Dr. C.V. Raman University
Kargi Road, Kota, BILASPUR, (C. G.),
Ph. : +07753-253801, +07753-253872
E-mail : info@cvru.ac.in | Website : www.cvru.ac.in

Business Statistics

Institute of **Open** and **Distance Education**

Faculty of Management

Business Statistics

3BBA5

DR. C.V. RAMAN UNIVERSITY
Chhattisgarh, Bilaspur
A STATUTORY UNIVERSITY UNDER SECTION 2(F) OF THE UGC ACT

3BBA5

Business Statistics

3BBA5
Business Statistics

Credit - 4

Subject Expert Team

Dr. Vivek Bajpai,
Dr. C.V. Raman University, Kota, Bilaspur, Chhattisgarh

Dr. Rajeev H. Peters,
Dr. C.V. Raman University, Kota, Bilaspur, Chhattisgarh

Dr. Niket Shukla,
Dr. C.V. Raman University, Kota, Bilaspur, Chhattisgarh

Dr. Satish Sahu,
Dr. C.V. Raman University, Kota, Bilaspur, Chhattisgarh

Dr. Archana Agrawal,
Dr. C.V. Raman University, Kota, Bilaspur, Chhattisgarh

Dr. Vikas Kumar Tiwari,
Dr. C.V. Raman University, Kota, Bilaspur, Chhattisgarh

Course Editor:

- **Dr. Jiny Jacob, Assistant Professor**
  Rabindranath Tagore University, Bhopal, Madhya Pradesh

Unit Written By:

1. **Dr. Niket Shukla**
   Professor, Dr. C. V. Raman University
2. **Dr. Priyank Mishra**
   AssociateProfessor, Dr. C. V. Raman University
3. **Dr. Anshul Shrivastava**
   Assistant Professor, Dr. C. V. Raman University

Published by: Dr. C.V. Raman University Kargi Road, Kota, Bilaspur, (C. G.), Ph. +07753-253801,07753-253872 E-mail: inf o@cvru.ac.in Website: www.cvru.ac.in.

CONTENTS

BLOCK - 1

UNIT - I

AN INTRODUCTION TO STATISTICS

CONTENTS

1.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Explain the meaning, definitions and characteristics of statistics
- Discuss the importance and scope of statistics
- Recognise the limitations, development and classification of statistics
- Identify the role of statistics in decision-making and research
- Explain the laws of statistics

1.1 INTRODUCTION

Modern age is the age of science which requires that every aspect, whether it pertains to natural phenomena, politics, economics or any other field, should be expressed in an unambiguous and precise form. A phenomenon expressed in ambiguous and vague terms might be difficult to understand in proper perspective. Therefore, in order to provide an accurate and precise explanation of a phenomenon or a situation, figures are often used. The statement that prices in a country are increasing conveys only incomplete information about the nature of the problem. However, if the figures of prices of various years are also provided, we are in a better position to understand the nature of the problem. In addition to this, these figures can also be used to compare the extent of price changes in a country vis-a-vis the changes in prices of some other country. Using these figures, it might be possible to estimate the possible level of prices at some future date so that some policy measures can be suggested to tackle the problem. The subject which deals with such type of figures, called data, is known as Statistics.

Information derived from good statistical analysis is always precise and never useless. One of the primary tasks of a manager is decision-making. Decision-making is usually based on the past experience and future projections. In many situations, decision-making purely based on personal experience, subjective judgment and intuition, is rather difficult and inefficient. Statistical techniques offer powerful tools in the decision-making process. These tools have power to interpret quantitative information in a scientific and an objective manner. These tools also provide certain conceptual framework to the decision maker and enable him/her to comprehend qualitative information in a more objective way.

1.2 MEANING, DEFINITIONS AND CHARACTERISTICS OF STATISTICS

The meaning of the word 'Statistics' is implied by the pattern of development of the subject. Since the subject originated with the collection of data, in later years, the techniques of analysis and interpretation were developed, the word 'statistics' has been used in both the plural and the singular sense. Statistics, in plural sense, means a set of numerical figures or data. In the singular sense, it represents a method of study and therefore, refers to statistical principles and methods developed for analysis and interpretation of data. Statistics has been defined in different ways by different authors. These definitions can be broadly classified into two categories. In the first category are those definitions which lay emphasis on statistics as data whereas the definitions in second category emphasise statistics as a scientific method.

