Chapter 1

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

—H. G. Wells

Statistics: An Overview

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- present a broad overview of statistics as a subject.
- bring out applications of statistics and its usefulness in managerial decision-making.
- describe the data collection process.
- understand basic concepts of questionnaire design and measurement scales.

1.1 REASONS FOR LEARNING STATISTICS

H. G. Wells' statement that the 'statistical thinking will one day be as necessary as the ability to read and write' is valid in the context of today's competitive business environment where many organizations find themselves data-rich but information-poor. Thus for decision-makers it is important to develop the ability to extract meaningful information from raw data to make better decisions. It is possible only through the careful analysis of data guided by statistical thinking.

Reasons for analysis of **data** is an understanding of *variation and its causes* in any phenomenon. Since variation is present in all phenomena, therefore knowledge of it leads to better decisions about a phenomenon that produced the data. It is from this perspective that the learning of statistics enables the decision-maker to understand how to

Data: A collection of observations of one or more variables of interest.

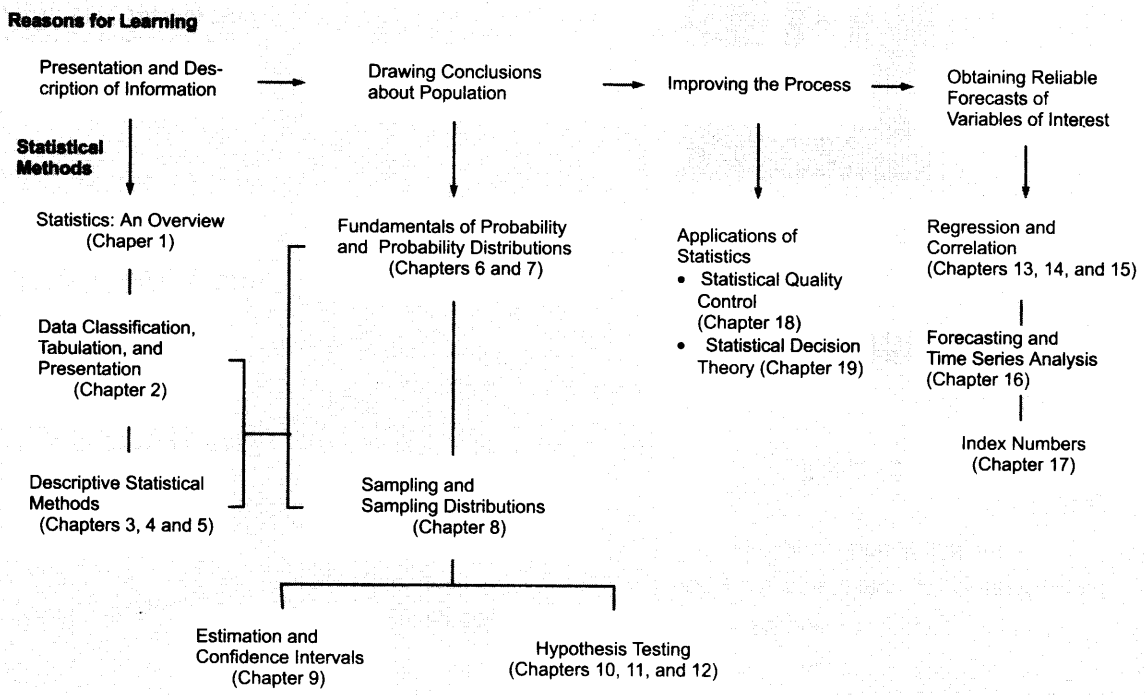
- present and describe information (data) so as to improve decisions.
- draw conclusions about the large **population** based upon information obtained from samples.
- seek out relationship between pair of variables to improve processes.
- obtain reliable forecasts of statistical variables of interest.

Population: A collection of all elements (units or variables) of interest.

Thus a statistical study might be simple exploration enabling us to gain an insight into a virtually unknown situation or it might be sophisticated analysis to produce numerical confirmation or refection of some widely held belief.

As shown in Fig. 1.1, the text matter of the book has been organized keeping in view these four reasons for learning statistics.

Figure 1.1
Flow Chart of Reasons For Learning Statistics



1.2 GROWTH AND DEVELOPMENT OF STATISTICS

Statistics: The art and science of collecting, analysing, presenting, and interpreting data.

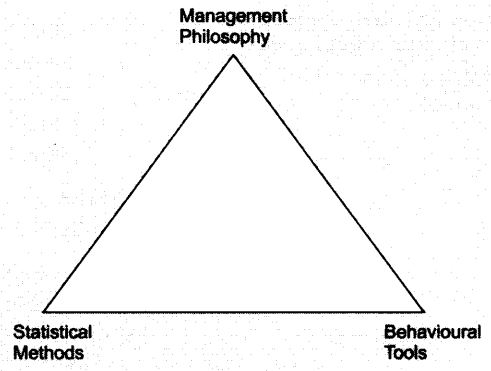
The views commonly held about **statistics** are numerous, but often incomplete. It has different meanings to different people depending largely on its use. For example, (i) for a cricket fan, statistics refers to numerical information or data relating to the runs scored by a cricketer; (ii) for an environmentalist, statistics refers to information on the quantity of pollution released into the atmosphere by all types of vehicles in different cities; (iii) for the census department, statistics consists of information about the birth rate per thousand and the sex ratio in different states; (iv) for a share broker, statistics is the information on changes in share prices over a period of time; and so on.

The average person perceives statistics as a column of figures, various types of graphs, tables and charts showing the increase and/or decrease in per capita income, wholesale price index, industrial production, exports, imports, crime rate and so on. The sources of such statistics for a common man are newspapers, magazines/journals, reports/bulletins, radio, and television. In all such cases the relevant data are collected, numbers manipulated and information presented with the help of figures, charts, diagrams, and pictograms; probabilities are quoted, conclusions reached, and discussions held. Efforts to understand and find a solution (with certain degree of precision) to problems pertaining to social, political, economic, and cultural activities, seem to be unending. All such efforts are guided by the use of methods drawn from the field of statistics.

The development of mathematics in relation to the probability theory and the advent of fast-speed computers have substantially changed the field of statistics in the last few decades. The use of computer software, such as SAS and SPSS, have brought about a technological revolution. The increasing use of spreadsheet packages like Lotus 1-2-3 and Microsoft Excel have led to the incorporation of statistical features in these packages. These softwares have made the task of statistical analysis quite convenient and feasible.

1.3 STATISTICAL THINKING AND ANALYSIS

Figure 1.2
Quality Improvement Process Model

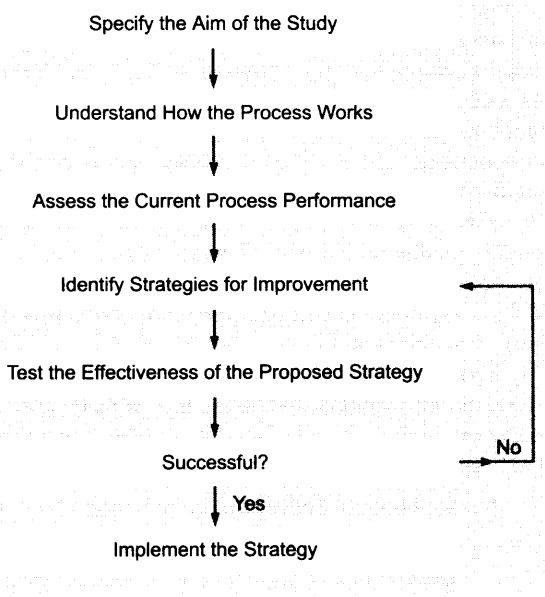


An integral part of the managerial approach focuses on the quality of products manufactured or services provided by an organization. This approach requires the application of certain statistical methods and the statistical thinking by the management of the organization. *Statistical thinking can be defined as the thought process that focuses on ways to identify, control, and reduce variations present in all phenomena.* A better understanding of a phenomenon through statistical thinking and use of statistical methods for data analysis, enhances opportunities for improvement in the quality of products or services. Statistical thinking also allows one to recognize and make interpretations of the variations in a process.

As shown in Fig. 1.2, *management philosophy* acts as a guide for laying a solid foundation for total quality improvement efforts. However, use of *behavioural tools* such as brainstorming, team-building, and nominal group decision-making, and *statistical methods* such as tables, control charts, and descriptive statistics, are also necessary for understanding and improving the processes.

The steps of statistical thinking necessary for increased understanding of and improvement in the processes are summarized in Fig. 1.3.

Figure 1.3
Flow Chart of Process Improvement



1.4 STATISTICS DEFINED

As Statistical Data The word statistics refers to a special discipline or a collection of procedures and principles useful as an aid in gathering and analysis of numerical information for the purpose of drawing conclusions and making decisions. Since any numerical figure, or figures, cannot be called statistics owing to many considerations which decide its use, statistical data or mere data is more appropriate expression to indicate numerical facts.

A few definitions which describe the statistics characteristics are as follows:

- The classified facts respecting the condition of the people in a state . . . especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement. —Webster

This definition is quite narrow as it confines the scope of statistics only to such facts and figures which are related to the conditions of the people in a state.

Quantitative data: Numerical data measured on the interval or ratio scales to describe 'how much' or 'how many'.

- By statistics we mean **quantitative data** affected to a marked extent by multiplicity of causes. —Yule and Kendall
- 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 predetermined purpose and placed in relation to each other. —Horace Secrist

The definition given by Horace is more comprehensive than those of Yule and Kendall. This definition highlights the following important characteristics

(i) statistics are aggregates of facts,
(ii) statistics are effected to a marked extent by multiplicity of causes,
(iii) statistics are numerically expressed,
(iv) statistics are enumerated or estimated according to reasonable standards of accuracy,
(v) statistics are collected in a systematic manner for a pre-determined purpose, and
(vi) statistics are placed in relation to each other.

As Statistical Methods Methods adopted as aids in the collection and analyses of numerical information or statistical data for the purpose of drawing conclusions and making decisions are called *statistical methods*.

Statistical methods, also called statistical techniques, are sometimes loosely referred to cover 'statistics' as a subject in whole. There are two branches of statistics: (i) *Mathematical statistics* and (ii) *Applied statistics*. Mathematical statistics is a branch of mathematics and is theoretical. It deals with the basic theory about how a particular statistical method is developed. Applied statistics, on the other hand, uses statistical theory in formulating and solving problems in other subject areas such as economics, sociology, medicine, business/industry, education, and psychology.

The field of applied statistics is not easy because the rules necessary to solve a particular problem are not always obvious although the guiding principles that underlie the various methods are identical regardless of the field of their application. As a matter of fact, experience and judgment are otherwise more necessary to execute a given statistical investigation.

The purpose of this book is limited to discussing the fundamental principles and methods of applied statistics in a simple and lucid manner so that readers with no previous formal knowledge of mathematics could acquire the ability to use statistical methods for making managerial decisions.

A few relevant definitions of statistical methods are given below:

- 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. —Seligman
- The science of statistics is the method of judging, collecting natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates. —King

A. L. Bowley has given the following three definitions keeping in mind various aspects of statistics as a science:

- Statistics may be called the science of counting.
- Statistics may be called the science of average.
- Statistics is the science of the measurement of social organism regarded as a whole in all its manifestations.

These definitions confine the scope of statistical analysis only to 'counting, average, and applications' in the field of sociology alone. Bowley realized this limitation and himself said that statistics cannot be confined to any science. Another definition of statistics given by Croxton and Cowden is as follows:

- Statistics may be defined as a science of collection, presentation, analysis and interpretation of numerical data. —Croxton and Cowden

This definition has pointed out four stages of statistical investigation, to which one more stage 'organization of data' rightly deserves to be added. Accordingly, statistics may be defined as the science of collecting, organizing, presenting, analysing, and interpreting numerical data for making better decisions.

1.5 TYPES OF STATISTICAL METHODS

Statistical methods, broadly, fall into the following two categories:

(i) Descriptive statistics, and
(ii) Inferential statistics

Descriptive statistics: It consists of procedures used to summarize and describe the characteristics of a set of data.

Descriptive statistics includes statistical methods involving the collection, presentation, and characterization of a set of data in order to describe the various features of that set of data.

In general, methods of descriptive statistics include graphic methods and numeric measures. Bar charts, line graphs, and pie charts comprise the graphic methods, whereas numeric measures include measures of central tendency, dispersion, skewness, and kurtosis.

Inferential statistics: It consists of procedures used to make inferences about population characteristics on the basis of sample results.

Inferential statistics includes statistical methods which facilitate estimating the characteristic of a population or making decisions concerning a population on the basis of sample results. **Sample** and population are two relative terms. The larger group of units about which inferences are to be made is called the population or universe and a sample is a fraction, subset, or portion of that universe.

Inferential statistics can be categorized as *parametric* or *non-parametric*. The use of parametric statistics is based on the assumption that the population from which the sample is drawn, is normally distributed. Parametric statistics can be used only when data are collected on an interval or ratio scale. Non-parametric statistics makes no explicit assumption regarding the normality of distribution in the population and is used when the data are collected on a nominal or ordinal scale.

Sample: A subset (portion) of the population.

The need for sampling arises because in many situations data are sought for a large group of elements such as individuals, companies, voters, households, products, customers, and so on to make inferences about the population that the sample represents. Thus, due to time, cost, and other considerations data are collected from only a small portion of the population called *sample*. The concepts derived from probability theory help to ascertain the likelihood that the analysis of the characteristic based on a sample do reflect the characteristic of the population from which the sample is drawn. This helps the decision-maker to draw conclusions about the characteristics of a large population under study.

Following definitions are necessary to understand the concept of inferential statistics:

- A *process* is a set of conditions that repeatedly come together to transform inputs into outcomes. Examples includes a business process to serve customers, length of time to complete a banking transaction, manufacturing of goods, and so on.
- A *population* (or *universe*) is a group of elements or observations relating to a phenomenon under study for which greater knowledge and understanding is needed. The observations in population may relate to employees in a company, a large group of manufactured items, vital events like births and deaths or road accidents. A population can be *finite* or *infinite* according to the number of observations under statistical investigation.
- A *statistical variable* is an operationally defined characteristic of a population or process and represents the quantity to be observed or measured.
- A *sample* is a group of some, but not all, of the elements or observations of a population or process. The individual elements of a sample are called *sampling* or *experimental units*.
- A *parameter* is a descriptive or summary measure (a numerical quantity) associated with a statistical variable that describes a characteristic of the entire population.

- A *statistic* is a numerical quantity that describes the characteristic of a sample drawn from a population.

For example, a manufacturer who produces electrical coils wanted to learn the average resistance of coils. For this he selects a sample of coils at regular intervals of time and measures the resistance of each. If the sample average does not fall within the specified range of variations, the process controls are checked and adjustments are made. In this example, the population or universe would be all the coils being produced by the manufacturing process; the statistical variable is the resistance of a coil; statistic is the average resistance of coils in a given sample; parameters of interest are the average resistance and variation in resistance among manufactured coils; and sampling units are the coils selected for the sample.

1.6 IMPORTANCE AND SCOPE OF STATISTICS

The scope of applications of statistics has assumed unprecedented dimensions these days. Statistical methods are applicable in all diversified fields such as economics, trade, industry, commerce, agriculture, bio-sciences, physical sciences, educations, astronomy, insurance, accountancy and auditing, sociology, psychology, meteorology, and so on. Bringing out its wide applications, Carrol D. Wright (1887), United States Commissioner of the Bureau of Labour, has explained the importance of statistics in saying so:

> To a very striking degree our culture has become a statistical culture. Even a person who may never have heard of an index number is affected by those index numbers which describe the cost of living. It is impossible to understand Psychology, Sociology, Economics or a Physical Science without some general idea of the meaning of an average, of variation, of concomitance of sampling, of how to interpret charts and tables.

In the recent past, statistics has acquired its importance as a subject of study in the curricula of many other disciplines. According to the statistician, Bowley, '*A knowledge of statistics is like a knowledge of foreign language or of algebra, it may prove of use at any time under any circumstances*'.

Given below is a brief discussion on the importance of statistics in a few other important disciplines.

1.6.1 Statistics and the State

A state in the modern setup collects the largest amount of statistics for various purposes. It collects data relating to prices, production, consumption, income and expenditure, investments, and profits. Popular statistical methods such as time-series analysis, index numbers, forecasting, and demand analysis are extensively practised in formulating economic policies. Governments also collect data on population dynamics in order to initiate and implement various welfare policies and programmes.

In addition to statistical bureaus in all ministries and government departments in the Central and state governments, other important agencies in the field are the Central Statistical Organisation (CSO), National Sample Survey Organization (NSSO), and the Registrar General of India (RGI).

1.6.2 Statistics in Economics

Statistical methods are extensively used in all branches of economics. For example:

(i) Time-series analysis is used for studying the behaviour of prices, production and consumption of commodities, money in circulation, and bank deposits and clearings.
(ii) Index numbers are useful in economic planning as they indicate the changes over a specified period of time in (a) prices of commodities, (b) imports and exports, (c) industrial/agricultural production, (d) cost of living, and the like.
(iii) Demand analysis is used to study the relationship between the price of a commodity and its output (supply).

(iv) Forecasting techniques are used for curve fitting by the principle of least squares and exponential smoothing to predict inflation rate, unemployment rate, or manufacturing capacity utilization.

