that *"Statistics is a most useful servant but only of great value to those who understand its proper use."*

REVIEW EXERCISES

1. Comment on the following statements :
   (*a*) "Statistics is the science of averages."
   (*b*) "Statistics is the science of counting."
   (*c*) "Statistics is the science of estimates and probabilities."
2. "Statistics is a body of methods for making wise decisions in the face of uncertainty". Comment on the statement bringing out clearly how does statistics help in business decision-making.
3. "Statistics are like clay of which you can make a God or Devil, as you please". Explain.
4. "Statistics are numerical statements of facts but all facts numerically stated are not statistics." Comment upon the statement and state briefly which numerical statements of facts are not statistics.
5. "Statistical methods are most dangerous tools in the hands of the inexpert." Elucidate.
6. Science without statistics bears no fruit and statistics without sciences have no root." Explain the above statement with necessary comments.
7. Statistical thinking will, one day, be as necessary for efficient citizenship as the ability to read and write".
8. Discuss briefly the role of statistical methods in economic planning with special reference to India.
9. Write a note on the importance of statistics to a businessman, an economist, a social worker and the Government.
10. Discuss the meaning and scope of statistics, bringing out its importance particularly in the field of trade and commerce.
11. Discuss briefly the possible applications of statistical methods in business, pointing out the limitations, if any.
12. Discuss the importance of the study of statistics and how it can help the extension of scientific knowledge the establishment of a sound business and the introduction of social and political reform.
13. "Statistics should not be used as a blind man does, a lamppost for support instead of for illumination". Discuss.
14. Without adequate understanding of statistics, the investigator in social sciences may frequently be like the blind man groping in a dark closet for a black cat that is not there". Comment. Can the statement be extended to the field of natural sciences also ?
15. Define statistics. Why do some people look at this science with an eye of distrust ?

3
COLLECTION OF DATA

3.1 PRIMARY AND SECONDARY DATA

Statistical data, as we have already seen, can be either primary or secondary. Primary data are those which are collected for the first time and are, thus, **original** in character, whereas secondary data are those which have already been collected by some other persons and which have passed through the statistical machine at least once. Primary data are in the shape of raw materials to which statistical methods are applied for the purpose of analysis and interpretation. Secondary data are usually in the shape of finished products since they have been treated statistically in some form or the other. After statistical treatment, the primary data loose their original shape and become secondary data. On a closer examination, it will be found that the distinction between primary data and secondary data in many cases is one of degree only. Data which are secondary in the hands of one may be primary for others. Statistics of agricultural production are secondary data for the Agriculture Department of a Government but for the purpose of calculation of national income these data are primary, because they will have to go through further analysis and their shape will not remain the same.

Factor affecting choice of method : It is obvious that the methods of the collection of primary data and secondary data would not be exactly identical because in one case, the data have to be originally collected while in the other, the work is of the nature of compilation. There are various methods of the collection of primary and secondary data and choice of the method depends on a number of factors. *Nature, object and scope* of the enquiry are the most important things on which the selection of the method depends. The method selected should be such that it suites the type of enquiry that is being conducted.

Availability of finance is another factor which influences the selection of the method of collection of data. When financial resources at the disposal of the investigator are scanty, he shall have to leave aside expensive methods even though they are better than others which are comparatively cheap.

Availability of time has also to be taken into account. Some methods involve a long duration of enquiry while with others, the enquiry can be conducted in a comparatively shorter duration. The time at the disposal of the investigator, thus, affects the selection of the technique by which data are to be collected.

3.2 METHODS OF COLLECTING PRIMARY DATA

The following methods of collection of the primary data are in common use :

(*a*) Direct personal investigation

(*b*) Indirect oral investigation

(*c*) By local reports

(*d*) By schedules and questionnaires

We shall briefly discuss each of them in turn.

Direct Personal Investigation

In direct personal investigation, as the name suggests, the investigator has to collect the information personally from the sources concerned. He has to be on the spot for conducting the enquiry and has to meet people from whom data have to be collected. It is necessary that in such cases, the investigator has a keen sense of observation and he is very polite and courteous. He should further acquaint himself with local conditions, customs and traditions so that he is in a position to identify himself fully with the persons from whom the information is sought. The investigator has to be very tactful and cautious in such cases. He should put easy and simple questions which are capable of being answered precisely and in a language which is not vague and fully understandable by the source. In some cases, it may not be possible or worthwhile to contact directly the persons concerned and in such cases, the investigator has to cross-examine other persons who are closely in touch with the sources of data. The information elicited in such a manner should be carefully used and the investigator should make sure that the persons from whom data are being collected actually know the facts fully and can deliver him the goods.