Statistics used in the plural sense implies a set of numerical figures collected with reference to a certain problem under investigation. It may be noted here that any set of numerical figures cannot be regarded as statistics. There are certain characteristics which must be satisfied by a given set of numerical figures in order that they may be termed as statistics. Before giving these characteristics it will be advantageous to go through the definitions of statistics in the plural sense, given by noted scholars.

"Statistics are numerical facts in any department of enquiry placed in relation to each other."

–A.L. Bowley

The main features of the above definition are:

- Statistics (or Data) implies numerical facts.
- Numerical facts or figures are related to some enquiry or investigation.
- Numerical facts should be capable of being arranged in relation to each other.

On the basis of the above features, we can say that data are those numerical facts which have been expressed as a set of numerical figures related to each other and to some area of enquiry or research. We may, however, note here that all the characteristics of data are not covered by the above definition.

"By statistics, we mean quantitative data affected to a marked extent by multiplicity of causes."

— Yule & Kendall

This definition covers two aspects, i.e., the data are quantitative and affected by a large number of causes.

"Statistics are classified facts respecting the conditions of the people in a state – especially those facts which can be stated in numbers or in tables of numbers or in any other tabular or classified arrangement."

— Webster

On the basis of the above definitions, we can now state the following characteristics of statistics as data:

- ***Statistics are numerical facts:*** In order that any set of facts can be called as statistics or data, it must be capable of being represented numerically or quantitatively. Ordinarily, the facts can be classified into two categories:
  - ❖ Facts that are measurable and can be represented by numerical measurements. Measurement of heights of students in a college, income of persons in a locality, yield of wheat per acre in a certain district, etc., is examples of measurable facts.
  - ❖ Facts that are not measurable but we can feel the presence or absence of the characteristics. Honesty, colour of hair or eyes, beauty, intelligence, smoking habit, etc., are examples of immeasurable facts. Statistics or data can be obtained in such cases also, by counting the number of individuals in different categories.

  Example: The population of a country can be divided into three categories on the basis of Notes complexion of the people such as white, whitish or black.

- ***Statistics are aggregate of facts:*** A single numerical figure cannot be regarded as statistics. Similarly, a set of unconnected numerical figures cannot be termed as statistics. Statistics means an aggregate or a set of numerical figures which are related to one another. The number of cars sold in a particular year cannot be regarded as statistics. On the other hand, the figures of the number of cars sold in various years of the last decade are statistics because it is an aggregate of related figures. These figures can be compared and we can know whether the sale of cars has increased, decreased or remained constant during the last decade. It should also be noted here that different figures are comparable only if they are expressed in same units and represent the same characteristics under different situations. In the above example, if we have the number of Ambassador cars sold in 1981 and the number of Fiat cars sold in 1982, etc., then it cannot be regarded as statistics. Similarly, the figures of, say, measurement of weight of students should be expressed in the same units in order that these figures are comparable with one another.

- ***Statistics are affected to a marked extent by a multiplicity of factors:*** Statistical data refer to measurement of facts in a complex situation, e.g., business or economic phenomena are very complex in the sense that there are a large number

of factors operating simultaneously at a given point of time. Most of these factors are even difficult to identify. We know that quantity demanded of a commodity, in a given period, depends upon its price, income of the consumer, prices of other commodities, taste and habits of the consumer. It may be mentioned here that these factors are only the main factors but not the only factors affecting the demand of a commodity. Similarly, the sale of a firm in a given period is affected by a large number of factors. Data collected under such conditions are called statistics or statistical data.