1.6.3 Statistics in Business Management

According to Wallis and Roberts, 'Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty.' Ya-Lin-Chou gave a modified definition over this, saying that 'Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks.' These definitions reflect the applications of statistics in the development of general principles for dealing with uncertainty.

Statistical reports provide a summary of business activities which improves capability of making more effective decisions regarding future activities. Discussed below are certain activities of a typical organization where statistics plays an important role in their efficient execution.

Marketing Before a product is launched, the market research team of an organization, through a pilot survey, makes use of various techniques of statistics to analyse data on population, purchasing power, habits of the consumers, competitors, pricing, and a hoard of other aspects. Such studies reveal the possible market potential for the product.

Analysis of sales volume in relation to the purchasing power and concentration of population is helpful in establishing sales territories, routing of salesman, and advertising strategies to improve sales.

Production Statistical methods are used to carry out R&D programmes for improvement in the quality of the existing products and setting quality control standards for new ones. Decisions about the quantity and time of either self-manufacturing or buying from outside are based on statistically analysed data.

Finance A statistical study through correlation analysis of profit and dividend helps to predict and decide probable dividends for future years. Statistics applied to analysis of data on assets and liabilities and income and expenditure, help to ascertain the financial results of various operations.

Financial forecasts, break-even analysis, investment decisions under uncertainty—all involve the application of relevant statistical methods for analysis.

Personnel In the process of manpower planning, a personnel department makes statistical studies of wage rates, incentive plans, cost of living, labour turnover rates, employment trends, accident rates, performance appraisal, and training and development programmes. Employer-employee relationships are studied by statistically analysing various factors—wages, grievances handling, welfare, delegation of authority, education and housing facilities, and training and development.

1.6.4 Statistics in Physical Sciences

Currently there is an increasing use of statistical methods in physical sciences such as astronomy, engineering, geology, meteorology, and certain branches of physics. Statistical methods such as sampling, estimation, and design of experiments are very effective in the analysis of quantitative expressions in all fields of most physical sciences.

1.6.5 Statistics in Social Sciences

The following definitions reflect the importance of statistics in social sciences.

- Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations. —Bowley
- The science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis, enumeration or collection of estimates. —W. I. King

Some specific areas of applications of statistics in social sciences are as listed below:

(i) Regression and correlation analysis techniques are used to study and isolate all those factors associated with each social phenomenon which bring out the changes in data with respect to time, place, and object.

(ii) Sampling techniques and estimation theory are indispensable methods for conducting any social survey pertaining to any strata of society, and drawing valid inferences.

(iii) In sociology, statistical methods are used to study mortality (death) rates, fertility (birth rates) trends, population growth, and other aspects of vital statistics.

1.6.6 Statistics in Medical Sciences

The knowledge of statistical techniques in all natural sciences—zoology, botany, meteorology, and medicine—is of great importance. For example, for proper diagnosis of a disease, the doctor needs and relies heavily on factual data relating to pulse rate, body temperature, blood pressure, heart beats, and body weight.

An important application of statistics lies in using the *test of significance* for testing the efficacy of a particular drug or injection meant to cure a specific disease. Comparative studies for effectiveness of a particular drug/injection manufactured by different companies can also be made by using statistical techniques such as the t-test and F-test.

To study plant life, a botanist has to rely on data about the effect of temperature, type of environment, and rainfall, and so on.

1.6.7 Statistics and Computers

Computers and information technology, in general, have had a fundamental effect on most business and service organizations. Over the last decade or so, however, the advent of the personal computer (PC) has revolutionized both the areas to which statistical techniques are applied. PC facilities such as spreadsheets or common statistical packages have now made such analysis readily available to any business decision-maker. Computers help in processing and maintaining past records of operations involving payroll calculations, inventory management, railway/airline reservations, and the like. Use of computer softwares, however, presupposes that the user is able to interpret the computer outputs that are generated.

Remark We discussed above the usefulness of statistical techniques in some important fields. However, the scope of statistics is not limited to these only. Statistical data and methods are useful to banking, research and development, insurance, astronomy, accountancy and auditing, social workers, labour unions, chambers of commerce, and so on.

1.7 LIMITATIONS OF STATISTICS

Although statistics has its applications in almost all sciences—social, physical, and natural—it has its own limitations as well, which restrict its scope and utility.

1.7.1 Statistics Does Not Study Qualitative Phenomena

Since statistics deals with numerical data, it cannot be applied in studying those problems which can be stated and expressed quantitatively. For example, a statement like 'Export volume of India has increased considerably during the last few years' cannot be analysed statistically. Also, qualitative characteristics such as honesty, poverty, welfare, beauty, or health, cannot directly be measured quantitatively. However, these subjective concepts can be related in an indirect manner to numerical data after assigning particular scores or quantitative standards. For example, attributes of intelligence in a class of students can be studied on the basis of their Intelligence Quotients (IQ) which is considered as a quantitative measure of the intelligence.

1.7.2 Statistics Does Not Study Individuals

According to Horace Secrist '*By statistics we mean aggregate of facts effected to a marked extent by multiplicity of factors . . . and placed in relation to each other.*' This statement implies that a single or isolated figure cannot be considered as statistics, unless it is part of the aggregate of facts relating to any particular field of enquiry. For example, price of a single commodity or increase or decrease in the share price of a particular company does not constitute statistics. However, the aggregate of figures representing prices, production, sales volume, and profits over a period of time or for different places do constitute statistics.

1.7.3 Statistics Can be Misused

Statistics are liable to be misused. For proper use of statistics one should have enough skill and experience to draw accurate and sensible conclusions. Further, valid results cannot be drawn from the use of statistics unless one has a proper understanding of the subject to which it is applied.

The greatest danger of statistics lies in its use by those who do not possess sufficient experience and ability to analyse and interpret statistical data and draw sensible conclusions. Bowley was right when he said that '*statistics only furnishes a tool though imperfect which is dangerous in the hands of those who do not know its use and deficiencies.*' For example, the conclusion that smoking causes lung cancer, since 90 per cent of people who smoke die before the age of 70 years, is statistically invalid because here nothing has been mentioned about the percentage of people who do not smoke and die before reaching the age of 70 years. According to W. I. King, '*statistics are like clay of which you can make a God or a Devil as you please.*' He also remarked, '*science of statistics is the useful servant but only of great value to those who understand its proper use.*'

1.8 HOW TO LIE WITH STATISTICS

Despite the happy use of statistics and statistical methods in almost every profession, it is still distrusted. Statistics is considered one of the three types of lies: lies, damn lies, and statistics. Listed below may be two reasons for such a notion being held by people about statistics.

(i) Figures being innocent and convincing, are easily believable.
(ii) Figures which support a particular statement may not be true. Such figures may be incomplete, inaccurate, or deliberately manipulated by prejudiced persons in an attempt to deceive the user or attain ones own motive.

Table 1.1 lists some of the personal qualities and attributes considered necessary for an individual to be an effective statistician:

Table 1-1 Personal Qualities and Attributes For A Statistician*

An effective statistician	
• is well-trained in the theory and practice of statistics.	• has a pleasing personality and is able to work with others.
• is an effective problem-solver.	• gets highly involved in solving organizational problems.
• has good oral and written communication skills.	• is able to extend and develop statistical methodology.
• can work within the constraints of real-life.	• can adapt quickly to new problems and challenges.
• knows how to use computers to solve problems.	• produces high quality work in an orderly and timely fashion.
• understands the realities of statistical practices.	

* *Source:* "Preparing Statistics for Cancers in Industry," *The American Statistician*, Vol. 34, No. 2, May 1980.

Conceptual Questions 1A

1. What is statistics? How do you think that the knowledge of statistics is essential in management decisions. Give examples.
2. Write a brief note on the application of statistics in business and industry.
3. Discuss the meaning and scope of statistics, bringing out its importance particularly in the field of trade and commerce.
4. (a) How far can statistics be applied for business decisions? Discuss briefly bringing out limitations, if any
   (b) Define 'statistics' and give its main limitations.
5. (a) Explain how statistics plays an important role in management planning and decision-making?
   (b) Define statistics and statistical methods. Explain the uses of statistical methods in modern business.
   [Vikram Univ., MBA, 1996]
6. Statistical methods are the most dangerous tools in the hands of an inexpert. Examine this statement. How are statistics helpful in business and industry? Explain.
   [Delhi Univ., MBA, 1999]
7. (a) Define statistics. Discuss its applications in the management of business enterprises. What are its limitation, if any.
   [Jodhpur Univ., MBA; HP Univ., MBA, 1996]
   (b) Explain the utility of statistics as a managerial tool. Also discuss its limitations.
   [Osmania Univ., MBA, 1998]
8. What role does Business Statistics play in the management of a business enterprise? Examine its scope and limitations. *[Delhi Univ., MBA, 1998]*
9. (a) Statistics are like clay of which you can make a God or Devil, as you please. Explain.
   (b) There are three known lies : lies, dam-lies and statistics. Comment on this statement and point out the limitations of statistics.
10. Discuss briefly the applications of Business Statistics, pointing out their limitations, if any. *[Delhi Univ., MBA, 1997]*
11. Describe the main areas of business and industry where statistics are extensively used.
12. Statistics affects everybody and touches life at many points. It is both a science and an art. Explain this statement with suitable examples.
13. With the help of few examples explain the role of statistics as a managerial tool.
14. 'Statistics in the science of estimates and probabilities'. Explain the statement and discuss the role of statistics in the management of business enterprises.
15. Are statistical methods likely to be of any use to a business firm ? Illustrate your answer with some typical business problems and the statistical techniques to be used there.
    [HP Univ., MBA, 1996; Delhi Univ., MBA, 2000; Roorkee Univ., MBA, 2000]
16. 'Statistics is a body of methods for making wise decisions in the face of uncertainty'. Comment on the statement bringing out clearly how statistics helps in business decision-making. *[Osmania Univ., MBA, 1996]*
17. 'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write' Comment. Also give two examples, of how the science of statistics could be of use in managerial decision-making.
    [HP Univ., MBA, 1996]
18. 'Statistics is a method of decision-making in the face of uncertainty on the basis of numerical data and calculated risks'. Comment and explain with suitable illustrations.
    [Delhi Univ., MBA, 1992, 1993]
19. 'Without adequate understanding of statistics, the investigator in social sciences may frequently be like the blind man grouping in a dark closet for a black cat that is not there'. Comment. Give two examples of the use and abuse of statistics in business.
20. One can say that statistical inference includes an interest in statistical description as well, since the ultimate purpose of statistical inference is to describe population data. Does statistical inference differ from statistical description? Discuss.
21. What characteristics are inevitable in virtually all data and why is the understanding of it important?
22. 'Modern statistical tools and techniques are important for improving the quality of managerial decisions'. Explain this statement and discuss the role of statistics in the planning and control of business. *[HP Univ., MBA, 1998]*
23. 'The fundamental gospel of statistics is to push back the domain of ignorance, rule of thumb, arbitrary or prepare decisions, traditions, and dogmatism, and to increase the domain in which decisions are made and principles are formulated on the basis of analysed quantitative facts'. Explain the statement with the help of a few business examples. *[Osmania Univ., MBA, 1999]*
24. 'Statistics are numerical statements of facts but all facts numerically stated are not statistics'. Comment upon the statement.
25. (a) Define statistics. Why do some people look at this science with an eye of distrust?
   (b) 'The science of statistics is the most useful servant but only of great value to those who understand its proper use'. Discuss.
26. Bring out the applications of statistics in economics and business administration as a scientific tool. Also point out any two limitations of statistics.
    [CA Foundation, May 1996]
27. Give an example of the use of descriptive statistics and inferential statistics in each of the following areas of application in a business firm.
   (a) Production management
   (b) Financial management
   (c) Marketing management
   (d) Personnel management
28. Discuss the differences between statistics as numerical facts and as a discipline or field of study.
29. ORG conducts weekly surveys of television viewing throughout the country. The ORG statistical ratings indicate the size of the viewing audience for each major network television programme. Rankings of the television

programmes and of the viewing audience market shares for each network are published each week.

(a) What is the organization, ORG, attempting to measure?
(b) What is the population?
(c) Why would a sample be used for this situation?
(d) What kinds of decisions or actions are based on the ORG studies?

1.9 NEED FOR DATA

Statistical data are the basic material needed to make an effective decision in a particular situation. The main reasons for collecting data are as listed below:

(i) To provide necessary inputs to a given phenomenon or situation under study.
(ii) To measure performance in an ongoing process such as production, service, and so on.
(iii) To enhance the quality of decision-making by enumerating alternative courses of action in a decision-making process, and selecting an appropriate one.
(iv) To satisfy the desire to understand an unknown phenomenon.
(v) To assist in guessing the causes and probable effects of certain characteristics in given situations.

For any statistical analysis to be useful, the collection and use of input data is extremely important. One can collect an enormous amount of data on a subject of interest in a compact and usuable form from the internet. However, the reliability of such data is always doubtful. Thus, before relying on any interpreted data, either from a computer, internet or other source, we should study answers to the following questions: (i) Have data come from an unbaised source, that is, source should not have an interest in supplying the data that lead to a misleading conclusion, (ii) Do data represent the entire population under study i.e. how many observations should represent the population, (iii) Do the data support other evidences already available. Is any evidence missing that may cause to arrive at a different conclusion? and (iv) Are data support the logical conclusions drawn. Have we made conclusions which are not supported by data.

Nowadays computers are extensively used for processing data so as to draw logical conclusions. Since a computer is only a machine used for fast processing of input data, the output data received are only as accurate as the data that is fed in. The decision-maker thus needs to be careful that the data he is using comes from a valid source and evidences that might cause him to arrive at a different conclusion are not missing.

In order to design an experiment or conduct a survey one must understand the different types of data and their measurement levels.

1.9.1 Types of Data

Statistical data are the outcome of a continuous process of measuring, counting, and/or observing. These may pertain to several aspects of a phenomenon (or a problem) which are measurable, quantifiable, countable, or classifiable. While conducting a survey or making a study, an investigator develops a method to ask several questions to deal with the variety of characteristics of the given population or universe. These characteristics which one intends to investigate and analyse are termed as *variables*. The data, which are the observed outcomes of these variables, may vary from response to response. Consumer behaviour (attitude), profit/loss to a company, job satisfaction, drinking and/or smoking habits, leadership ability, class affiliation or status are examples of a variable.

Table 1.2 summarizes the types of variables which can be studied to yield the observed outcomes in relation to the nature of data, information, and measurement.

Table 1.2 Nature of Data, Information, and Measurement

Data Type		Information Type	Measurement Type
Categorical		→ Do you practice Yoga?	Yes ☐ No ☐
Numerical	Discrete	→ How many books do you have in your library?	Number
	Continuous	→ What is your height?	Centimetres or Inches

It may be noted from Table 1.2 that categorical variables are those which are not expressed in numerical terms. Sex, religion, and language are a few examples of such variables. The numerical variables are classified into two categories:

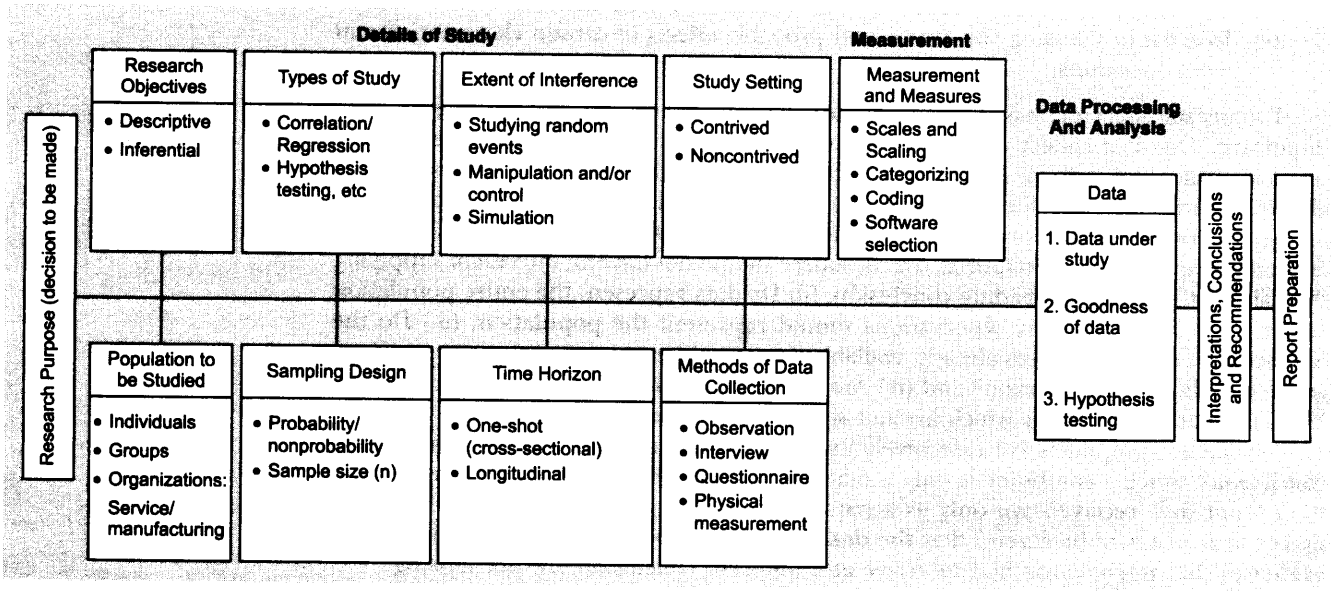
(i) Discrete variables—which can only take certain fixed integer numerical values. The number of cars sold by Maruti Udyog Ltd. in 2001, or the number of employees in an organization are examples of discrete variables.