The method of direct personal investigation is suitable only for *intensive investigations. It involves enormous cost* and usually requires a *long time*. It is naturally not suitable for extensive enquiries where the scope of investigation is wide. Further, in this method, the bias or *prejudice of the investigator* can do a lot of damage as he is in sole charge of the collection of data. This method, however, gives very satisfactory results if the scope of the enquiry is narrow and if the investigator is fully dependable and is completely unbiased.

Indirect Oral Investigation

When the above mentioned method cannot be used either on account of the reluctance of persons to part with information when approached directly, or on account of the extensive scope of the enquiry or on account of some other reason, an indirect oral examination can be conducted. In this method, data are not collected directly from the persons concerned but through indirect sources. Persons who are supposed to have knowledge about the problem under investigation are interrogated and the desired information is collected. Usually, in such enquiries, a small *list of questions* relating to the investigation is prepared and these questions are put to different persons (known as witnesses) and their answers are recorded. Most of the commissions and committees appointed by the Government to collect statistical data or to carry on such investigations in which factual data have to be compiled, make use of this method. They request different types of people to come and give evidences and on the basis of these records, facts about different problems are ascertained. In such enquiries the evidence of *one* person should not be relied upon and the views of a number of persons should be ascertained to find out the real position. In this method, the accuracy of data collected would largely depend on the *type of persons* whose evidences are being recorded. It is, therefore, necessary to be very cautious in the selection of these persons. Invariably it should be seen that the person who is being questioned :

(*a*) knows full facts of the problem under investigation ;

(*b*) is not prejudiced;

(*c*) is capable of expressing himself correctly and can give a true account; and

(*d*) is not motivated to give colour to the facts.

Proper allowance made for the inherent *optimism or pessimism of the informants i.e.*, their inherent psychology. For example if there are two drunkards (one optimist and the other pessimist), each of whom was left with half a glass of wine illustrates the point very clearly. The optimist said, "What do I care for the world, I have *yet* half the glass with me" and the pessimist remarked, "What can I do in this world, I have *only* half the glass with me." Both of them were stating facts correctly and yet the two statements give entirely different impressions.

Local Reports

The last method of collection of primary data is through local reports. In this method, data are not formally collected by enumerators but by the local correspondents or agents in their own fashion and to their own likings. Obviously, such data cannot be very reliable and, as such, this method is used in those cases where the purpose of investigation can be served with rough estimates only and where a high degree of precision is not necessary. This method has the advantage of being least expensive and it also saves the botheration usually associated with statistical investigation of other types.

Schedules and Questionnaires

This is an important method for the collection of data which is followed usually by private individuals,

research workers, non-official institutions and, sometimes, the Government also. In this method, a list of questions relating to the problem under investigation is prepared and printed and information is collected from various sources in any of the following ways :

(*a*) *By sending the questionnaire to the persons concerned and requesting them to answer the questions and return the questionnaire.*

The main advantage of this method is that it is least expensive, and with it information can be collected from a wide area in a comparatively short period of time. If the investigation is properly conducted, the method can easily ensure a reasonable standard of accuracy. Success in this method depends on the co-operation and the attitude of the informants. If the informants are indifferent in their attitude and do not take interest in answering the questions, the method may give any fruit for results.

However, this method cannot be used if the informants are illiterate. If they are literate but adopt an indifferent attitude, then also the method should be used with utmost caution as, in such cases, likelihood of error is very great.

(*b*) *By sending the questionnaires through enumerators to help the informants in filling the answers.*

In this method, the enumerators go to the informants along with the questionnaires and help them in recording their answers. The enumerators explain the aims and objects of the investigation to the informants and also emphasise the necessity and usefulness of correct answers. They also remove the difficulties which any informant may feel in understanding the implications of a particular question or the definition or concept of difficult terms. This method is very useful in extensive enquiries and with it, fairly dependable results can be expected. It is, however, very expensive and usually such enquiries can be conducted only by the Government. Population census all over the world is conducted by this method. In such enquiries, it is necessary that not only the questions are simple and few in number but the enumerators are also courteous and polite and have proper training.