- ***Statistics are either enumerated or estimated with reasonable standard of accuracy:*** This characteristic is related to the collection of data. Data are collected either by counting or by measurement of units or individuals. For example, the number of smokers in a village is counted while height of soldiers is measured. We may note here that if the area of investigation is large or the cost of measurement is high, the statistics may also be collected by examining only a fraction of the total area of investigation. When statistics are being obtained by measurement of units, it is necessary to maintain a reasonable degree or standard of accuracy in measurements. The degree of accuracy needed in an investigation depends upon its nature and objectivity on the one hand and upon time and resources on the other.

  Example: In weighing of gold, even milligrams may be significant whereas, for weighing wheat, a few grams may not make much difference. Sometimes, a higher degree of accuracy is needed in order that the problem, to be investigated, gets highlighted by the data. Suppose the diameter of bolts produced by a machine is measured as 1.546 cms, 1.549 cms, 1.548 cms, etc. If, instead, we obtain measurements only up to two places after decimal, all the measurements would be equal and as such nothing could be inferred about the working of the machine. In addition to this, the degree of accuracy also depends upon the availability of time and resources.

  For any investigation, a greater degree of accuracy can be achieved by devoting more time or resources or both. As will be discussed later, in statistics, generalisations about a large group (known as population) are often made on the basis of small group (known as sample). It is possible to achieve this by maintaining a reasonable degree of accuracy of measurements. Therefore, it is not necessary to always have a high degree of accuracy but whatever degree of accuracy is once decided must be uniformly maintained throughout the investigation.

- ***Statistics are collected in a systematic manner and for a predetermined purpose:*** In order that the results obtained from statistics are free from errors, it is necessary that these should be collected in a systematic manner. Haphazardly collected figures are not desirable as they may lead to wrong conclusions. Moreover, statistics should be collected for a well-defined and specific objective, otherwise it might happen that the unnecessary statistics are collected while the necessary statistics are left out. Hence, a given set of numerical figures cannot be termed as statistics if it has been collected in a haphazard manner and without proper specification of the objective.

- ***Statistics should be capable of being placed in relation to each other:*** This characteristic requires that the collected statistics should be comparable with reference to time or place or any other condition. In order that statistics are comparable it is essential that they are homogeneous and pertain to the same investigation. This can be achieved by collecting data in identical manner for different periods or for different places or for different conditions.

Hence, any set of numerical facts possessing the above mentioned characteristics can be termed as statistics or data. The use of the word 'STATISTICS' in singular form refers to a science which provides methods of collection, analysis and interpretation of statistical data. Thus, statistics as a science is defined on the basis of its functions and different scholars have defined it in a different way. In order to know about various aspects of statistics, we now state some of these definitions.

"Statistics is the science of counting."

— A.L. Bowley

"Statistics may rightly be called the science of averages."

— A.L. Bowley

"Statistics is the science of measurement of social organism regarded as a whole in all its manifestations."

— A.L. Bowley

"Statistics is the science of estimates and probabilities."

— Boddington

All of the above definitions are incomplete in one sense or the other because each considers only one aspect of statistics. According to the first definition, statistics is the science of counting. However, we know that if the population or group under investigation is large, we do not count but obtain estimates. The second definition viz. statistics is the science of averages, covers only one aspect, i.e., measures of average but, besides this, there are other measures used to describe a given set of data. The third definition limits the scope of statistics to social sciences only. Bowley himself realised this limitation and admitted that scope of statistics is not confined to this area only. The fourth definition considers yet another aspect of statistics. Although, use of estimates and probabilities have become very popular in modern statistics but there are other techniques, as well, which are also very important. The following definitions cover some more but not all aspects of statistics.

"The science of statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates."

— W.I. King

"Statistics or statistical method may be defined as collection, presentation, analysis and interpretation of numerical data."

— Croxton and Cowden

This is a simple and comprehensive definition of statistics which implies that statistics is a scientific method.

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

— Lovitt

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

— Seligman

The definitions given by Lovitt and Seligman are similar to the definition of Croxton and Cowden except that they regard statistics as a science while Croxton and Cowden have termed it as a scientific method. With the development of the subject of statistics, the definitions of statistics given above have also become outdated. In the last few decades, the discipline of drawing conclusions and making decisions under uncertainty has grown which is proving to be very helpful to decision makers, particularly in the field of business. Although, various definitions have been given which include this aspect of statistics also, we shall now give a definition of statistics, given by Spiegel, to reflect this new dimension of statistics.