(ii) Continuous variables—which can take any numerical value. Measurement of height, weight, length, in centimetres/inches, grams/kilograms are a few examples of continuous variables.

Remark: *Discrete data* are numerical measurements that arise from a process of counting, while *continuous data* are numerical measurements that arise from a process of measuring.

A flow chart of the research process is shown in Fig. 1.4.

Figure 1.4
A Flowchart of the Research Process



1.10 PRINCIPLES OF MEASUREMENT

Just as there are rules or guidelines that have to be followed to ensure that the wording of the questionnaire is appropriate to minimize bias, so also are some principles of measurement that are to be followed to ensure that the data collected are appropriate to test our hypothesis. These principles of measurement ecompass the scales and scaling techniques used in measuring concepts, as well as the assessment of reliability and validity of the measures used. Appropriate scales have to be used depending on the type of data that need to be obtained. Once data are obtained, the "goodness of data" is assessed through tests of validity and reliability. Validity established how well a technique, or a process measures a particular concept, the reliability indicates how stably and consistently the technique measures the variable.

In general, the principles of measurement (scaling) has three characteristics:

1. Numbers are ordered. One number is less than, equal to or greater than another number.
2. Difference between numbers are ordered. The difference between any pair of numbers is greater than, less than or equal to the difference between any other pair of numbers.
3. The number series has a unique origin indicated by the number zero.

The combinations of these characteristics of *order, distance* and *origin* provide the following widely used classification of measurement scales:

Scale of Measurement	Characteristics of Data	Determination
• Nominal	No order, distance or unique origin	Determination of categorical information. Numbers only identify groups which cannot be ordered
• Ordinal	Order but no distance or unique origin	Determination of greater or lesser values. Num-bers allow ranking but no arithmetic
• Interval	Both order and distance but not unique	Determination of equality of intervals or differences. Intervals between numbers are meaningful
• Ratio	Order, distance and unique origin	Determination of equality of ratios. Intervals between numbers are meaningful and also their ratio as the lowest value is a meaningful zero.

Nominal Scale In nominal scaling the numerical values are either named or categorized in such as way that these values are mutually exclusive and collectively exhaustive. For example, shirt numbers in a football or cricket match are measured at a nominal level. A player wearing a shirt number 24 is not more of anything than a player wearing a shirt number 12 and is certainly not twice the number 12. In other words, if we use numbers to identify categories, they are recognised as levels only and have no qualitative value.

Nominal scale: A scale of measurement for a variable that uses a label (or name) to identify an attribute of an element of the data set.

Nominal classifications may consists of any other number to separate groups if such groups are mutually exclusive and collectively exhaustive. For example, based on a nominal scale: each of the respondent has to fit into one of the six categories of nationality and scale will allow computation of the percentage of respondents who fit into each of these six categories

- Indian
- Nepalise
- Pakistanis
- Srilankan
- Bhottanis
- Others

Nominal scale is said to be least powerful among four scales because this scale suggest no order or distance relationship and have no arithmetic origin. Few examples of nominal scaling are: sex, blood type, religion, nationality, etc.

Nominal scale is usually used for obtaining personal data such as gender, place of work, and so on, where grouping of individuals or objects is useful, as illustrated below.

1. Your gender
   - Male
   - Female
2. Your place of work
   - Production
   - Sales
   - Finance
   - Personnel

Ordinal Scale In ordinal scaling the numerical values are categorised to denote qualitative differences among the various categories as well as rank-ordered the categories in some meaningful way according to some preference. The preferences would be ranked from best to worst, first to last, numbered 1, 2, and so on.

Ordinal scale: A scale of measurement for a variable that is used to rank (or order) observations in the data set.

The ordinal scale not only indicates the differences in the given items but also gives some information as to how respondents distinguish among these items by rank ordering them. However, the ordinal scale does not give any indication of the magnitude of the differences among the ranks, i.e. this scale implies a statement of 'greater than' or 'less than' (an equality statement is also acceptable) without stating how much greater or less. The real difference between ranks 1 and 2 may be more or less than the difference between ranks 4 and 5. The interval between values is not interpretable in an ordinal measure.

Besides 'greater than' and 'less than' measurements other measurements such as 'superior to', 'happier than' or 'above' may also be used as ordinal scale.

Oridinal scale is usually used to rate the preference or usage of various brands of a product by individuals and to rank individuals, objects, or events. For example, rank the following personal computers with respect to their usage in your office, assigning the number 1 to the most used system, 2 to the next most used, an so on. If particular system is not used at all in your office, put a 0 against it.

IBM/AT	Compaq
IBM/XT	AT&T
Apple II	Tandy 2000
Macintosh	Zenith

Interval scale: A scale of measurement for a variable in which the interval between observations is expressed in terms of a fixed standard unit of measurement.

Interval Scale An interval scale allows us to perform certain arithmetical operations on the data collected from the respondents. Whereas the nominal scale only allows us to qualitatively distinguish groups by categorizing them into mutually exclusive and collectively exhaustive sets, the oridinal scale allows us to rank-order the preferences, and the interval scale allows us to compute the mean and the standard deviation of the responses on the variables. In other words, the interval scale not only classify individuals according to certain categories and determines order of these categories; it also measure the magnitude of the differences in the preferences among the individuals.

In interval measurement the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30–40 is same as distance from 70–80. The interval values in interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for oridinal scales.

Interval scale is used when responses to various questions that measure a variable can be determined on a five-point (or seven-point or any other number of points) scale. For example, respondents may be asked to indicate their response to each of the questions by circling the number that best describes their feeling.

	Strongly Disagree 1	*Disagree* 2	*Neutral* 3	*Agree* 4	*Strongly Agree* 5
1. My job offers me a chance to test my abilities.	1	2	3	4	5
2. Mastering this job meant a lot to me.	1	2	3	4	5
3. Doing this job well is a reward in itself.	1	2	3	4	5
4. Considering the time spent on the job, feel thoroughly familiar with my tasks and responsibilities.	1	2	3	4	5

Ratio scale: A scale of measurement for a variable that has interval measurable is standard unit of measurement and a meaningful zero, i.e. the ratio of two values is meaningful.

Ratio Scale The ratio scale has an absolute measurement point. Thus the ratio scale not only measures the magnitude of the differences between points on the scale but also provides the proportions in the differences. It is the most powerful of the four scales because it has a unique zero origin. For example, a person weighing 90 kg is twice as heavy as one who weighs 45 kg. Since multiplying or dividing both of these numbers (90 and 45) by any given number will preserve the ratio of 2 : 1. The measure of central tendency of the ratio scale could either be arithmetic or geometric mean and the measure of dispersion could either be standard deviation or variance, or coefficient of variation.

Ratio scales are usually used in organizational research when exact figures on objective (as opposed to subjective) factors are desired. Few examples are as under:

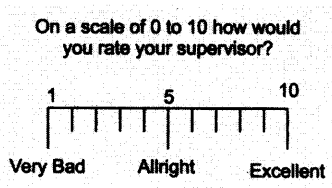
1. How many other organizations did you work for before joining this job?
2. Please indicate the number of children you have in each of the following categories:

- over 6 years but under 12
- 12 years and over

3. How many retail outlets do you operate?

The responses could range from 0 to any figure.

Graphic Rating Scale A graphical representation helps the respondent to indicate the response to a particular question by placing a mark at the appropriate point on the line as in the adjoining example.



Itemized Rating Scale This scale helps the respondent to choose one option that is most relevant for answering certain questions as in the following examples.

(*a*)

	Not at all interested	*Somewhat interested*	*Moderately interested*	*Very much interested*
How would you rate your interest in changing organizational policies?	1	2	3	4

(*b*)

	Extremely Poor	*Rather Poor*	*Quite Well*	*Very Well*	*Excellent*
How well is the new distribution channel working?	1	2	3	4	5

Other Measurement Scales

(a) **Continuous rating scales**

Type A

0 10 20 30 40 50 60 70 80 90 100

Unfavourable Neutral Favourable

Type B

Unfavourable Favourable

(b) **Itemized rating scale**

Type A

Favourable Unfavourable

_____ : _____ : _____ : _____ : _____ : _____ : _____

extremely quite slightly neither slightly quite extremely

Type B

Favourable Unfavourable

_____ : _____ : _____ : _____ : _____ : _____ : _____

Type C

Favourable

Unfavourable

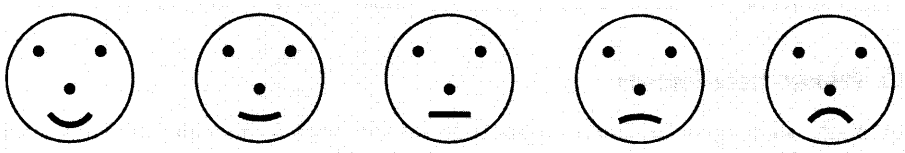
_____ : _____ : _____ : _____ : _____ : _____ : _____

7 6 5 4 3 2 1

Type D

Favourable Unfavourable

7 6 5 4 3 2 1



(c) **Stapel scale**

Perfectly 7 6 5 4 3 2 1 Not at all

For example describe your visit to Shimla during January

safe	_____	boring	_____
pleasant	_____	status	_____
risky	_____	enjoyable	_____
necessary	_____	old	_____
useless	_____	valuable	_____
attractive	_____	cold	_____

Similarly, given on the next page are five characteristics of an automobile. Allocate 100 points among the characteristics such that the allocation represents the importance of each characteristic to you. The more points a characteristic receives, the more important it is. If the characteristic is not at all important, it is possible to assign zero points. If a characteristic is twice as important as some other, then it should receive twice as many points.

Characteristics	*Number of Points*
• Styling	50
• Ride	10
• Petrol mileage	35
• Warranty	5
• Closeness to dealer	0
	100

(d) **Semantic differential scale**

Describe going to Delhi during the summer vacations:

important	_____ : _____ : _____ : _____ : _____ : _____ : _____	unimportant
worthless	_____ : _____ : _____ : _____ : _____ : _____ : _____	valuable
good	_____ : _____ : _____ : _____ : _____ : _____ : _____	bad
rewarding	_____ : _____ : _____ : _____ : _____ : _____ : _____	punishing
useful	_____ : _____ : _____ : _____ : _____ : _____ : _____	useless
pessimistic	_____ : _____ : _____ : _____ : _____ : _____ : _____	optimistic
hard	_____ : _____ : _____ : _____ : _____ : _____ : _____	soft
boring	_____ : _____ : _____ : _____ : _____ : _____ : _____	interesting
active	_____ : _____ : _____ : _____ : _____ : _____ : _____	passive
compulsory	_____ : _____ : _____ : _____ : _____ : _____ : _____	voluntary
serious	_____ : _____ : _____ : _____ : _____ : _____ : _____	humorous
pleasant	_____ : _____ : _____ : _____ : _____ : _____ : _____	unpleasant

1.11 SOURCES OF DATA

The choice of a data collection method from a particular source depends on the facilities available, the extent of accuracy required in analyses, the expertise of the investigator, the time span of the study, and the amount of money and other resources required for data collection. When the data to be collected are very voluminous and require huge amounts of money, manpower, and time, reasonably accurate conclusions can be drawn by observing even a small part of the population provided the concept of sampling is used objectively.

Data sources are classified as (i) primary sources, and (ii) secondary sources.

1.11.1 Primary Data Sources

Individuals, focus groups, and/or panels of respondents specifically decided upon and

set up by the investigator for data collection are examples of primary data sources. Any one or a combination of the following methods can be chosen to collect primary data:

(i) Direct personal observations
(ii) Direct or indirect oral interviews
(iii) Administrating questionnaires

The methods which may be used for primary data collection are briefly discussed below:

Observation In observational studies, the investigator does not ask questions to seek clarifications on certain issues. Instead he records the behaviour, as it occurs, of an event in which he is interested. Sometimes mechanical devices are also used to record the desired data.

Studies based on observations are best suited for researches requiring non-self report descriptive data. That is, when respondents' behaviours are to be understood without asking them to part with the needed information. Diverse opinions in the diagnosis of a particular disease could be an example of an observational study.

Certain difficulties do arise during the collection of such data on account of (i) the observer's training, philosophy, opinions, and expectations, (ii) the interdependence of observations and inferences, and (iii) the inadequacies of the sense organs causing significant variations in the observations of the same phenomenon.

Interviewing Interviews can be conducted either face-to-face or over telephone. Such interviews provide an opportunity to establish a rapport with the interviewer and help extract valuable information. Direct interviews are expensive and time-consuming if a big sample of respondents is to be personally interviewed. Interviewers' biases also come in the way. Such interviews should be conducted at the exploratory stages of research to handle concepts and situational factors.

Telephone interviews help establish contact with interviewers spread over distantly separated geographic locations and obtain responses quickly. This method is effective only when the interviewer has specific questions to ask the needs and responses promptly. Since the interviewer in this case cannot observe the non-verbal responses at the other end, the respondent can unilaterally terminate the interview without warning or explanation.

Questionnaire It is a formalized set of questions for extracting information from the target respondents. The form of the questions should correspond to the form of the required information. The three general forms of questions are: *dichotomous* (yes/no response type); *multiple choice*, and *open-ended*. A questionnaire can be administered personally or mailed to the respondents. It is an efficient method of collecting primary data when the investigator knows what exactly is required and how to measure such variables of interest as:

Questionnaire: A set of questions for extracting information from the target respondents.

- Behaviour—past, present, or intended.
- Demographic characteristics—age, sex, income, and occupation.
- Level of knowledge.
- Attitudes and opinions.

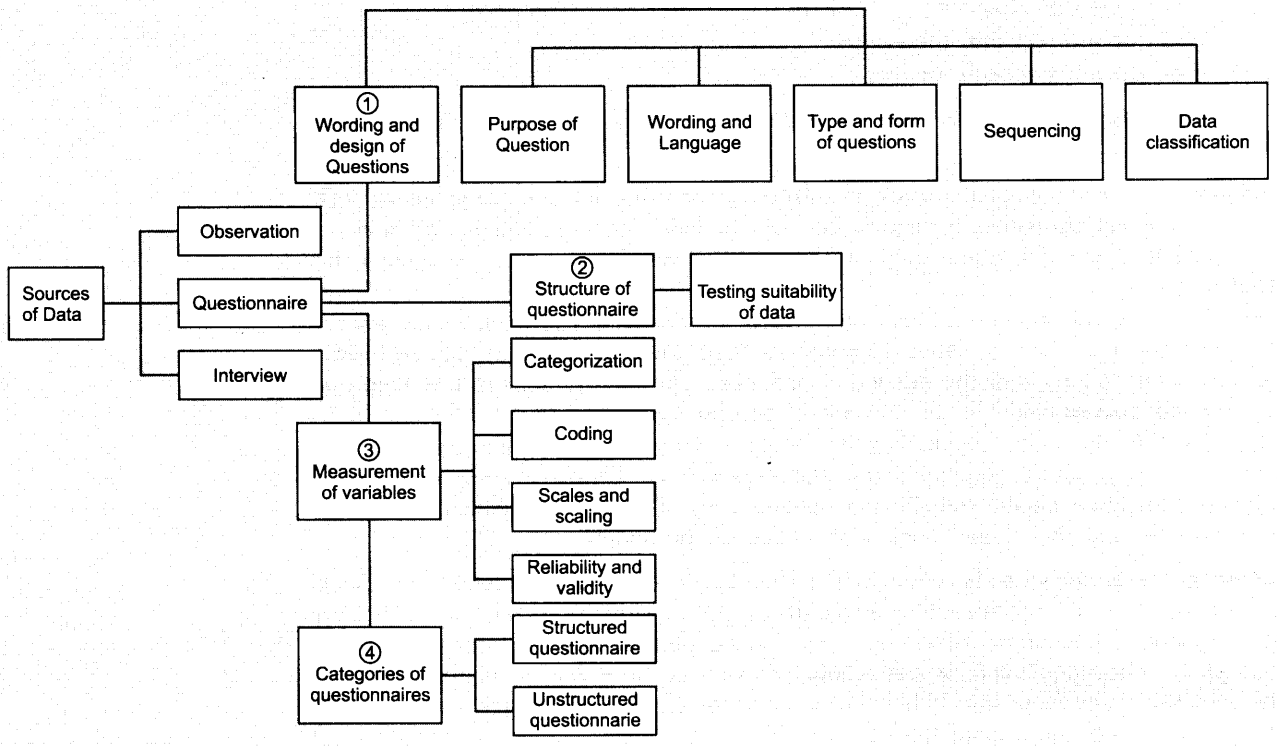
As such there are no set principles that must be used to design a questionnaire. However, general principles of questionnaire design based on numerous studies and experiences of survey researchers are shown in Fig. 1.5. A good questionnaire does, however, require the application of common sense, concern for the respondent, a clear concept of the information needed, and a thorough pre-testing of the questionnaire.

1. *The wording and design of questions* The writing of good questions is an art, and a time-consuming art at that! In order to obtain valid and reliable responses one needs well-worded questions. There are a number of pitfalls to be avoided:

- *Open Ended Versus Closed Questions:* Open-ended questions allow respondents to answer them in any way they choose. Examples of an open-ended question are :

(*i*) State five things that are interesting and challenging in the job,
(*ii*) What you like about your supervisors or work environment,
(*iii*) What is you opinion about investment portfolio of your organization.