The selection of enumerators is very important and should be carefully done. The enumerators should be explained the nature, scope and subject of the investigation thoroughly and they should properly understand the implications of the different questions put and the definitions of the various terms used. The enumerators should have intelligence and capacity of cross-examination for the purpose of finding out the truth and they should be persons who are hard-working and should have patience and perseverance.

3.3 DRAFTING A QUESTIONNAIRE

The success of the questionnaire method depends to a large extent on proper drafting of the questionnaire. It is a highly specialised job and needs lot of skill and experience.

Though there is no hard and fast rule for preparing a questionnaire, yet some broad principles followed in this regard are given below :

1. (*i*) A very polite covering letter should be sent to the respondents with the questionnaire, emphasising the need and usefulness of the information that is being collected and requesting them to give their co-operation by sending correct replies. A self-addressed stamped envelope should also be enclosed for the respondent's convenience in sending the reply.

(*ii*) An assurance must be given to the respondents that if they wish the information given by them would be treated as confidential and not disclosed to anybody, it will be taken care of.

(*iii*) If possible, some inducement could also be given to the respondent in the form of small gifts so that there is greater chance of getting a response.

(*iv*) The respondents could also be given a promise that if they so desire a copy of the results of the survey would be sent to them,

2. Number of questions should not be very large : Unnecessary details should be avoided and only relevant questions should be framed. There is no hard and fast rule with regard to the number of questions as it would depend on the nature of the enquiry.

3. The questions should be short and clear : There should be no ambiguity in the questions. The questions should not be confusing and should be capable of a straight answer. *As far as possible, the questions should be capable of objective answers.* A set of possible answers may be given against each question and the respondent may be asked to tick the appropriate answer. For example if we have to ask a respondent about the percentage of his income spent by him on his children's education, we can frame the question as follows :

What percentage of your total income is spent on your children's education ? Tick the appropriate square

(*a*) Less than 5% ☐
(*b*) 5% to 10% ☐
(*c*) 10% to 15% ☐
(*d*) 15% to 20% ☐
(*e*) more than 20% ☐

Note : If you spend exactly 10% of your income, please tick 10% to 15% and, similarly, if it is exactly 15%, then tick 15 to 20 per cent.

Another example could be when you wish to find out the mode of cooking in a household. The question should be :

Which of the following modes of cooking do you use ?

(*a*) Wood ☐
(*b*) Coal ☐
(*c*) Kerosine ☐
(*d*) Gas ☐
(*e*) Electricity ☐

Please tick the relevant square. If you use more than one method, please tick the one which is mostly used.

4. If the opinion of a respondent is sought on a particular point, if possible, the question should be so framed that *he can answer it in yes or no*. For example, there could be a question like :

Do you own a house ? The answer could be 'yes' or 'no'. Such questions are not desirable when the answer cannot be a clear-cut one. For example, if a respondent's opinion is sought about some government policies, the answer may not be clear-cut yes or no. Such questions should, as far as possible, be avoided.

5. Questions of a personal nature should not be asked because the respondent would either not give information or give incorrect information. Questions relating to income tax returns or perks given to certain employees may not be liked by the respondents. In fact, no question which elicits some sort of confidential information should be included in the questionnaire.

6. No such question which hurts the sentiments of the respondent should be asked : For example questions on indebtedness, private life, litigation, etc., should be avoided.

7. There should be such questions which are corroboratory in nature : Such questions are meant for cross-checks to the answers given by the respondent. If the answers given are correct, there would be no contradiction in answers.

8. Questions whose answers require calculation should not be generally asked : For example, if the annual income of a salaried class person is asked or if monthly profits of a concern are called for, or if the percentage of part-time employees to total employees is asked, a lot of calculation work is involved. The respondent may not wish to undertake this arduous task.

9. The questionnaire should look attractive and impressive : It always helps if it is so. There should be sufficient space for answering the questions and the quality of paper and printing should also be good.

3.4 PRE-TESTING A QUESTIONNAIRE

Before the questionnaire is finalised, it is always worthwhile to get it pre-tested. A small sample from the relevant universe may be picked up and the questionnaire tested on it.

This has many advantages. We are in a position to know the type of response that we may ultimately get from the respondents. Some questions may be found to be inappropriate and may need a change. Some concepts may need clarification as the sample respondents might have displayed confusion while answering questions.