"Statistics is concerned with scientific method for collecting, organising, summarising, presenting and analysing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis."

On the basis of the above definitions, we can say that statistics, in singular sense, is a science which consists of various statistical methods that can be used for collection, classification, presentation and analysis of data relating to social, political, natural, economical, business or any other phenomena. The results of the analysis can be used further to draw valid conclusions and to make reasonable decisions in the face of uncertainty.

1.2.1 Statistics as a Scientific Method

We have seen above that, statistics as a non-experimental science can be used to study and analyse various problems of social sciences. It may, however, be pointed out that there may be situations even in natural sciences, where conducting of an experiment under hundred per cent controlled conditions is rather impossible. Statistics, under such conditions, finds its use in natural sciences, like physics, chemistry, etc. In view of the uses of statistics in almost all the disciplines of natural as well as social sciences, it will be more appropriate to regard it as a scientific method rather than a science. Statistics as a scientific method can be divided into the following two categories:

1. ***Theoretical Statistics:*** Theoretical statistics can be further sub-divided into the following three categories:

   (a) *Descriptive Statistics:* All those methods which are used for the collection, classification, tabulation, diagrammatic presentation of data and the methods of calculating average, dispersion, correlation and regression, index numbers, etc., are included in descriptive statistics.

   (b) *Inductive Statistics:* It includes all those methods which are used to make generalisations about a population on the basis of a sample. The techniques of forecasting are also included in inductive statistics.

   (c) *Inferential Statistics:* It includes all those methods which are used to test certain hypotheses regarding characteristics of a population.

2. ***Applied Statistics:*** It consists of the application of statistical methods to practical problems. Design of sample surveys, techniques of quality control, decision-making in business, etc., are included in applied statistics.

1.2.2 Statistics as a Science or an Art

We have seen above that statistics is a science. Now we shall examine whether it is an art or not. We know that science is a body of systematized knowledge. How this knowledge is to be used for solving a problem is work of an art. In addition to this, art also helps in achieving certain objectives and to identify merits and demerits of methods that could be used. Since statistics possesses all these characteristics, it may

be reasonable to say that it is also an art. Thus, we conclude that since statistical methods are systematic and have general applications, therefore, statistics is a science. Further since the successful application of these methods depends, to a considerable degree, on the skill and experience of a statistician, therefore, statistics is an art also.

R.A. Fisher is a notable contributor to the field of statistics. His book 'Statistical Methods for Research Workers', published in 1925, marks the beginning of the theory of modern statistics.

1.3 IMPORTANCE OF STATISTICS

It is perhaps difficult to imagine a field of knowledge which can do without statistics. To begin with, the state started the use of statistics and now it is being used by almost every branch of knowledge such as physics, chemistry, biology, sociology, geography, economics, business, etc. The use of statistics provides precision to various ideas and can also suggest possible ways of tackling a problem relating to any of the above subjects. The importance of statistics has been summarized by A.L. Bowley as, "Knowledge of statistics is like a knowledge of foreign language or of algebra. It may prove of use at any time under any circumstances." We shall discuss briefly, the importance of statistics in the following major areas:

- ***Importance to the State:*** We know that the subject of statistics originated for helping the ancient rulers in the assessment of their military and economic strength. Gradually its scope was enlarged to tackle other problems relating to political activities of the State. In modern era, the role of State has increased and various governments of the world also take care of the welfare of its people. Therefore, these governments require much greater information in the form of numerical figures for the fulfilment of welfare objectives in addition to the efficient running of their administration.

  In a democratic form of government, various political groups are also guided by the statistical analysis regarding their popularity in the masses. Thus, it can be said that it is impossible to think about the functioning of modern state in the absence of statistics.