Figure 1.5
Principles of Questionnaire Design



A *closed* question, would ask the respondents to make choices among a set of alternatives. For instance, instead of asking the respondent to state any five aspects of the job that are interesting and challenging, the researcher might list ten or fifteen characteristics that might seem interesting or challenging in jobs and ask the respondent to rank the first five among these.

Closed questions help the respondent to make quick decision by making a choice among the several alternatives that are provided. They also help the researcher to code the information easily for subsequent analysis. Of course, care has to be taken to ensure that the alternatives are mutually exclusive and collectively exhaustive. If there are overlapping categories, or if all possible alternatives are not given (i.e., the categories are not exhaustive), the respondents might get confused and the advantage of making a quick decision may be lost.

- *Positively and Negatively Worded Questions:* Instead of phrasing all questions positively, it is advisable to include some negatively worded questions also, so that it minimizes the tendency in respondents to mechanically circle the points toward one end of the scale. For example, a set of six questions are used to measure the variable 'perceived success' on a five-point scale, with 1 being 'very low' and 5 being 'very high' on the scale. A respondent who is not particularly interested in completing the questionnaire is more likely to stay involved and remain alert while answering the questions when positively and negatively worded questions are interspersed in the questionnaire. For instance, if the respondent had circled 5 for a positively worded question such as, 'I feel I have been able to accomplish a number of different things in my job' he cannot circle number 5 again to the negatively worded questions, 'I do not feel I am very effective in my job.' The use of double negatives and excessive they tend to confuse respondents. For instance, it is better to say 'Coming to work is not great fun' than to say 'Not coming to work is greater fun than coming to work.' Likewise, it is better to say 'The strong people need no tonics' than to say 'Only the strong should be given no tonics.'
- *Double-Barreled Questions:* A question that lends itself to different possible answers to its subparts is called a double-barreled question. Such questions should be avoided

and two or more separate questions should be ask. For example, the question "Do you think there is a good market for the product and that it will sell well? could bring a 'yes' response to the first part (i.e., there is a good market for the product) and a 'no' response to the latter part (i.e., it will not sell well—for various other reasons). In this case, it would be better to ask two questions such as: (a) 'Do you think there is a good market for the product?' (b) 'Do you think the product will sell well?'

- *Ambigious Questions:* Questions that can be interpreted different by different respondents should be avoided. For example, for the question such as: 'To what extent would you say you are happy?', the respondent might not be sure whether the question refers to his feelings at the workplace, or at home, or in general. Because, respondent might presume that the question relates to the workplace. Yet the researcher might have intended to inquire about the, overall degree of satisfaction that the respondent experiences in everyday life–a feeling not specific to the workplace alone or at home.
- *Level of Wording:* It is important to tailor the level of wording of questions in accordance with the understanding respondents'. Jargons are to be avoided, and it should be established in the pilot study that the respondents understand the concepts. For instance, asking questions about 'Trisomy 21' might be inappropriate while 'mongolism' or 'Down syndrome' could be intelligible. Using double negatives should be avoided. In general, the questions should be simple and concise.
- *Biased and Leading Questions:* The wording of the questions should not lead the respondent to feel committed to respond in a certain way. For example, the question 'How often do you go to church?' may lead the respondent to respond in a way that is not entirely truthful if they, in fact, never go to church. Not only can the wording of a question be leading but the response format may also be leading. For example, if a 'never' response were excluded from the available answers to the above question, the respondent would be led to respond in an inaccurate way.

Bias might also arise from possible carry-over effects from answering a pattern of questions. For instance, a questionnaire on health workers' attitudes to abortion might include the questions 'Do you value human life?' followed by 'Do you think unborn babies should be murdered in their mothers' wombs?'. In this case, the respondent is being led both by the context in which the second question is asked and the bias involved in the emotional wording of the questions. Surely, one would have to be a monster to answer 'yes' to the second question, given the way it was asked.

Finally, it should be kept in mind that even a good questionnaire might be invalidly administered. For instance, a survey on 'Attitudes to migration' might be answered less than honestly by respondents if the interviewer is obviously of immigrant background.

2. The structure of questionnaire A questionnaire may be structured in different ways, but typically the following components are included:

- *Introductory Statement:* The introductory statement describes the purpose of the questionnaire, the information sought, and how it is to be used. It also introduces the researchers and explains whether the information is confidential and/or anonymous.
- *Demographic Questions:* It is usual to collect information about the respondents, including details such as age, sex, education, and so on. It is best to position these questions first as they are easily answered and serve as a 'warm-up' to what follows.
- *Factual Questions:* It is generally easier for respondents to answer direct factual questions such as, 'Do you have a driver's licence?' than to answer opinion questions. Often, this type of question is positioned early on in the questionnaire—also to help 'warm up'!

  (iv) *Opinion Questions:* Questions that require reflection on the part of the respondent are usually positioned after the demographic and factual questions.
- *Closing Statements and Return Instructions:* The closing statements in a questionnaire usually thank the respondent for their participation, invite the respondent to take up any issues they feel have not been satisfactorily addressed in the questionnaire, and provide information on how to return the questionnaire.

It is best to avoid complicated structures involving, for example, many conditional questions such as 'If you answered yes to Question 6 and no to Question 9, please answer Question 10'. Conditional questions usually confuse respondents and ought be avoided where possible.

3. Categories of questionnaires

- *Structured Questionnaire:* It is a formal list of questions to be posed to the respondents in a predetermined order. The responses permitted are also completely predetermined. Such questions are often called *closed questions* since the respondents are asked to make choices among a set of alternatives given by the investigator.

  A structured questionnaire can also be *disguised* and *non-disguised*. This classification is based on whether the objectives of the study are disclosed or not disclosed to the respondents. A *structured undisguised questionnaire* is one where the purpose of the study and the particulars of the sponsor are disclosed to the respondent. In such cases, the questionnaire contains a list of questions in a predetermined order and freedom of response is limited only to the stated alternatives. Such questions help the respondent to make quick decisions by making a choice among the given alternatives. The alternatives provided have to be mutually exclusive and collectively exhaustive.

  In the case of a *structured disguised questionnaire*, the objectives of the study and its sponsor are not disclosed to the respondents. Such questionnaires are not often used because it is felt necessary to have the respondents taken into confidence so that they appreciate the relevance of the desired information needed and willingly offer accurate answers.

- *Unstructured Questionnaire:* In this case, the investigator does not offer a limited set of response choices, but provides only a frame of reference within which the respondents are expected to answer. Such questionnaires are sometimes referred to as *open-ended questions*. Examples of open-ended questions are:

  (i) State three things that are interesting and challenging in your job.
  (ii) State about the behaviour of a supervisor or the work environment.

  These questions encourages the respondents to share as much information as possible in a free environment. The investigator may also provide extra guidance to the respondents by using a set of questions to promote discussion and elaboration.

  The unstructured questionnaire is used in exploratory research studies or where the investigator is dealing with a complex phenomenon which does not lend itself to structured questioning. Such questionnaires are also useful when the investigator requires to know the respondent's emotions, needs, motivation level, attitude, and values. Obviously, using a questionnaire of this type, needs more time per interview and, therefore, raises the cost of the study. Editing and tabulation of these questionnaires also impose practical difficulties. Interestingly, unstructured questionnaires could also be of two types— *disguised* and *undisguised*.

Examples of questionnaire design

Two sets of questionnaire having most of the qualities of a good questionnaire are as under:

Questionnaire 1: *Consumer Preferences*

Name: _____ Age: _____

Address: _____

City: _____ Pin: _____ Phone: _____

Marital status: ☐ Married ☐ Single Occupation: _____

Family type: ☐ Joint ☐ Nuclear

Family members: ☐ Adults ☐ Children

Family income: ☐ Less than 10,000 ☐ 10,000 to 15,000
☐ 15,000 to 20,000 ☐ More than 20,000

Remarks (if any):

Place and Date:

1. What kind of food do you normally eat at home?
   ☐ North Indian ☐ South Indian ☐ Mughlai ☐ Chinese
   ☐ Continental ☐ Italian ☐ Fast Food ☐ Others _____
2. How frequently do you eat out?
   In a week ☐ Once ☐ Twice ☐ Thrice ☐ More than thrice
   In a fortnight ☐ Once ☐ Twice ☐ Thrice ☐ More than thrice
   In a month ☐ Once ☐ Twice ☐ Thrice ☐ More than thrice
3. You usually go out with:
   ☐ Family ☐ Friends ☐ Colleagues ☐ Others _____
4. Any specific days when you go out:
   ☐ Weekdays ☐ Weekends ☐ Holidays ☐ Special Occassions
   ☐ No specific days
5. You generally go out for:
   ☐ Lunch ☐ Snacks ☐ Dinner ☐ Party/Picnics
   ☐ Others _____
6. Where do you usually go:
   ☐ Restaurant ☐ Chinese joint ☐ Fast food joint ☐ Others _____
7. Who decides on the place to go:
   ☐ Husband ☐ Wife ☐ Children ☐ Others _____
8. How much do you spend on eating out (one time):
   ☐ Below 200 ☐ 200-500 ☐ 500-800 ☐ More than 800
9. What are the factors that determine your choice for the restaurant/joint?
   Rank the following from 1-9 (9-highest score):
   ☐ Restaurant ☐ Chinese joint ☐ Fast food joint ☐ Others _____
10. Name the fast food giants that you are aware of (in Delhi):
    ☐ Nirula's ☐ Wimpy's ☐ McDonalds ☐ Pizza Hut
    ☐ Dominos ☐ Slice of Italy ☐ Pizza Express ☐ Others _____
11. How frequently do you go out/order for fast food?
    ☐ Very frequently ☐ Often ☐ Sometimes ☐ Never
12. What do you prefer: ☐ Going Out ☐ Home Delivery ☐ Take Away
13. Which of the places mentioned above in Q.10 are visited by you—(and why):
    (a) Most frequently _____
    (b) Sometimes _____
    (c) Never _____
14. What are the distinguishing factors you look for in fast food service:
    (Rank from 1 to 8, 8-highest score)
    ☐ Quality ☐ Service ☐ Location ☐ Wide Menu Range
    ☐ Price ☐ Taste ☐ Home Delivery ☐ Others _____
15. What your order normally consists of:
    ☐ Pizza ☐ Burgers ☐ Footlong ☐ Curries & Breads
    ☐ Soups ☐ Pasta ☐ Desert ☐ Others _____
16. The price paid by you for the above is:

Outlets	*Very High*	*Little High*	*Just Right*
Nirula's			
Wimpy's			
Pizza Hut			
Domino's			
Slice of Italy			
Pizza Express			
McDonalds			
Others			

17. If you feel that the price paid by you is very high, what should be the price according to you:

Items	*Vegetarian*	*Non-Vegetarian*
Pizza		
Burger		
Footlong		
Others		

Questionnaire 2: *Journal outlets for Production/Operations Management (POM) Research*

If you have *not received* a Ph.D. degree, and have *not accepted* a full-time teaching position yet, mark the tick (✓) and stop. You need not complete the questionnaire.

If you have *not received* a Ph.D. degree but have *accepted* a full-time teaching position somewhere, mark the tick (✓). Skip Question 11, and answer all other questions.

1. How relevant do you consider the journal as a Production/Operations Management (POM)-related research outlet? Use the scale below.

1	2	3	4	5	6	7	8	9
Most relevant		Quite relevant		Relevant		Somewhat relevant		Not relevant

2. Based on the quality of the POM-related articles published, how would you rate the journal? (Use the scale below)

1	2	3	4	5	6	7	8	9	0
Level A		Level A–		Level B		Level B–		Level C	Not possible to rate

3. How does your institution/college rate this journal? (Use the scale in Question 2)
4. How many articles have you authored or coauthored in this journal (include any articles that are currently in the press)?

Academy of Management Journal	____	____	____	____	____
Academy of Management Review	____	____	____	____	____
Computers and Industrial Engineering	____	____	____	____	____
Computers and Operations Research	____	____	____	____	____
Decision Sciences	____	____	____	____	____
European Journal of Operational Research	____	____	____	____	____
Harvard Business Review	____	____	____	____	____
Interfaces	____	____	____	____	____
Journal of Operations Management	____	____	____	____	____
Journal of Operational Research Society	____	____	____	____	____
Journal of Purchasing and Materials Management	____	____	____	____	____
Management Science	____	____	____	____	____
Naval Research Logistics Quarterly	____	____	____	____	____
Omega	____	____	____	____	____
Operations Research	____	____	____	____	____
Production and Inventory Management	____	____	____	____	____
Production and Operations Management	____	____	____	____	____
(List below any other journal that you consider related to POM research)					
_____	____	____	____	____	____
_____	____	____	____	____	____
_____	____	____	____	____	____

5. Using the scale below, please indicate the importance of the following factors in your assessment of the quality of a POM journal.

1	2	3	4	5	6	7	8	9
Most important		Very important		Important		Somewhat important		Not important at all

___ Acceptance rate
___ Methodological rigour of the published work
___ Editor and editorial board members
___ Authors who publish in the journal
___ Number of issues per year
___ Age of the journal
___ Professional organization that sponsors the journal
___ Other (please specify)

6. At this stage of your career, how important to your career advancement is the quality of the journals in which your articles appear? (Use the scale below)

1	2	3	4	5	6	7	8	9
Most important		Very important		Important		Somewhat important		Not important at all

7. At this stage of your career, how important to your career advancement is the quality of articles you author/coauthor? (use the scale below)

1	2	3	4	5	6	7	8	9
Most important		Very important		Important		Somewhat important		Not important at all

8. How much weightage does your institution/college place on research and publication in evaluating your annual performance? _____ (use a number between 0 and 100%)
9. What business degree(s) is (are) offered by the institution in which you teach? (tick all that apply)
   ☐ Undergraduate ☐ Masters level (MBA, MCA, M.Tech, etc.)
   ☐ Doctoral; (M.Phil, Ph.D.)
10. What is your academic rank?
    ☐ Full professor ☐ Associate professor ☐ Assistant professor
    ☐ Other (e.g., instructor, lecturer, etc.)
11. In which year was your Ph.D. degree granted? _____
12. How many POM-related articles have you authored/coauthored in referenced journals? (include any articles that are currently in the press) _____

1.11.2 Secondary Data Sources

Secondary data refer to those data which have been collected earlier for some purpose other than the analysis currently being undertaken. Besides newspapers and business magazines, other sources of such data are as follows:

1. External secondary data sources

- Government publications, which include
  (i) The National Accounts Statistics, published by the Central Statistical Organization (CSO). It contains estimates of national income for several years, growth rate, and rate on major economic activities such as agriculture, industry, trade, transport, and so on;
  (ii) Wholesale Price Index, published by the office of the Economic Advisor, Ministry of Commerce and Industry;
  (iii) Consumer Price Index;
  (iv) Reserve Bank of India bulletins;
  (v) Economic Survey.
- Non-Government publications include publications of various industrial and trade associations such as
  (i) The Indian Cotton Mills Association

(ii) The various Chambers of Commerce
(iii) The Bombay Stock Exchange, which publishes a directory containing financial accounts, key profitability and other relevant data.

- Various syndicate services such as Operations Research Group (ORG). The Indian Market Research Bureau (IMRB) also collects and tabulates abundant marketing information to suit the requirements of individual firms, making the same available at regular intervals.
- International organizations which publish data are:
  (i) The International Labour Organization (ILO)—which publishes data on the total and active population, employment, unemployment, wages, and consumer prices.
  (ii) The Organization for Economic Cooperation and Development (OECD)—which publishes data on foreign trade, industry, food, transport, and science and technology.
  (iii) The International Monetary Fund (IMF)—which publishes reports on national and international foreign exchange regulations and other trade barriers, foreign trade, and economic developments.

2. Internal secondary data sources The data generated within an organization in the process of routine business activities, are referred to as internal secondary data. Financial accounts, production, quality control, and sales records are examples of such data. However, data originating from one department of an organization may not be useful for another department in its original form. It is, therefore, desirable to condense such data into a form needed by the other.