By pre-testing a questionnaire, it is also possible to find out the ways in which greater co-operation of the respondents may be achieved.

After the questionnaire has been pretested, it should be modified in the light of the experience gathered and then used for the purpose of collecting information.

4. Rules of Tabulation : There are no hard and fast rules for the tabulation of data. ***Bowley*** very rightly pointed out that *"In collection and tabulation commonsense is the chief requisite and experience, the chief teacher."* Constructing a good table is an art and practical experience is of great help in designing the structure of a table. However, the following general rules should be observed while tabulating statistics' data :

(*a*) Table should be precise and easy to understand. It should not be necessary to go through footnotes or explanation to properly understand a table.

(*b*) If the data are very large they should not be crowded in a single table. This would increase the chances of mistakes and would make the table unwieldy and inconvenient. Such data can be presented in a number of tables. Each table should be complete in itself and should serve a particular purpose.

(*c*) The table should suit the size of the paper and, therefore, the width of the column should be decided beforehand.

(*d*) The number of main headings should be few, though, there is no harm if the number of sub-headings is large. This will help in understanding the main points of the table.

(*e*) Captions, headings or sub-headings of columns and headings and sub-headings of rows must be self explanatory.

(*f*) Those columns whose data are to be compared should be kept side by side. Similarly, percentages, totals and averages must be kept close to the data.

(*g*) As far as possible figures should be approximated before tabulation. This would reduce unnecessary details.

(*h*) The units of measurement under each heading or sub-heading must always be indicated.

(*i*) Total of rows should be placed in the extreme right column; though sometimes they are placed in the first column after the vertical captions on the left. The totals of columns should ordinarily be placed at the foot though in some cases it is helpful to place them at the top of the table.

(*j*) Items should be arranged either in alphabetical, chronological or geographical order or according to size, importance, emphasis or casual relationship to facilitate comparison.

(*k*) If certain figures are to be emphasised they should be in distinctive type or in a "box" or "circle" or between thick lines.

(*l*) When percentages are given side by side with original figures they should be in a separate type—preferably italics.

(*m*) If some portion of collected data cannot be classified in any class or division a miscellaneous class should be created and the data shown in it.

(*n*) There should be a proper title to each table. It should tell what exactly the table presents.

(*o*) Indicate a zero quantity by a zero and do not use zero to indicate such information which is not available. Information which is not available should be indicated by the letters N. A. or by dash (—).

(*p*) Abbreviations should be avoided, particularly in titles and sub-titles.

(*q*) The data should be tabulated in an explicit fashion. The expression, "etc.", should not be used in a table, since the reader may not easily find out what it refers to.

(*r*) Ditto marks should not be used in a table. Sometimes, it creates confusion.

It may be difficult to follow all these guidelines in preparing a single table, but their purpose should always be kept in mind.

J.C. Capt has summarised general rules of tabulation in the following words :

"In the final analysis, there are only two rules in tabular presentation that should be applied rigidly. First, the use of common sense when planning a table, and second, the viewing of the proposed table from the standpoint of the user. The details of mechanical arrangement must be governed by a single objective, that is, to make the statistical table as easy to read and to understand as the nature of the material will permit."

5. Types of Tabulation (*i*) Simple and complex tabulation : Broadly speaking, tabulation of data can be either simple or complex. Simple tabulation gives information about one or more groups of independent questions. Complex tabulation shows the division of data in two or more categories and, as such, is meant to give information about one or more sets of interrelated questions.

One-way tables : Simple tabulation usually gives rise to single or one-way tables. One way tables supply answers to questions about one characteristic of data only. The following table will illustrate the point:

The above table can supply us information about (1) marks obtained by students, (2) the distribution of these students sex-wise, and (3) the distribution of the students on the basis of residence.

Higher order tables : The tables can also be the *manifold or of higher order*. Such tables supply information about a large number of interrelated questions. If in the above table, additional information is given about civil conditions (married or unmarried) of the students it would become a four-way table and, similarly, tables can be of still higher order-five-way, six-way, and so on. All such tables are called manifold or higher order tables. Table 4 is an example of a manifold table :

Table 4 : Marks obtained by students (sex-wise, on the basis of civil conditions, and residences)

		Number of students								
Residence	Marks	Male			Female			Total		
		Married	Unmarried	Total	Married	Unmarried	Total	Married	Unmarried	Total
Hostellers	30–40									
	40–50									
	50–60									
	60–70									
	70–80									
	Total									
Day scholar	30–40									
	40–50									
	50–60									
	60–70									
	70–80									
	Total									
Total	30–40									
	40–50									
	50–60									
	60–70									
	70–80									
Grand Total										

The above table gives information about a large number of interrelated questions regarding students, namely, about the marks obtained, sex-wise distribution, civil conditions and residence. Manifold tables are very useful in presenting population census data.