- ***Importance in economics:*** Statistics is an indispensable tool for a proper understanding of various economic problems. It also provides important guidelines for the formulation of various economic policies. Almost every economic problem is capable of being expressed in the form of numerical figures, e.g., the output of agriculture or of industry, volume of exports and imports, prices of commodities, income of the people, distribution of land holding, etc.

  In each case, the data are affected by a multiplicity of factors. Further, it can be shown that the other conditions prescribed for statistical data are also satisfied. Thus, we can say that the study of various economic problems is essentially the one of a statistical nature. Inductive method of generalisation, popularly used in economics, is also based on statistical principles. Various famous laws in economics such as, the law of diminishing marginal utility, the law of diminishing marginal returns, the theory of revealed preference, etc., are based on generalisations from observation of economic behaviour of a large number of individuals. Statistical methods are also useful in estimating a mathematical relation between various economic variables.

  Example: The data on prices and corresponding quantities demanded of a commodity can be used to estimate the mathematical form of the demand relationship between two variables. Further, the validity of a generalisation or relation between variables can also be tested by using statistical techniques.

Statistical analysis of a given data can also be used for the precise understanding of an economic problem. For example, to study the problem of inequalities of income in a society, we can classify the relevant data and, if necessary, compute certain measures to bring the problem into focus.

Using statistics, suitable policy measures can also be adopted for tackling this problem. Similarly, statistical methods can also be used to understand and to suggest a suitable solution for problems in other areas such as industry, agricultural, human resource development, international trade, etc. Realising the importance of statistics in economics, a separate branch of economics, known as econometrics, has been developed during the recent years. The techniques of econometrics are based upon the principles of economics, statistics and mathematics.

- ***Importance in national income accounting:*** The system of keeping the accounts of income and expenditure of a country is known as national income accounting. These accounts contain information on various macro-economic variables like national income, expenditure, production, savings, investments, volume of exports and imports, etc. The national income accounts of a country are very useful in having an idea about the broad features of its economy or of a particular region. The preparation of these accounts requires data, regarding various variables, at the macro-level. Since such information is very difficult, if not impossible, to obtain, is often estimated by using techniques and principles of statistics.

- ***Importance in planning:*** Planning is indispensable for achieving faster rate of growth through the best use of a nation's resources. It also requires a good deal of statistical data on various aspects of the economy. One of the aims of planning could be to achieve a specified rate of growth of the economy. Using statistical techniques, it is possible to assess the amounts of various resources available in the economy and accordingly determine whether the specified rate of growth is sustainable or not.

  The statistical analysis of data regarding an economy may reveal certain areas which might require special attention, e.g., a situation of growing unemployment or a situation of rising prices during past few years. Statistical techniques and principles can also guide the Government in adopting suitable policy measures to rectify such situations. In addition to this, these techniques can be used to assess various policies of the Government in the past.

  Thus, it is rather impossible to think of a situation where planning and evaluation of various policies can be done without the use of statistical techniques. In view of this, it is sometimes said that, "Planning without statistics is a ship without rudder and compass". Hence statistics is an important tool for the quantification of various planning policies.

- ***Importance in business:*** With the increase in size of business of a firm and with the uncertainties of business because of cut throat competition, the need for statistical information and statistical analysis of various business situations has increased tremendously. Prior to this, when the size of business used to be smaller without much complexities, a single person, usually owner or manager of the firm, used to take all decisions regarding its business. For example, he used to decide, from where the necessary raw materials and other factors of production were to be obtained, how much of output will be produced, where it will be sold, etc. This type of decision-making was usually based on experience and expectations of this single man and as such had no scientific basis.

The modern era is an era of mass production in which size and number of firms have increased enormously. The increase in the number of firms has resulted into cut throat

competition among various firms and, consequently, the uncertainties in business have become greater than before. Under such circumstances, it has become almost impossible for a single man to take decisions regarding various aspects of a rather complex business. It is precisely this point from where the role of statistics started in business.