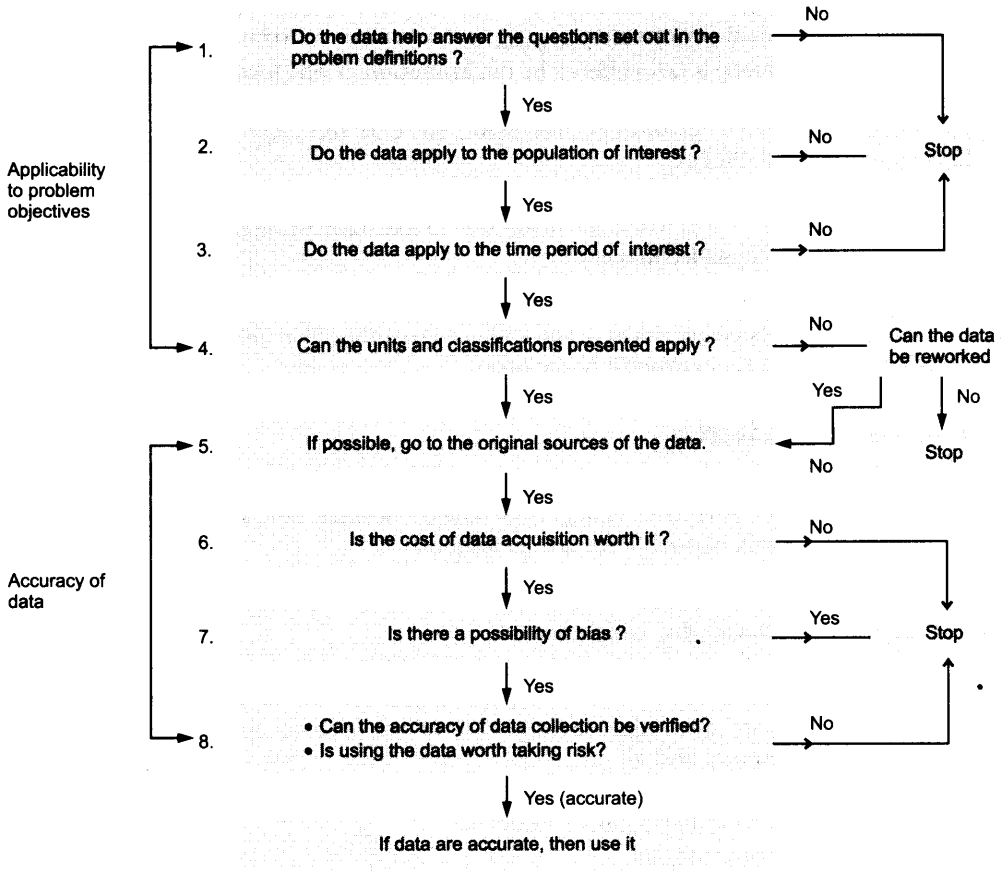


Figure 1.6
Flow Chart Showing the Procedure for Evaluating Secondary Data

Advantages and Disadvantages of Secondary Data

Secondary data have their own advantages and disadvantages. The advantages are that such data are easy to collect and involve relatively lesser time and cost. Deficiencies and gaps can be identified easily and steps taken promptly to overcome the same.

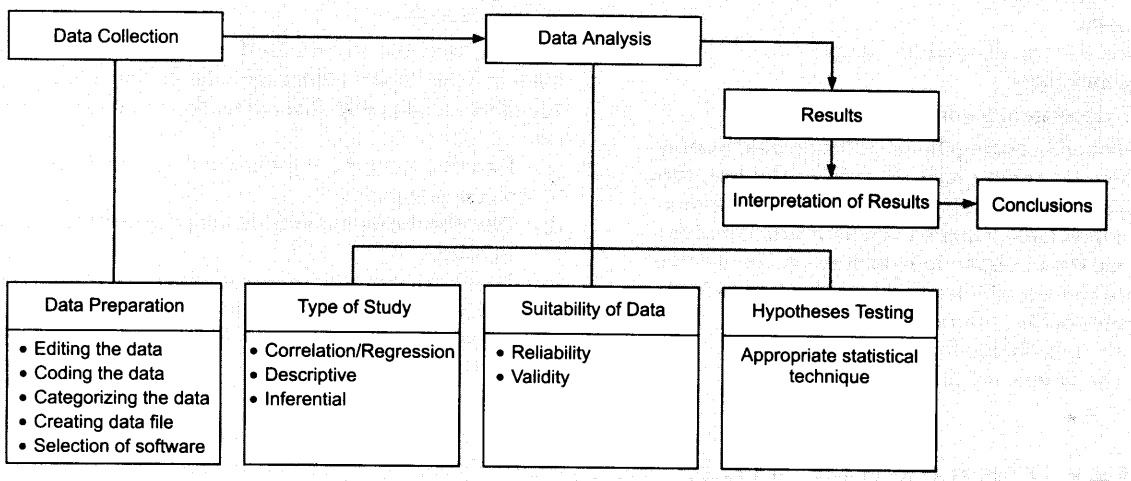
Their disadvantage is that the unit of measurement may not be the same as required by the users. For example, the size of a firm may be stated in terms of either number of employees, gross sales, gross profit, or total paid-up capital.

The scale of measurement may also be different from the one desired. For example, dividend declared by various companies may have breakup of 'less than 10 per cent' '10–15 per cent'; '15–20 per cent,' and so on. For a study requiring to know the number of companies who may have declared dividend of '16 per cent and above', such secondary data are of no use.

Robert W. Joselyn in his book *Designing and Marketing Research Project*, Petrocelli/ Charter, 1977, New York, suggested an approach for evaluating the usefulness of secondary data and understanding their limitations. The flow chart showing the steps to be taken for evaluating the secondary data is shown in Fig. 1.6.

After data have been collected from a representative sample of the population, the next step is to analyse the data so that the research hypotheses can be tested. For this, some preliminary steps need to followed. These steps help to prepare the data for analysis, ensure that the data obtained are reasonably good, and allow the results to be meaningfully interpreted. A flow diagram in Fig. 1.7 shows the data analysis process.

Figure 1.7
Flow Diagram of Data Analysis Process



Conceptual Questions 1B

30. A manager of a large corporation has recommended that a Rs 1000 raise be given to keep a valued subordinate from moving to another company. What internal and external sources of data might be used to decide whether such a salary increase is appropriate?

31. In the area of statistical measurement instruments such as questionnaires, *reliability* refers to the consistency of the measuring instrument and *validity* refers to the accuracy of the instrument. Thus, if a questionnaire yields comparable or similar results when completed by two equivalent groups of respondents, then the questionnaire can be described as being reliable. Does the fact that an instrument is reliable guarantee that it is also a valid instrument? Discuss.

32. Describe the three basic steps involved in the development and use of a written questionnaire prior to actual data analysis.

33. Describe the three general forms of questions that can be included in a questionnaire and give an example of each in the context of a political poll.

34. One can say that statistical inference includes an interest in statistical description as well, since the ultimate purpose of statistical inference is to *describe* population data. How then, does statistical inference differ from statistical description? Discuss.

35. In a recent study of causes of death in men 60 years of age and older, a sample of 120 men indicated that 48 died as a result of some form of heart disease.

(a) Develop a descriptive statistic that can be used as an estimate of the percentage of men 60 years of age or older who die from some form of heart disease

(b) Are the data on the causes of death qualitative or quantitative?

(c) Discuss the role of statistical inference in this type of medical research

36. Determine whether each of the following random variables is categorical or numerical. If it is numerical, determine whether the phenomenon of interest is discrete or continuous.
    (a) Amount of time the personal computer is used per week
    (b) Number of persons in the household who use the personal computer
    (c) Amount of money spent on clothing in the last month
    (d) Favourite shopping centre.
    (e) Amount of time spent shopping for clothing in the last month
37. State whether each of the following variables is qualitative or quantitative and indicate the measurement scale that is appropriate for each.
    (a) Age
    (b) Gender
    (c) Class rank
    (d) Make of automobile
    (e) Annual sales
    (f) Soft-drink size (small, medium, large)
    (g) Earnings per share
    (h) Method of payment (cash, check, credit card)
38. A firm is interested in testing the advertising effectiveness of a new television commercial. As part of the test, the commercial is shown on a 6:30 p.m. local news programme in Delhi. Two days later, a market research firm conducts a telephone survey to obtain information on recall rates (percentage of viewers who recall seeing the commercial) and impressions of the commercial.
    (a) What is the population for this study?
    (b) What is the sample for this study?
    (c) Why would a sample be used in this situation? Explain.
39. Suppose the following information is obtained from a person on his application for a home loan from a bank:
    (a) Place of residence: GK II, New Delhi
    (b) Type of residence: Single-family home
    (c) Date of birth: 14 August 1975
    (d) Monthly income: Rs 25,000
    (e) Occupation: Systems Engineer
    (f) Employer: Telecom company
    (g) Number of years at job: 5
    (h) Other income: Rs 30,000 per year
    (i) Marital status: Married
    (j) Number of children: 1
    (k) Loan requested: Rs 5,00,000
    (l) Term of Loan: 10 years
    (m) Other loan: Car
    (n) Amount of other loan: Rs 1,00,000

    Classify each of the 14 responses by type of data and level of measurement.
40. Suppose that the Rotary Club was planning to survey 2000 of its members primarily to determine the percentage of its membership that currently own more than one car.
    (a) Describe both the population and the sample of interest to the club
    (b) Describe the type of data that the club primarily wishes to collect
    (c) Develop the questionnaire needed by writing a series of five categorical questions and five numerical questions that you feel would be appropriate for this survey

Chapter Concepts Quiz

True or False

1. The scale of measurement of a variable is a nominal scale when data are labels to identify an attribute of the element. (T/F)
2. The statistical method used to summarize data depends upon whether the data are qualitative or quantitative. (T/F)
3. Statistical studies can be classified as either experimental or observational. (T/F)
4. Learning statistics does not help to improve processes. (T/F)
5. Statistics cannot be misused. (T/F)
6. All facts numerically stated are not statistics. (T/F)
7. Statistical thinking focuses on ways to understand, manage, and reduce variation. (T/F)
8. An average value computed from the set of all observations in the population is called a statistic. (T/F)
9. Inferential statistics help in generalizing the results of a sample to the entire population. (T/F)
10. Descriptive statistical methods are used for presenting and characterizing data. (T/F)
11. A statistic is a summary measure that describes the characteristic of a population. (T/F)
12. A descriptive measure computed from a sample of the population is called a parameter. (T/F)
13. Enumerative studies involve decision-making regarding a population and/or its characteristics. (T/F)
14. Analytical studies involve taking some action on a process to improve performance in the future. (T/F)
15. Data are needed to satisfy our curiosity. (T/F)
16. A continuous variable can also be used for quantitative data when every value within some interval is a possible result. (T/F)
17. A summary measure computed from sample data is called statistic. (T/F)
18. The summary numbers for either a population or a sample are called descriptive statistics. (T/F)

Concepts Quiz Answers

1. T	2. T	3. T	4. F	5. F	6. T	7. T	8. F	9. T
10. T	11. F	12. F	13. T	14. T	15. T	16. T	17. T	18. T

Chapter 2

It's not the figures themselves . . ., it's what you do with them that matters.
—K. A. C. Manderville

If you torture the data long enough, it will confess.
—Ronald Coase

Data Classification, Tabulation and Presentation

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand types of data and the basis of their classification.
- use techniques of organizing data in tabular and graphical form in order to enhance data analysis and interpretation.

2.1 INTRODUCTION

In Chapter 1, we learned how to collect data through primary and/or secondary sources. Whenever a set of data that we have collected contains a large number of observations, the best way to examine such data is to present it in some compact and orderly form. Such a need arises because data contained in a questionnaire are in a form which does not give any idea about the salient features of the problem under study. Such data are not directly suitable for *analysis* and *interpretation*. For this reason the data set is organized and summarized in such a way that patterns are revealed and are more easily interpreted. Such an arrangement of data is known as the *distribution* of the data. Distribution is important because it reveals the pattern of variation and helps in a better understanding of the phenomenon the data present.

2.2 CLASSIFICATION OF DATA

Classification of data is the process of arranging data in groups/classes on the basis of certain properties. The classification of statistical data serves the following purposes:

(i) It condenses the raw data into a form suitable for statistical analysis.
(ii) It removes complexities and highlights the features of the data.
(iii) It facilitates comparisons and in drawing inferences from the data. For example, if university students in a particular course are divided according to sex, their results can be compared.

(iv) It provides information about the mutual relationships among elements of a data set. For example, based on literacy and criminal tendency of a group of peoples, it can be established whether literacy has any impact or not on criminal tendency.

(v) It helps in statistical analysis by separating elements of the data set into homogeneous groups and hence brings out the points of similarity and dissimilarity.

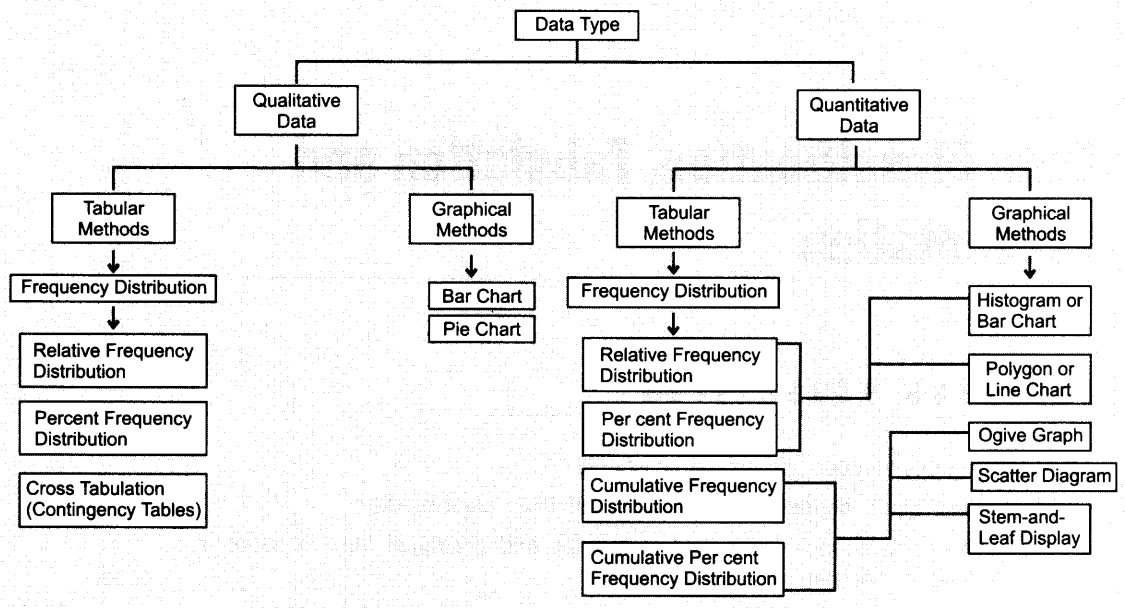


Figure 2.1
Tabular and Graphical Methods For Summarizing Data

2.2.1 Requisites of Ideal Classification

The classification of data is decided after taking into consideration the nature, scope, and purpose of the investigation. However, an ideal classification should have following characteristics:

It should be unambiguous It is necessary that the various classes should be so defined that there is no room for confusion. There must be only one class for each element of the data set. For example, if the population of the country is divided into two classes, say literates and illiterates, then an exhaustive definition of the terms used would be essential.

Classes should be exhaustive and mutually exclusive Each element of the data set must belong to a class. For this, an extra class can be created with the title 'others' so as to accommodate all the remaining elements of the data set.

Each class should be mutually exclusive so that each element must belong to only one class. For example, classification of students according to the age: below 25 years and more than 20 years, is not correct because students of age 20 to 25 may belong to both the classes.

It should be stable The classification of a data set into various classes must be done in such a manner that if each time an investigation is conducted, it remains unchanged and hence the results of one investigation may be compared with that of another. For example, classification of the country's population by a census survey based on occupation suffers from this defect because various occupations are defined in different ways in successive censuses and, as such, these figures are not strictly comparable.

It should be flexible A classification should be flexible so that suitable adjustments can be made in new situations and circumstances. However, flexibility does not mean instability. The data should be divided into few major classes which must be further subdivided. Ordinarily there would not be many changes in the major classes. Only small sub-classes

may need a change and the classification can thus retain the merit of stability and yet have flexibility.

The term stability does not mean rigidity of classes. The term is used in a relative sense. One-time classification can not remain stable forever. With change in time, some classes become obsolete and have to be dropped and fresh classes have to be added. The classification may be called ideal if it can adjust itself to these changes and yet retain its stability.

2.2.2 Basis of Classification

Statistical data are classified after taking into account the nature, scope, and purpose of an investigation. Generally, data are classified on the basis of the following four bases:

Geographical Classification In geographical classification, data are classified on the basis of geographical or locational differences such as—cities, districts, or villages between various elements of the data set. The following is an example of a geographical distribution

City	:	Mumbai	Kolkata	Delhi	Chennai
Population density (per square km)	:	654	685	423	205

Such a classification is also known as *spatial classification*. Geographical classifications are generally listed in alphabetical order. Elements in the data set are also listed by the frequency size to emphasize the importance of various geographical regions as in ranking the metropolitan cities by population density. The first approach is followed in case of reference tables while the second approach is followed in the case of summary tables.

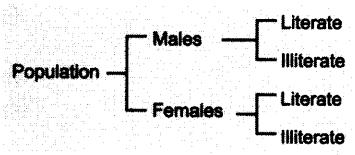
Chronological Classification When data are classified on the basis of time, the classification is known as chronological classification. Such classifications are also called *time series* because data are usually listed in chronological order starting with the earliest period. The following example would give an idea of chronological classification:

Year	:	1941	1951	1961	1971	1981	1991	2001
Population (crore)	:	31.9	36.9	43.9	54.7	75.6	85.9	98.6

Qualitative Classification In qualitative classification, data are classified on the basis of descriptive characteristics or on the basis of attributes like sex, literacy, region, caste, or education, which cannot be quantified. This is done in two ways:

(i) *Simple classification:* In this type of classification, each class is subdivided into two sub-classes and only one attribute is studied such as: male and female; blind and not blind, educated and uneducated, and so on.

(ii) *Manifold classification:* In this type of classification, a class is subdivided into more than two sub-classes which may be sub-divided further. An example of this form of classification is shown in the box:



Quantitative Classification In this classification, data are classified on the basis of some characteristics which can be measured such as height, weight, income, expenditure, production, or sales.

Quantitative variables can be divided into the following two types. The term variable refers to any quantity or attribute whose value varies from one investigation to another.

(i) *Continuous variable* is the one that can take any value within the range of numbers. Thus the height or weight of individuals can be of any value within the limits. In such a case data are obtained by measurement,

(ii) *Discrete* (also called *discontinuous*) *variable* is the one whose values change by steps or jumps and can not assume a fractional value. The number of children in a family, number of workers (or employees), number of students in a class, are few examples of a discrete variable. In such a case data are obtained by counting.