2. General and Special Purpose Tables : General purpose tables are also called *reference tables or repository tables*, and they provide information for general use and reference. **Croxton** and **Crowden** have identified the purpose of such tables in the following words :

"Primarily and usually the sole purpose of a reference table is to present data in such a manner that individual items may be found readily by a reader."

According to **Horse Secrist** :

"Reference tables contain ungrouped data basic for a particular report, usually containing a large amount of data and frequently related to a tabular appendix."

Tables published by various governmental agencies like C. S. O. in our country are particularly of this type. The purpose of these tables is to present detailed statistical information relating to a general subject under study.

Specific purpose tables also known as ***text tables*** and ***summary tables*** present information relating to a specific subject under study. For example, production figures of a particular industrial unit for a number of years or figures relating to its profits would give us specific purpose tables. Such tables are, sometimes, in the nature of *derivative tables* where information contained in them is derived from the general tables. For example, if from the general table, certain ratios, percentages and other measures are derived, the table containing them would be a derivative table.

Given below are some illustrations explaining the preparation of some types of tables discussed above.

Example 6. Present the following information in a suitable tabular form :

In 1995 out of a total of 1,750 workers of a factory, 1,200 were members of a trade union.

The number of women employed was 200 of which 175 did not belong to a trade union. In 2000 the number of union workers increased to 1,580, of which 1,290 were men. On the other hand, the number of non-union workers fell down to 208, of which 180 were men.

In 2005, there were 1,800 employees who belonged to a trade union and 50 who did not belong to a trade union. Of all the employees in 2005, 300 were women of whom only 8 did not belong to a trade union.

Solution. Table Showing The Sex-wise Distribution of Union and Non-union Members for 1965, 1970 and 1975

Category	1995			2000			2005		
	M	F	Total	M	F	Total	M	F	Total
Members	1,175	25	1200	1290	290	1580	1508	292	1800
Non-members	375	175	550	180	28	208	42	8	50
Total	1,550	200	1750	1470	318	1788	1550	300	1850

Example 7. Present the following information in a suitable form supplying the figures not directly given. In 2005, out of a total of 4,000 workers in a factory, 3,300 were members of a trade union. The number of women workers employed was 500 out of which 400 did not belong to any union.

In 2004, the number of workers in the union was 3,450 of which 3,200 were men. The number of non-union workers was 760 of which 330 were women.

Solution. Table Showing Union and non-union Workers By Sex

Union	2004			2005		
	Males	Females	Total	Males	Females	Total
Members	3,200	250	3,450	3,200	100	3,300
Non-members	430	330	760	300	400	700
Total	3,630	580	4,210	3,500	500	4,000

Example 8. In a trip organised by a college there were 80 persons, each of whom paid Rs.15.50 on an average. There were 60 Students each of whom paid Rs.16. Members of the teaching staff were charged at a higher rate. The number of servants was 6 (all males) and they were not charged anything. The number of ladies was 20% of the total of which one was a lady staff member. Tabulate the above information.

Solution. Table Showing the Type of Participants, Sex and Contribution Made

Types of Participants	Sex			Contribution per member (Rs.)	Total contribution (Rs.)
	Male	Female	Total		
Students	45	15	60	16.00	960
Teaching Staff	13	1	14	20.00	280
Servants	6	—	6	—	—
Total	64	16	80	—	1,240

Notes. 1. Total contribution = average contribution × No. of Persons who joined the trip
$= 15.5 \times 80 = 1{,}240$

2. Contribution of the staff per head has been obtained by deducting the contribution of students from the total and dividing the difference by the number of teaching staff, *i.e.*,

$$\frac{1240 - (60 \times 16)}{14} = \frac{1240 - 960}{14} = \frac{280}{14} = \text{Rs. } 20$$

Example 9. Of the 1,125 students studying in a college during a year, 720 were Hindus, 628 were boys,