Now a days no business, large or small, public or private, can prosper without the help of statistics. Statistics provides necessary techniques to a businessman for the formulation of various policies with regard to his business. In fact, the process of collection and analysis of data becomes necessary right from the stage of launching a particular business. Some of the stages of business where statistical analysis has become necessary are briefly discussed below:

- ***Decisions regarding business, its location and size:*** Before starting a business it is necessary to know whether it will be worthwhile to undertake this. This involves a detailed analysis of its costs and benefits which can be done by using techniques and principles of statistics. Furthermore, statistics can also provide certain guidelines which may prove to be helpful in deciding the possible location and size of the proposed business.
- ***Planning of production:*** After a business is launched, the businessman has to plan its production so that he is able to meet the demand of its product and incurs minimum losses on account of over or under production. For this, he has to estimate the pattern of demand of the product by conducting various market surveys. Based upon these surveys, he might also forecast the demand of the product at various points of time in future. In addition to this, the businessman has to conduct market surveys of various resources that will be used in the production of the given output. This may help him in the organisation of production with minimum costs.
- ***Inventory control:*** Sometimes, depending upon the fluctuations in demand and supply conditions, it may not be possible to keep production in pace with demand of the product. There may be a situation of no demand resulting in over production and consequently the firm might have to discontinue production for some time. On the other hand, there may be a sudden rise in the demand of the product so that the firm is able to meet only a part of the total demand. Under such situations, the firm may decide to have an inventory of the product for the smooth running of its business. The optimum limits of inventory, i.e., the minimum and maximum amount of stock to be kept, can be decided by the statistical analysis of the fluctuations in demand and supply of the product.
- ***Quality control:*** Statistical techniques can also be used to control the quality of the product manufactured by a firm. This consists of the preparation of control charts by means of the specification of an average quality. A control chart shows two limits, the lower control limit and the upper control limit for variation in the quality of the product. The samples of output, being produced, are taken at regular intervals and their quality is measured. If the quality falls outside the control limits, steps are taken to rectify the manufacturing process.
- ***Accounts writing and auditing:*** Every business firm keeps accounts of its revenue and expenditure. All activities of a firm, whether big or small, are reflected by these accounts. Whenever certain decisions are to be taken or it is desired to assess the performance of the firm or of its particular section or sections, these accounts are required to be summarised in a statistical way. This may consist of the calculation of typical measures like average production per unit of labour, average production per hour, average rate of return on investment, etc. Statistical methods may also be helpful in generalising relationships between two or more of such

variables. Further, while auditing the accounts of a big business, it may not be possible to examine each and every transaction. Statistics provides sampling techniques to audit the accounts of a business firm. This can save a lot of time and money. Statistics should not be used in the same way as a drunken man uses lamppost for support rather than for illumination.

- ***Banks and insurance companies:*** Banks use statistical techniques to take decisions regarding the average amount of cash needed each day to meet the requirements of day to day transactions. Furthermore, various policies of investment and sanction of loans are also based on the analysis provided by statistics.

  The business of insurance is based on the studies of life expectancy in various age groups. Depending upon these studies, mortality tables are constructed and accordingly the rates of premium to be charged by an insurance company are decided. All this involves the use of statistical principles and methods.

The science of statistics received contributions from notable economists such as Augustin Cournot (1801-1877), Leon Walras (1834-1910), Vilfredo Pareto (1848-1923), Alfred Marshall (1842-1924), Edgeworth, A.L. Bowley, etc. They gave an applied form to the subject.

1.4 SCOPE OF STATISTICS

Statistics is used to present the numerical facts in a form that is easily understandable by human mind and to make comparisons, derive valid conclusions, etc., from these facts. R.W. Buges describes the functions of statistics in these words, "The fundamental gospel of statistics is to push back the domain of ignorance, prejudice, rule of thumb, arbitrary or premature decisions, tradition and dogmatism and to increase the domain in which decisions are made and principles are formulated on the basis of analysed quantitative facts."