The following are examples of continuous and discrete variables in a data set:

Table 2.1

Discrete Series		*Continuous Series*	
Number of Children	*Number of Families*	*Weight (kg)*	*Number of Persons*
0	10	100 to 110	10
1	30	110 to 120	20
2	60	120 to 130	25
3	90	130 to 140	35
4	110	140 to 150	50
5	20		
	320		140

2.3 ORGANIZING DATA USING DATA ARRAY

The best way to examine a large set of numerical data is first to organize and present it in an appropriate tabular and graphical format.

Table 2.2 presents the total number of overtime hours worked for 30 consecutive weeks by machinists in a machine shop. The data displayed here are in *raw form*, that is, the numerical observations are not arranged in any particular order or sequence.

Table 2.2 Raw Data Pertaining to Total Time Hours Worked by Machinists

94	89	88	89	90	94	92	88	87	85
88	93	94	93	94	93	92	88	94	90
93	84	93	84	91	93	85	91	89	95

These raw data are not amenable even to a simple reading and do not highlight any characteristic/trend, such as the highest, the lowest, and the average weekly hours. Even a careful look at these data do not easily reveal any significant trend regarding the nature and pattern of variations therein. As such no meaningful inference can be drawn, unless these data are reorganized to make them more useful. For example, if we are to ascertain a value around which most of the overtime hours cluster, such a value is difficult to obtain from the raw data.

Moreover, as the number of observations gets large, it becomes more and more difficult to focus on the specific features in a set of data. Thus we need to organize the observation so that we can better understand the information that the data are revealing.

The raw data can be reorganized in a data array and frequency distribution. Such an arrangement enables us to see quickly some of the characteristics of the data we have collected.

When a raw data set is arranged in rank order, from the smallest to the largest observation or vice-versa, the ordered sequence obtained is called an *ordered array*. Table 2.3 reorganizes data given in Table 2.2 in the ascending order

Table 2.3 Ordered Array of Total Overtime Hours Worked by Machinists

84	84	85	85	87	88	88	88
88	89	89	89	90	90	91	91
92	92	93	93	93	93	93	93
94	94	94	94	94	95		

It may be observed that an ordered array does not summarize the data in any way as the number of observations in the array remains the same. However, a few advantages of ordered arrays are as under:

Advantages and Disadvantages of Ordered Array

Advantages The following are a few advantages of an ordered array:

(i) It provides a quick look at the highest and lowest observations in the data within which individual values vary.
(ii) It helps in dividing the data into various sections or parts.
(iii) It enables us to know the degree of concentration around a particular observation.
(iv) It helps to identify whether any values appear more than once in the array.

Disadvantages In spite of various advantages on converting a set of raw data into an ordered array, an array is a cumbersome form of presentation which is tiresome to construct. It neither summarizes nor organizes the data to present them in a more meaningful way. It also fails to highlight the salient characteristics of the data which may be crucial in terms of their relevance to decision-making.

The above task cannot be accomplished unless the observations are appropriately condensed. The best way to do so is to display them into a convenient number of groupings with the number of observations falling in different groups indicated against each. Such tabular summary presentation showing the number (frequency) of observations in each of several non-overlapping classes or groups is known as *frequency distribution* (also referred to as *grouped data*).

2.3.1 Frequency Distribution

Frequency distribution: A tabular summary of data showing the number (frequency) of observations in each of several non overlapping class intervals.

A frequency distribution divides observations in the data set into conveniently established, numerically ordered classes (groups or categories). The number of observations in each class is referred to as *frequency* denoted as *f*.

Few examples of instances where frequency distributions would be useful are when (i) a marketing manager wants to know how many units (and what proportions or percentage) of each product sells in a particular region during a given period, (ii) a tax consultant desires to keep count of the number of times different size of firs are audited, and (iii) a financial analyst wants to keep track of the number of times the shares of manufacturing and service companies to be or gain order a period of time.

Advantages and Disadvantages of Frequency Distribution

Advantages The following are a few advantages of grouping and summarizing raw data in this compact form:

(i) The data are expressed in a more compact form. One can get a deeper insight into the salient characteristics of the data at the very first glance.
(ii) One can quickly note the pattern of distribution of observations falling in various classes.
(iii) It permits the use of more complex statistical techniques which help reveal certain other obscure and hidden characteristics of the data.

Disadvantages A frequency distribution suffers from some disadvantages as stated below:

(i) In the process of grouping, individual observations lose their identity. It becomes difficult to notice how the observations contained in each class are distributed. This applies more to a frequency distribution which uses the tally method in its construction.
(ii) A serious limitation inherent in this kind of grouping is that there will be too much clustering of observations in various classes in case the number of classes is too small. This will cause some of the essential information to remain unexposed.

Hence, it is important that summarizing data should not be at the cost of losing essential details. The purpose should be to seek an appropriate compromise between having too much of details or too little. To be able to achieve this compromise, certain criteria are discussed for constructing a frequency distribution.

The frequency distribution of the number of hours of overtime given in Table 2.2 is shown in Table 2.4.

Table 2.4 Array and Tallies

Number of Overtime Hours	*Tally*	*Number of Weeks (Frequency)*
84	||	2
85	||	2
86	—	0
87	|	1
88	||||	4
89	|||	3
90	||	2
91	||	2
92	||	2
93	𝍸	6
94	𝍸|	5
95	|	1
		30

Constructing a Frequency Distribution As the number of observations obtained gets large, the method discussed above to condense the data becomes quite difficult and time-consuming. Thus to further condense the data into frequency distribution tables, the following steps should be taken:

(i) Select an appropriate number of non-overlapping class intervals
(ii) Determine the width of the class intervals
(iii) Determine class limits (or boundaries) for each class interval to avoid overlapping.

1. *Decide the number of class intervals* The decision on the number of class groupings depends largely on the judgment of the individual investigator and/or the range that will be used to group the data, although there are certain guidelines that can be used. As a general rule, a frequency distribution should have at least five class intervals (groups), but not more than fifteen. The following two rules are often used to decide approximate number of classes in a frequency distribution:

(i) If k represents the number of classes and N the total number of observations, then the value of k will be the smallest exponent of the number 2, so that $2^k \geq N$.

In Table 2.3 we have N = 30 observations. If we apply this rule, then we shall have

$$2^3 = 8 \ (< 30)$$
$$2^4 = 16 \ (< 30)$$
$$2^5 = 32 \ (> 30)$$

Thus we may choose $k = 5$ as the number of classes.

(ii) According to Sturge's rule, the number of classes can be determined by the formula

$$k = 1 + 3.222 \log_e N$$

where k is the number of classes and $\log_e N$ is the logarithm of the total number of observations.

Applying this rule to the data given in Table 2.3, we get

$$k = 1 + 3.222 \log 30$$
$$= 1 + 3.222 \ (1.4771) = 5.759 \cong 5$$

2. *Determine the width of class intervals* When constructing the frequency distribution it is desirable that the width of each class interval should be equal in size. The size (or width) of each class interval can be determined by first taking the difference between the largest and smallest numerical values in the data set and then dividing it by the number of class intervals desired.

$$\text{Width of class interval } (h) = \frac{\text{Largest numerical value} - \text{Smallest numerical value}}{\text{Number of classes desired}}$$

The value obtained from this formula can be rounded off to a more convenient value based on the investigator's preference.

From the ordered array in Table 2.3, the range is: 95 – 84 = 11 hours. Using the above formula with 5 classes desired, the width of the class intervals is approximated as:

$$\text{Width of class interval} = \frac{11}{5} = 2.2 \text{ hours}$$

For convenience, the selected width (or interval) of each class is rounded to 3 hours.

3. Determine class limits (Boundaries) The limits of each class interval should be clearly defined so that each observation (element) of the data set belongs to one and only one class.

Each class has two limits—*a lower limit* and an *upper limit*. The usual practice is to let the lower limit of the first class be a convenient number slightly below or equal to the lowest value in the data set. In Table 2.3, we may take the lower class limit of the first class as 82 and the upper class limit as 85. Thus the class would be written as 82–85. This class interval includes all overtime hours ranging from 82 upto but not including 85 hours. The various other classes can be written as:

Overtime Hours (Class intervals)	*Tallies*	*Frequency*
82 but less than 85	||	2
85 but less than 88	|||	3
88 but less than 91	𝍸 ||||	9
91 but less than 94	𝍸 𝍸	10
94 but less than 97	𝍸 |	6
		30

Mid-point of Class Intervals The main advantage of using the above summary table is that the major data characteristics become clear to the decision-maker. However, it is difficult to know how the individual values are distributed within a particular class interval without access to the original data. The **class midpoint** is the point halfway between the boundaries (both upper and lower class limits) of each class and is representative of all the observations contained in that class.

Class midpoint: The point in each class that is halfway between the lower and upper class limits.

Arriving at the correct class mid-points is important, for these are used as representative of all the observations contained in their respective class while computing many important statistical measures. A mid-point is obtained by dividing the sum of the upper and lower class limits by two. Problems in computing mid-points arise when the class limits are ambiguous and not clearly defined.

The width of the class interval should, as far as possible, be equal for all the classes. If this is not possible to maintain, the interpretation of the distribution becomes difficult. For example, it will be difficult to say whether the difference between the frequencies of the two classes is due to difference in the concentration of observations in the two classes or due to the width of the class intervals being different.

Further, to facilitate computation of the summary measures discussed in Chapter 3, the width of the class intervals should preferably be not only the same throughout, it should also be a convenient number such as 5, 10, or 15. A width given by integers 7, 13, or 19 should be avoided.

2.3.2 Methods of Data Classification

There are two ways in which observations in the data set are classified on the basis of class intervals, namely

(i) Exclusive method, and
(ii) Inclusive method

Exclusive Method When the data are classified in such a way that the upper limit of a class interval is the lower limit of the succeeding class interval (i.e. no data point falls into more than one class interval), then it is said to be the exclusive method of classifying data. This method is illustrated in Table 2.5.

Table 2.5 Exclusive Method of Data Classification

Dividend Declared in per cent (Class Intervals)	Number of Companies (Frequencies)
0–10	5
10–20	7
20–30	15
30–40	10

Such classification ensures continuity of data because the upper limit of one class is the lower limit of succeeding class. As shown in Table 2.5, 5 companies declared dividend ranging from 0 to 10 per cent, this means a company which declared exactly 10 per cent dividend would not be included in the class 0–10 but would be included in the next class 10–20. Since this point is not always clear, therefore to avoid confusion data are displayed in a slightly different manner, as given in Table 2.6.

Table 2.6

Dividend Declared in per cent (Class Intervals)	Number of Companies (Frequencies)
0 but less than 10	5
10 but less than 20	7
20 but less than 30	15
30 but less than 40	10

Inclusive Method When the data are classified in such a way that both lower and upper limits of a class interval are included in the interval itself, then it is said to be the inclusive method of classifying data. This method is shown in Table 2.7.

Table 2.7 Inclusive Method of Data Classification

Number of Accidents (Class Intervals)	Number of Weeks (Frequencies)
0– 4	5
5– 9	22
10–14	13
15–19	8
20– 24	2

Remarks: 1. An exclusive method should be used to classify a set of data involving continuous variables and an inclusive method should be used to classify a set of data involving discrete variables.

2. If a continuous variable is classified according to the inclusive method, then certain adjustment in the class interval is needed to obtain continuity as shown in Table 2.8.

Table 2.8

Class Interval	Frequency
30 – 44	28
45 – 59	32
60 – 74	45
75 – 89	50
90 – 104	35

To ensure continuity, first calculate correction factor as:

$$x = \frac{\text{Upper limit of a class} - \text{Lower limit of the next higher class}}{2}$$

and then subtract it from the lower limits of all the classes and add it to the upper limits of all the classes.

From Table 2.8, we have $x = (45 - 44) \div 2 = 0.5$. Subtract 0.5 from the lower limits of all the classes and add 0.5 to the upper limits. The adjusted classes would then be as shown in Table 2.9.

Table 2.9

Class Intervals	Frequency
29.5 – 44.5	28
44.5 – 59.5	32
59.5 – 74.5	45
74.5 – 89.5	50
89.5 – 104.5	35

3. Class intervals should be of equal size to make a meaningful comparison between classes. In a few cases, extreme values in the data set may require the inclusion of *open-end classes* and this distribution is known as an *open-end distribution.* Such open-end classes do not pose any problem in data analysis as long as only a few frequencies (or values) lie in these classes. However, an open-end distribution is not fit for further mathematical calculations because *mid-value* which is used to represent the class, cannot be determined for an open-end class. An example of an open-end distribution is given in Table 2.10.

Table 2.10

Age (Years)	Population (Millions)
Under 5	17.8
5 – 17	44.7
18 – 24	29.9
25 – 44	69.6
45 – 64	44.6
65 and above	27.4
	234.0

Table 2.11 provides a tentative guide to determine an adequate number of classes.

Table 2.11 Guide to Determine the Number of Classes to Use

Number of Observations, N	Suggested Number of Classes
20	5
50	7
100	8
200	9
500	10
1000	11

Example 2.1: The following set of numbers represents mutual fund prices reported at the end of a week for selected 40 nationally sold funds.

10	17	15	22	11	16	19	24	29	18
25	26	32	14	17	20	23	27	30	12
15	18	24	36	18	15	21	28	33	38
34	13	10	16	20	22	29	29	23	31

Arrange these prices into a frequency distribution having a suitable number of classes.

Solution: Since the number of observations are 40, it seems reasonable to choose 6 ($2^6 > 40$) class intervals to summarize values in the data set. Again, since the smallest value is 10 and the largest is 38, therefore the class interval is given by

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{38-10}{6} = \frac{28}{6} = 4.66 \simeq 5$$

Now performing the actual tally and counting the number of values in each class, we get the frequency distribution by exclusive method as shown in Table 2.12:

Table 2.12: Frequency Distribution

Class Interval (Mutual Fund Prices, Rs)	*Tally*	*Frequency (Number of Mutual Funds)*
10 – 15	𝍸 |	6
15 – 20	𝍸 𝍸 |	11
20 – 25	𝍸 ||||	9
25 – 30	𝍸 ||	7
30 – 35	𝍸	5
35 – 40	||	2
		40

Example 2.2: The take-home salary (in Rs) of 40 unskilled workers from a company for a particular month was.

2482	2392	2499	2412	2440	2444
2446	2540	2394	2365	2412	2458
2482	2394	2450	2444	2440	2494
2460	2425	2500	2390	2414	2365
2390	2460	2422	2500	2470	2428

Construct a frequency distribution having a suitable number of classes.

Solution: Since the number of observations are 30, we choose 5($2^5 > 30$) class intervals to summarize values in the data set. In the data set the smallest value is 2365 and the largest is 2500, so the width of each class interval will be

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{2540 - 2365}{5} = \frac{175}{5} = 35$$

Sorting the data values into classes and counting the number of values in each class, we get the frequency distribution by exclusive method as

Table 2.13 Frequency Distribution

Class Interval (Salary, Rs)	*Tally*	*Frequency (Number of Workers)*
2365 – 2400	𝍸 |	6
2400 – 2435	𝍸 ||	7
2435 – 2470	𝍸 𝍸	10
2470 – 2505	𝍸 |	6
2505 – 2540	|	1
		30

Example 2.3: A computer company received a rush order for as many home computers as could be shipped during a six-week period. Company records provide the following daily shipments:

22	65	65	67	55	50	65
77	73	30	62	54	48	65
79	60	63	45	51	68	79
83	33	41	49	28	55	61
65	75	55	75	39	87	45
50	66	65	59	25	35	53

Group these daily shipments figures into a frequency distribution having the suitable number of classes.

Solution: Since the number of observations are 42, it seems reasonable to choose 6 ($2^6 > 42$) classes. Again, since the smallest value is 22 and the largest is 87, therefore the class interval is given by

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{87 - 22}{6} = \frac{65}{6} = 10.833 \text{ or } 11$$

Now performing the actual tally and counting the number of values in each class, we get the following frequency distribution by inclusive method as shown in Table 2.14.

Table 2.14 Frequency Distribution

Class Interval (Number of Computers)	*Tally*	*Frequency (Number of Days)*
22 – 32	||||	4
33 – 43	||||	4
44 – 54	𝍸 ||||	9
55 – 65	𝍸 𝍸 ||||	14
66 – 76	𝍸 |	6
77 – 87	𝍸	5
		42

Example 2.4: Following is the increase of D.A. in the salaries of employees of a firm at the following rates.

Rs 250 for the salary range up to Rs 4749
Rs 260 for the salary range from Rs 4750
Rs 270 for the salary range from Rs 4950
Rs 280 for the salary range from Rs 5150
Rs 290 for the salary range from Rs 5350

No increase of D.A for salary of Rs 5500 or more. What will be the additional amount required to be paid by the firm in a year which has 32 employees with the following salaries (in Rs)?

5422	4714	5182	5342	4835	4719	5234	5035
5085	5482	4673	5335	4888	4769	5092	4735
5542	5058	4730	4930	4978	4822	4686	4730
5429	5545	5345	5250	5375	5542	5585	4749

Solution: Performing the actual tally and counting the number of employees in each salary range (or class), we get the following frequency distribution as shown in Table 2.15.

Table 2.15 Frequency Distribution

Class Interval (Pay Range)	*Tally*	*Frequency, f (Number of Employees)*	*Rate of D.A. (Rs x)*	*Total Amount Paid (Rs f x)*
upto 4749	𝍸 | | |	8	250	2000
4750 – 4949	𝍸	5	260	1300
4950 – 5149	𝍸	5	270	1350
5150 – 5349	𝍸 |	6	280	1680
5350 – 5549	𝍸 | | |	8	290	2320
		32		8650

Hence additional amount required by the firm for payment of D.A. is Rs 8650.

Example 2.5: Following are the number of items of similar type produced in a factory during the last 50 days

21	22	17	23	27	15	16	22	15	23
24	25	36	19	14	21	24	25	14	18
20	31	22	19	18	20	21	20	36	18
21	20	31	22	19	18	20	20	24	35
25	26	19	32	22	26	25	26	27	22

Arrange these observations into a frequency distribution with both inclusive and exclusive class intervals choosing a suitable number of classes.

Solution: Since the number of observations are 50, it seems reasonable to choose 6 ($2^6 > 50$) or less classes. Since smallest value is 14, and the largest is 36 therefore the class interval is given by

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{36 - 14}{6} = \frac{22}{6} = 3.66 \text{ or } 4$$

Performing the actual tally and counting the number of observations is each class, we get the following frequency distribution with inclusive class intervals as shown in Table 2.16.

Table 2.16 Frequency Distribution with Inclusive Class Intervals

Class Intervals	*Tally*	*Frequency (Number of Items Produced)*
14 – 17	𝍸 |	6
18 – 21	𝍸 𝍸 𝍸 | | |	18
22 – 25	𝍸 𝍸 𝍸	15
26 – 29	𝍸	5
30 – 33	| | |	3
34 – 33	| | |	3
		50

Converting the class intervals shown in Table 2.16 into exclusive class intervals is shown in Table 2.17.

Table 2.17 Frequency Distribution with Exclusive Class Intervals

Class Intervals	*Mid-Value of Class Intervals*	*Frequency (Number of Items Produced)*
13.5 – 17.5	15.5	6
17.5 – 21.5	19.5	18
21.5 – 25.5	23.5	15
25.5 – 29.5	27.5	5
29.5 – 33.5	31.5	3
33.5 – 37.5	34.5	3

2.3.3 Bivariate Frequency Distribution

The frequency distributions discussed so far involved only one variable and therefore called *univariate frequency distributions*. In case the data involve two variables (such as profit and expenditure on advertisements of a group of companies, income and expenditure of a group of individuals, supply and demand of a commodity, etc.), then frequency distribution so obtained as a result of cross classification is called *bivariate frequency distribution*. It can be summarized in the form of a *two-way (bivariate) frequency table* and the values of each variable are grouped into various classes (not necessarily same for each variable) in the same way as for univariate distributions.

If the data corresponding to one variable, say x, is grouped into m classes and the data corresponding to another variable, say y, is grouped into n classes, then bivariate frequency table will have $m \times n$ cells.

Frequency distribution of variable x for a given value of y is obtained by the values of x and vice-versa. Such frequencies in each cell are called *conditional frequencies*. The frequencies of the values of variables x and y together with their frequency totals are called the *marginal frequencies*.

Example 2.6: The following figures indicate income (x) and percentage expenditure on food (y) of 25 families. Construct a bivariate frequency table classifying x into intervals 200 – 300, 300 – 400, . . . and y into 10 – 15, 15 – 20, . . .

Write the marginal distribution of x and y and the conditional distribution of x when y lies between 15 and 20.

x	y	x	y	x	y	x	y	x	y
550	12	225	25	680	13	202	29	689	11
623	14	310	26	300	25	255	27	523	12
310	18	640	20	425	16	492	18	317	18
420	16	512	18	555	15	587	21	384	17
600	15	690	12	325	23	643	19	400	19

Solution: The two-way frequency table showing income (in Rs) and percentage expenditure on food is shown in Table 2.18.

Table 2.18

Expenditure (y) (Percentage)	*Income (x)* 200–300	300–400	400–500	500–600	600–700	*Marginal Frequencies, f_y*
10 – 15				|| (2)	|||| (4)	6
15 – 20		||| (3)	|||| (4)	|| (2)	|| (2)	11
20 – 25		| (1)		| (1)	| (1)	3
25 – 30	||| (3)	|| (2)				5
Marginal Frequencies, f_x	3	6	4	5	7	25

The conditional distribution of x when y lies between 15 and 20 per cent is as follows:

Income (x) :	200–300	300–400	400–500	500–600	600–700
15%–20% :	0	3	4	2	2

Example 2.7: The following data give the points scored in a tennis match by two players X and Y at the end of twenty games:

(10, 12) (7, 11) (7, 9) (15, 19) (17, 21) (12, 8) (16, 10) (14, 14) (22, 18) (16, 7)
(15, 16) (22, 20) (19, 15) (7, 18) (11, 11) (12, 18) (10, 10) (5, 13) (11, 7) (10, 10)

Taking class intervals as: 5–9, 10–14, 15–19 . . ., for both X and Y, construct

(i) Bivariate frequency table.

(ii) Conditional frequency distribution for Y given X > 15.

Solution: (i) The two-way frequency distribution is shown in Table 2.19.

Table 2.19 Bivariate Frequency Table

Player Y	*Player X* 5–9	10–14	15–19	20–24	*Marginal Frequencies, f_y*
5– 9	| (1)	| | (2)	| (1)	—	4
10–14	| | (2)	𝍸 (5)	| (1)	—	8
15–19	| (1)	| (1)	| | | (3)	| (1)	6
20–24	—	—	| (1)	| (1)	2
Marginal Frequencies, f_x	4	8	6	2	20

(ii) Conditional frequency distribution for Y given X > 15.

Player Y	*Player X* 15–19	20–24
5–9	1	—
10–14	1	—
15–19	3	1
20–24	1	1
	6	2

2.3.4 Types of Frequency Distributions

Cumulative frequency distribution: The cumulative number of observations less than or equal to the upper class limit of each class.

Cumulative Frequency Distribution Sometimes it is preferable to present data in a **cumulative frequency (cf) distribution** or simply a distribution which shows the cumulative number of observations below the upper boundary (limit) of each class in the given frequency distribution. A cumulative frequency distribution is of two types: (i) *more than* type and (ii) *less than* type.

In a *less than* cumulative frequency distribution, the frequencies of each class interval are added successively from top to bottom and represent the cumulative number of observations less than or equal to the class frequency to which it relates. But in the *more than* cumulative frequency distribution, the frequencies of each class interval are added successively from bottom to top and represent the cumulative number of observations greater than or equal to the class frequency to which it relates.

The frequency distribution given in Table 2.20 illustrates the concept of cumulative frequency distribution:

Table 2.20 Cumulative Frequency Distribution

Number of Accidents	*Number of Weeks (Frequency)*	*Cumulative Frequency (less than)*	*Cumulative Frequency (more than)*
0– 4	5	5	45 + 5 = 50
5– 9	22	5 + 22 = 27	23 + 22 = 45
10–14	13	27 + 13 = 40	10 + 13 = 23
15–19	8	40 + 8 = 48	2 + 8 = 10
20–24	2	48 + 2 = 50	2

From Table 2.20 it may be noted that cumulative frequencies are corresponding to the lower limit and upper limit of class intervals. The 'less than' cumulative frequencies are corresponding to the upper limit of class intervals and 'more than' cumulative frequencies are corresponding to the lower limit of class intervals shown in Table 2.21(a) and (b).

Table 2.21(a)

Upper Limits	*Cumulative Frequency (less than)*
less than 4	5
less than 9	27
less than 14	40
less than 19	48
less than 24	50

Table 2.21(b)

Lower Limits	*Cumulative Frequency (more than)*
0 andmore	50
5 and more	45
10 and more	23
15 and more	10
20 and more	2

Relative Frequency Distribution To enrich data analysis it is sometimes important to show what percentage of observations fall within each class of a distribution instead of showing the actual class frequencies. To convert a frequency distribution into a corresponding **relative frequency distribution**, we divide each class frequency by the total number of observations in the entire distribution. Each relative frequency is thus a proportion as shown in Table 2.22.

Cumulative relative frequency distribution: The cumulative number of observations less than or equal to the upper class limit of each class.

Table 2.22 Relative and Percentage Frequency Distributions

Number of Accidents	*Number of Weeks (Frequency)*	*Relative Frequency*	*Percentage Frequency*
0 – 4	5	$\frac{5}{50} = 0.10$	$\frac{5}{50} \times 100 = 10$
5 – 9	22	$\frac{22}{50} = 0.44$	$\frac{22}{50} \times 100 = 44$
10 – 14	13	$\frac{13}{50} = 0.26$	$\frac{13}{50} \times 100 = 26$
15 – 19	8	$\frac{8}{50} = 0.16$	$\frac{8}{50} \times 100 = 16$
20 – 24	2	$\frac{2}{50} = 0.04$	$\frac{2}{50} \times 100 = 4$
	50	1.00	100

Percentage Frequency Distribution A percentage frequency distribution is one in which the number of observations for each class interval is converted into a percentage frequency by dividing it by the total number of observations in the entire distribution. The quotient so obtained is than multiplied by 100, as shown in Table 2.22.

Cumulative percent frequency distribution: The cumulative percentage of observations less than or equal to the upper class limit of each class.

Example 2.8: Following are the number of two wheelers sold by a dealer during eight weeks of six working days each.

13	19	22	14	13	16	19	21
23	11	27	25	17	17	13	20
23	17	26	20	24	15	20	21
23	17	29	17	19	14	20	20
10	22	18	25	16	23	19	20
21	17	18	24	21	20	19	26

(a) Group these figures into a table having the classes 10–12, 13–15, 16–18, . . ., and 28–30.

(b) Convert the distribution of part (i) into a corresponding percentage frequency distribution and also a percentage cumulative frequency distribution.

Solution: (a) Frequency distribution of the given data is shown in Table 2.23.

Table 2.23 Frequency Distribution

Number of Automobiles Sold (Class Intervals)	*Tally*	*Number of Days (Frequency)*
10 – 12	||	2
13 – 15	𝍸 |	6
16 – 18	𝍸 𝍸	10
19 – 21	𝍸 𝍸 𝍸 |	16
22 – 24	𝍸 |||	8
25 – 27	𝍸	5
28 – 30	|	1
		48

(ii) **Table 2.24: Percentage and More Than Cumulative Percentage Distribution**

Number of Automobiles Sold (Class Intervals)	*Number of Days (Frequency)*	*Cumulative Frequency*	*Percentage Frequency*	*Percentage Cumulative Frequency*
10 – 12	2	2	4.17	4.17
13 – 15	6	8	12.50	16.67
16 – 18	10	18	20.83	37.50
19 – 21	16	34	33.34	70.84
22 – 24	8	42	16.67	87.51
25 – 27	5	47	10.41	97.92
28 – 30	1	48	2.08	100.00
	48		100.00	

Conceptual Questions 2A

1. Explain the characteristics of a frequency distribution.
2. Illustrate two methods of classifying data in class-intervals.
3. Distinguish clearly between a continuous variable and a discrete variable. Give two examples of continuous variables and two examples of discrete variables that might be used by a statistician.
4. State whether the statement is true of false: The heights of rectangles erected on class intervals are proportional to the cumulative frequency of the class.
5. What is the basic property of virtually all data that lead to methods of describing and analysing data? How is a frequency distribution related to this property of data?
6. What are the advantages of using a frequency distribution to describe a body of raw data? What are the disadvantages?
7. When constructing a grouped frequency distribution, should equal intervals always be used? Under what circumstances should unequal intervals be used instead?
8. What are the advantages and disadvantages of using open-end intervals when constructing a group frequency distribution?
9. When constructing a group frequency distribution, is it necessary that the resulting distribution be symmetric? Explain.
10. Why is it necessary to summarize data? Explain the approaches available to summarize data distributions.
11. What are the objections to unequal class and open class intervals? State the conditions under which the use of unequal class intervals and open class intervals are desirable and necessary.
12. (a) What do you understand by cumulative frequency distribution?
    (b) What do you understand by bivariate or two-way frequency distribution?

Self-Practice Problems 2A

2.1 A portfolio contains 51 stocks whose prices are given below:

67	34	36	48	49	31	61	34
43	45	38	32	27	61	29	47
36	50	46	30	40	32	30	33
45	49	48	41	53	36	37	47
47	30	50	28	35	35	38	36
46	43	34	62	69	50	28	44
43	60	39					

Summarize these stock prices in the from of a frequency distribution.

2.2 Construct a frequency distribution of the data given below, where class interval is 4 and the mid-value of one of the classes is zero.

−8	−7	10	12	6	4	3	0	7
−4	−3	−2	2	3	4	7	5	6
10	12	9	13	11	−10	−7	1	0
5	3	2	6	10	−6	−4		

2.3 Form a frequency distribution of the following data. Use an equal class interval of 4 where the lower limit of the first class is 10.

10	17	15	22	11	16	19	24	29
18	25	26	32	14	17	20	23	27
30	12	15	18	24	36	18	15	21
28	33	38	34	13	10	16	20	22
29	29	23	31					

2.4 If class midpoints in a frequency distribution of the ages of a group of persons are: 25, 32, 39,. 46, 53, and 60, find:

(a) the size of the class-interval

(b) the class boundaries

(c) the class limits, assuming that the age quoted is the age completed on the last birthdays

2.5 The distribution of ages of 500 readers of a nationally distributed magazine is given below:

Age (in Years)	*Number of Readers*
Below 14	20
15–19	125
20–24	25
25–29	35
30–34	80
35–39	140
40–44	30
45 and above	45

Find the relative and cumulative frequency distributions for this distribution.

2.6 The distribution of inventory to sales ratio of 200 retail outlets is given below:

Inventory to Sales Ratio	*Number of Retail Outlets*
1.0–1.2	20
1.2–1.4	30
1.4–1.6	60
1.6–1.8	40
1.8–2.0	30
2.0–2.2	15
2.2–2.4	5

Find the relative and cumulative frequency distributions for this distribution.

2.7 A wholesaler's daily shipments of a particular item varied from 1,152 to 9,888 units per day. Indicate the limits of nine classes into which these shipments might be grouped.

2.8 A college book store groups the monetary value of its sales into a frequency distribution with the classes, Rs 400–500, Rs 501–600, and Rs 601 and over, Is it possible to determine from this distribution the amount of sales

(a) less than Rs 601 (b) less than Rs 501

(c) Rs 501 or more?

2.9 The class marks of distribution of the number of electric light bulbs replaced daily in an office building are 5, 10, 15, and 20. Find (a) the class boundaries and (b) class limits.

2.10 The marks obtained by 25 students in Statistics and Economics are given below. The first figure in the bracket indicates the marks in Statistics and the second in Economics.

(14, 12)	(0, 2)	(1, 5)	(7, 3)	(15, 9)
(2, 8)	(12, 18)	(9, 11)	(5, 3)	(17, 13)
(19, 18)	(11, 7)	(10, 13)	(13, 16)	(16, 14)
(6, 10)	(4, 1)	(9, 15)	(11, 14)	(8, 3)
(13, 11)	(14, 17)	(10, 10)	(11, 7)	(15, 15)

Prepare a two-way frequency table taking the width of each class interval as 4 marks, the first being less than 4.

2.11 Prepare a bivariate frequency distribution for the following data for 20 students:

Marks in Law:	10	11	10	11	11
	14	12	12	13	10
Marks in Statistics:	20	21	22	21	23
	23	22	21	24	23
Marks in Law:	13	12	11	12	10
	14	14	12	13	10
Marks in Statistics:	24	23	22	23	22
	22	24	20	24	23

Also prepare

(a) A marginal frequency table for marks in law and statistics

(b) A conditional frequency distribution for marks in law when the marks in statistics are more than 22.

2.12 Classify the following data by taking class intervals such that their mid-values are 17, 22, 27, 32, and so on:

30	42	30	54	40	48	15	17	51
42	25	41	30	27	42	36	28	26
37	54	44	31	36	40	36	22	30
31	19	48	16	42	32	21	22	46
33	41	21						

[Madurai-Kamraj Univ., BCom, 1995]

2.13 In degree colleges of a city no teacher is less than 30 years or more than 60 years in age. Their cumulative frequencies are as follows:

Less than	:	60	55	50	45
		40	35	30	25
Total frequency	:	980	925	810	675
		535	380	220	75

Find the frequencies in the class intervals 25–30, 30–35, . . .

Hints and Answers

2.3 The classes for preparing frequency distribution by inclusive method will be

10–13, 14–17, 18–21, . . ., 34–37, 38–41

2.4 (a) Size of the class interval = Difference between the mid-values of any two consecutive classes = 7

(b) The class boundaries for different classes are obtained by adding (for upper class boundaries or limits) and subtracting (for lower class boundaries or limits) half the magnitude of the class interval that is, 7 ÷ 2 = 3.5 from the mid-values.

Class Intervals :	21.5–28.5	28.5–35.5	35.5–42.5
Mid-Values :	25	32	39
Class Intervals :	42.5–49.5	49.5–56.5	56.5–63.5
Mid-Values :	46	53	60

(c) The distribution can be expressed in inclusive class intervals with width of 7 as: 22–28, 29–35, . . ., 56–63.

2.7 One possibility is 1000–1999, 2000–2999, 3000–3999, . . . 9000–9999 units of the item.

2.13

Age (year)	*Cumulative Frequency*	*Age*	*Frequency*
Less than 25	75	20–25	75
Less than 30	220	25–30	220– 75 = 145
Less than 35	380	30–35	380–220 = 160
Less than 40	535	35–40	535–380 = 155
Less than 45	675	40–45	675–535 = 140
Less than 50	810	45–50	810–675 = 135
Less than 55	925	50–55	925–810 = 115
Less than 60	980	55–60	980–925 = 55

2.4 TABULATION OF DATA

Meaning and Definition Tabulation is another way of summarizing and presenting the given data in a systematic form in rows and columns. Such presentation facilitates comparison by bringing related information close to each other and helps in further statistical analysis and interpretation. Tabulation has been defined by two statisticians as:

- The logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and explanatory notes to make clear the full meaning, context and the origin of the data. —Tuttle

This definition gives an idea of the broad structure of statistical tables and suggests that tabulation helps organize a set of data in an orderly manner to highlight its basic characteristics.

- Tables are means of recording in permanent form the analysis that is made through classification and by placing in just a position things that are similar and should be compared. —Secrist

This definition defines tabulation as the process of classifying the data in a systematic form which facilitates comparative studies of data sets.

2.4.1 Objectives of Tabulation

The above two definitions indicate that tabulation is meant to summarize data in a simplest possible form so that the same can be easily analysed and interpreted. A few objectives of tabulation defined by few statisticians are as follows:

- Tabulation is the process of condensing classified data in the form of a table so that it may be more easily understood, and so that any comparison involved may be more readily made.

—D. Gregory and H. Ward

- It is a medium of communication of great economy and effectiveness for which ordinary prose is inadequate. In addition to its formation in simple presentation, the statistical table is also a useful tool of analysis.

—D. W. Paden and E. F. Lindquist

The major objectives of tabulation are:

1. *To simplify the complex data:* Tabulation presents the data set in a systematic and concise form avoiding unnecessary details. The idea is to reduce the bulk of information (data) under investigation into a simplified and meaningful form.
2. *To economize space:* By condensing data in a meaningful form, space is saved without sacrificing the quality and quantity of data.
3. *To depict trend:* Data condensed in the form of a table reveal the trend or pattern of data which otherwise cannot be understood in a descriptive form of presentation.
4. *To facilitate comparison:* Data presented in a tabular form, having rows and columns, facilitate quick comparison among its observations.
5. *To facilitate statistical comparison:* Tabulation is a phase between classification of data and its presentation. Various statistical techniques such as measures of average and disperson, correlation and regression, time series, and so on can be applied to analyse data and then interpretating the results.
6. *To help reference:* When data are arranged in tables in a suitable form, they can easily be identified and can also be used as reference for future needs.

2.4.2 Parts of a Table

Presenting data in a tabular form is an art. A statistical table should contain all the requisite information in a limited space but without any loss of clarity. Practice varies, but explained below are certain accepted rules for the construction of an ideal table:

1. **Table number:** A table should be numbered for easy identification and reference in future. The table number may be given either in the centre or side of the table but above the top of the title of the table. If the number of columns in a table is large, then these can also be numbered so that easy reference to these is possible.
2. **Title of the table:** Each table must have a brief, self-explanatory and complete title so that
   (a) it should be able to indicate nature of data contained.
   (b) it should be able to explain the *locality* (i.e., geographical or physical) of data covered.
   (c) it should be able to indicate the *time* (or period) of data obtained.
   (d) it should contain the *source* of the data to indicate the authority for the data, as a means of verification and as a reference. The source is always placed below the table.
3. **Caption and stubs:** The heading for columns and rows are called caption and stub, respectively. They must be clear and concise.

   Two or more columns or rows with similar headings may be grouped under a common heading to avoid repetition. Such arrangements are called sub-captions or sub-stubs. Each row and column can also be numbered for reference and to facilitate comparisons. The caption should be written at the middle of the column in small letters to save space. If different columns are expressed in different units, then the units should be specified along with the captions.

   The stubs are usually wider than column headings but must be kept narrow without sacrificing precision or clarity. When a stub occupies more than one line, the figures of the table should be written in the last line.

4. **Body:** The body of the table should contains the numerical information. The numerical information is arranged according to the descriptions given for each column and row.
5. **Prefactory or head note:** If need be, a prefactory note is given just below the title for its further description in a prominent type. It is usually enclosed in brackets and is about the unit of measurement.
6. **Foot notes:** Anything written below the table is called a footnote. It is written to further clarify either the title captions or stubs. For example if the data described in the table pertain to profits earned by a company, then the footnote may define whether it is profit before tax or after tax. There are various ways of identifying footnotes:
   (a) Numbering foot notes consecutively with small number 1, 2, 3, ... or letters a, b, c ... or star *, **, . . .
   (b) Sometimes symbols like @ or $ are also used to identify footnotes.

   A blank model table is given below:

Table Number and Title [Head or Prefactory Note (if any)]

Stub Heading	*Caption*				*Total (Rows)*
	Subhead		*Subhead*		
	Column-head	*Column-head*	*Column-head*	*Column-head*	
Stub Entries					
Total (Columns)					

Foot Note :
Source Note :

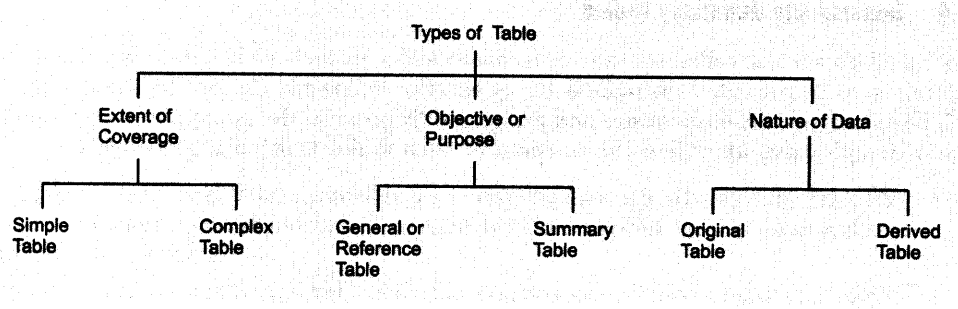
Remarks: 1. Information which is not available should be indicated by the letter N.A. or by dash (–) in the body of the table.

2. Ditto marks ("), 'etc.' and use of the abbreviated forms should be avoided in the table.
3. The requisites of a good statistical table given by various people are as flows:

- 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 stand point of 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. —J. C. Capt
- A good statistical table is not a mere careless grouping of columns and rows of figures, it is a triumph of ingenuity and technique, a master-piece of economy of space combined with a maximum of clearly presented information. To prepare a first class table, one must have a clear idea of the facts to be presented, the contrasts to be stressed, the points upon which emphasis is to be placed and lastly a familiarity with the technique of preparation. —Harry Jerome
- In collection and tabulation commonsense is the chief requisite and experience, the chief teacher. —A. L. Bowley

2.4.3 Types of Tables

The classification of tables depends on various aspects: objectives and scope of investigation, nature of data (primary or secondary) for investigation, extent of data coverage, and so on. The different types of tables used in statistical investigations are as follows:



Simple and Complex Tables In a *simple table* (also known as one-way table), data are presented based on only one characteristic. Table 2.25 illustrates the concept.

Table 2.25 Candidates Interviewed for Employment in a Company

Candidate's Profile	*Number of Candidates*
Experienced	50
Inexperienced	70
Total	120

The *complex table* also known as a manifold table is that in which data are presented according to two or more characteristics simultaneously. The complex tables are two-way or three-way tables according to whether two or three characteristics are presented simultaneously.

(a) *Double or Two-Way Table:* In such a table, the variable under study is further subdivided into two groups according to two inter-related characteristics. For example, if the total number of candidates given in Table 2.24 are further divided according to their sex, the table would become a two-way table because it would reveal information about two characteristics namely, male and female. The new shape of the table is shown in Table 2.26.

Table 2.26 Candidates Interviewed for Employment in a Company

Candidates Profile	*Number of Candidates*		*Total*
	Males	*Females*	
Experienced	35	15	50
Inexperienced	10	60	70
Total	45	75	120

(b) *Three-Way Table:* In such a table, the variable under study is divided according to three interrelated characteristics. For example, if the total number of males and females candidates given in Table 2.26 are further divided according to the marital status, the table would become a three-way. The new shape of the table is shown in Table 2.27.

Table 2.27 Candidates Interviewed for Employment in a Company

Candidates Profile	*Number of Candidates*						*Total*
	Males			*Females*			
	Married	*Unmarried*	*Total*	*Married*	*Unmarried*	*Total*	
Experienced	15	20	35	5	10	15	50
Inexperienced	2	8	10	10	50	60	70
Total	17	28	45	15	60	75	120

(c) *Manifold (or Higher Order) Table:* Such tables provide information about a large number of inter-related characteristics in the data set. For example, if the data given in Table 2.27 is also available for other companies, then table would become a manifold table.

2.4.4 General and Summary Tables

General tables are also called *reference* or *repository tables*. In such a table, data are presented in detail so as to provide information for general or reference use on the same subject. Such tables are usually large in size and are generally given in the appendix for reference. Various people have identified the purpose of such tables which are given below:

- Primary 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.
  —Croxton and Cowden
- Reference tables contain ungrouped data basic for a particular report, usually containing a large amount of data and frequently selected to a tabular appendix.
  —Horace Secrist
- These tables are those in which data are recorded not the detailed data which have been analysed but rather the results of the analysis. —John I. Griffin

Data published by various ministries, autonomous bodies, or institutions pertaining to employment, production, public expenditure, taxation, population, and so on are examples of such tables.

2.4.5 Original and Derived Tables

Original tables are also called *classification tables*. Such a table contains data collected from a primary source. But if the information given in a table has been derived from a general table, then such a table is called a *derivatived table*. For example, if from a general table, certain averages, ratios, or percentages are derived, then the table containing such information would be a derivatived table.

Example 2.9: A state government has taken up a scheme of providing drinking water to every village. During the first four years of a five-year plan, the government has installed 39,664 tubewells. Out of the funds earmarked for natural calamities the government has sunk 14,072 tubewells during the first four years of the plan. Thus, out of the plan fund 9245 and 8630 tubewells were sunk, respectively, in 2000–2001 and 2001–2002. Out of the natural calamities fund, the number of tubewells sunk in 1998–99 and 1999–2000 were 4511 and 637, respectively. The expenditure for 2000–2001 and 2001–2002 was Rs 863.41 lakh and Rs 1185.65 lakh, respectively.

The number of tubewells installed in 2002–2002 was 16,740 out of which 4800 were installed out of the natural calamities fund and the expenditure of sinking of tubewells during 2002–2003 was Rs 1411.17 lakh.

The number of tubewells installed in 2003–2004 was 13,973, out of which 9849 tubewells were sunk out of the fund for the plan and the total expenditure during the first four years was Rs 5443.05 lakh.

Represent this data in a tabular form.

Solution: The data of the problem is summarized in Table 2.28.

Table 2.28 Tubewells for Drinking Water for Villages in a State

Year	*Number of Tubewells*		*Expenditure (Rs in lakh)*
	Out of Fund Plan	*Out of Natural Calamities Fund*	
2001–2001	9245	4511	863.41
2001–2002	8630	637	1185.65
2002–2003	(16,740 – 4800) = 11,940	4800	1411.17
2003–2004	9849	(13,973 – 9849) = 4124	1982.82
Total	39,664	14,072	5,443.05

Example 2.10: In a sample study about coffee-drinking habits in two towns, the following information was received:

Town A : Females were 40 per cent. Total coffee drinkers were 45 per cent and male non-coffee drinkers were 20 per cent

Town B : Males were 55 per cent. Male non-coffee drinkers were 30 per cent and female coffee drinkers were 15 per cent.

Represent this data in a tabular form.

Solution: The given data is summarized in Table 2.29.

Table 2.29 Coffee Drinking Habit of Towns A and B (in percentage)

Attribute	*Town A*			*Town B*			
	Males	*Females*	*Total* (1)	*Males*	*Females*	*Total* (2)	*Total* (1) + (2)
Coffee drinkers	(45 – 5) = 40	(40 – 35) = 5	45	(55 – 30) = 25	15	40	85
Non-coffee drinkers	20	(55 – 20) = 35	(100 – 45) = 55	30	(60 – 30) = 30	(100 – 40) = 60	115
Total	(100 – 40) = 60	40	100	55	(100 – 55) = 45	100	200

Example 2.11: Industrial finance in India has showed great variation in respect of sources of funds during the first, second, and third five-year plans. There were two main sources—internal and external. The internal sources of funds are—depreciation, free reserves and surplus. The external sources of funds are—capital issues, borrowings.

During the first plan, internal and external sources accounted for 62 per cent and 38 per cent of the total, and of the depreciation, fresh capital, and other sources formed 29 per cent, 7 per cent, and 10.6 per cent respectively.

During the second plan, internal sources decreased by 17.3 per cent compared to the first plan, and depreciation was 24.5 per cent. The external finance during the same period consisted of 10.9 per cent fresh capital and 28.9 per cent borrowings.

Compared to the second plan, external finance during the third plan decreased by 4.4 per cent, and borrowings and 'other sources' were 29.4 per cent and 14.9 per cent respectively. During the third plan, internal finance increased by 4.4 per cent and free reserves and surplus formed 18.6 per cent.

Tabulate this information with the above details as clearly as possible observing the rules of tabulation.

Solution: The given information is summarized in Table 2.30.

Table 2.30 Pattern of Industrial Finance (in Percentage)

Five Year Plan	*Sources of Funds*						
	Internal			*External*			
	Depre-ciation	*Free Reserves and Surplus*	*Total*	*Capital Issues*	*Borrowings*	*Other Sources*	*Total*
First	29	62 – 29 = 33	62	7	38 – 7 – 10.6 = 20.4	10.6	38
Second	24.5	44.7 – 24.5 = 20.2	62 – 17.3 = 44.7	10.9	28.9	55.3 – 10.9 – 28.9 = 15.5	100 – 44.7 = 55.3
Third	49.1 – 18.6 = 30.5	18.6	44.7 + 4.4 = 49.1	50.9 – 29.4 – 14.9 = 6.6	29.4	14.9	55.3 – 4.4 = 50.9

Example 2.12: The following information about weather conditions at different stations were recorded at 8.30 a.m. on Thursday, 29 August 1990.

At Ahmednagar station, the maximum and minimum temperature in 24 hrs were 28°C and 20°C respectively. The rainfall in the past 24 hrs at Ahmednagar was nil. Since 1 June the rainfall was 185 mm which is 105 mm below normal.

Bangalore's minimum and maximum temperatures in 24 hrs for the day were 19°C and 23°C respectively. It had no rainfall in the past 24 hrs and since 1 June the rainfall was 252 mm which is 54 mm below normal.

The minimum temperature at Udaipur was 21°C and the rainfall in the past 24 hrs was nil. Since 1 June it experienced 434 mm of rainfall which is 24 mm below normal.

Panagarh's maximum temperature in 24 hrs was 28°C. It had 4 mm of rain in the past 24 hrs and since 1 June it had 955 mm of rain.

Kolkata's maximum and minimum day temperatures were 30°C and 26°C respectively. It had 3 mm of rainfall in the past 24 hrs. Since 1 June it experienced a rainfall of 1079 mm which is 154 mm above rainfall.

Present the above data in a tabular form.

Solution: The given information is summarized in Table 2.31.

Table 2.31 Weather Conditions at Different Stations (at 8.30 a.m. on 29 August 1990)

Stations	*In 24 Hours*		*Rainfall (mm)*			
	Temperature (°C)		*Past 24 hrs*	*Since June 1*	*Above Normal*	*Below Normal*
	Min.	*Max.*				
Ahmednagar	20	28	0	185	—	105
Bangalore	19	23	0	252	—	54
Udaipur	21	—	0	434	—	24
Panagarh	—	28	4	955	—	—
Kolkata	26	30	3	1079	154	—

Example 2.13: Present the following data in a tabular form:

A certain manufacturer produces three different products 1, 2, and 3. Product 1 can be manufactured in one of the three plants A, B, or C. However, product 2 can be manufactured in either plant B or C, whereas plant A or B can manufacture product 3. Plant A can manufacture in an hour 10 pieces of 1 or 20 pieces of 3, 20 pieces of 2, 15 pieces of 1, or 16 pieces of 3 can be manufactured per hour in plant B. C can produce 20 pieces of 1 or 18 pieces of 2 per hour.

Wage rates per hour are Rs 20 at A, Rs 40 at B and Rs 25 at C. The costs of running plants A, B, and C are respectively Rs 1000, 500, and 1250 per hour. Materials and other costs directly related to the production of one piece of the product are respectively Rs 10 for 1, Rs 12 for 2, and Rs 15 for 3. The company plans to market product 1 at Rs 15 per piece, product 2 at Rs 18 per piece and product 3 at Rs 20 per piece.

Solution: The given information is summarized in Table 2.32.

Table 2.32 Production Schedule of a Manufacturer

Plant	*Rate of Manufacturing Per Hour (Pieces)*			*Wage Rates Per Hour (Rs)*	*Cost of Running Plant Per Hour (Rs)*
	1	2	3		
A	10	—	20	20	1000
B	15	20	16	40	500
C	20	18	—	25	1250
Material and other direct costs per piece (Rs)	10	12	15		
Product price per piece (Rs)	15	18	20		