The following are the main functions of statistics:

- ***Presents facts in numerical figures:*** The first function of statistics is to present a given problem in terms of numerical figures. We know that the numerical presentation helps in having a better understanding of the nature of problem. Facts expressed in words are not very useful because they are often vague and are likely to be understood differently by different people. For example, the statement that a large proportion of total work force of a country is engaged in agriculture, is vague and uncertain. On the other hand, the statement that 70% of the total work force is engaged in agriculture is more specific and easier to grasp. Similarly, the statement that the annual rate of inflation in a country is 10% is more convincing than the statement that prices are rising.

- ***Presents complex facts in a simplified form:*** Generally a problem to be investigated is represented by a large mass of numerical figures which are very difficult to understand and remember. Using various statistical methods, this large mass of data can be presented in a simplified form. This simplification is achieved by the summarisation of data so that broad features of the given problem are brought into focus. Various statistical techniques such as presentation of data in the form of diagrams, graphs, frequency distributions and calculation of average, dispersion, correlation, etc., make the given data intelligible and easily understandable.

- ***Studies relationship between two or more phenomena:*** Statistics can be used to investigate whether two or more phenomena are related. For example, the relationship between income and consumption, demand and supply, etc., can be

studied by measuring correlation between relevant variables. Furthermore, a given mathematical relation can also be fitted to the given data by using the technique of regression analysis.

- ***Provides techniques for the comparison of phenomena:*** Many a times, the purpose of undertaking a statistical analysis is to compare various phenomena by computing one or more measures like mean, variance, ratios, percentages and various types of coefficients. For example, when we compute the consumer price index for a particular group of workers, then our aim could be to compare this index with that of previous year or to compare it with the consumer price index of a similar group of workers of some other city, etc. Similarly, the inequalities of income in various countries may be computed for the sake of their comparison.
- ***Enlarges individual experiences:*** An important function of statistics is that it enlarges human experience in the solution of various problems. In the words of A.L. Bowley, "the proper function of statistics, indeed is to enlarge individual experience." Statistics is like a master key that is used to solve problems of mankind in every field. It would not be an exaggeration to say that many fields of knowledge would have remained closed to the mankind forever but for the efficient and useful techniques and methodology of the science of statistics.
- ***Helps in the formulation of policies:*** Statistical analysis of data is the starting point in the formulation of policies in various economic, business and government activities. For example, using statistical techniques a firm can know the tastes and preferences of the consumers and decide to make its product accordingly. Similarly, the Government policies regarding taxation, prices, investments, unemployment, imports and exports, etc. are also guided by statistical studies in the relevant areas.

1.5 LIMITATIONS OF STATISTICS

Like every other science, statistics also has its limitations. In order to have maximum advantage from the use of statistical methods, it is necessary to know their limitations. According to Newshome, *"It (statistics) must be regarded as an instrument of research of great value, but having severe limitations, which are not possible to overcome and as such they need our careful attention."* The science of statistics suffers from the following limitations:

- ***Statistics deals with numerical facts only:*** Broadly speaking, there are two types of facts, (a) quantitative and (b) qualitative facts. Quantitative facts are capable of being represented in the form of numerical figures and therefore, are also known as numerical facts. These facts can be analysed and interpreted with the help of statistical methods. Qualitative facts, on the other hand, represent only the qualitative characteristics like honesty, intelligence, colour of eyes, beauty, etc. and statistical methods cannot be used to study these types of characteristics. Sometimes, however, it is possible to make an indirect study of such characteristics through their conversion into numerical figures. For example, we may assign a number 0 for a male and 1 for a female, etc.
- ***Statistics deals only with groups and not with individuals:*** Statistical studies are undertaken to study the characteristics of a group rather than individuals. These studies are done to compare the general behaviour of the group at different points of time or the behaviour of different groups at a particular point of time. For example, the economic performance of a country in a year is measured by its national income in that year and by comparing national income of various years, one can know whether performance of the country is improving or not. Further, by

comparing national income of different countries, one can know its relative position vis-a-vis other countries.

- ***Statistical results are true only on the average:*** Statistical results give the behaviour of the group on the average and these may not hold for an individual of that very group. Thus, the statement that average wages of workers of a certain factory is ₹ 1,500 p.m. does not necessarily mean that each worker is getting this wage. In fact, some of the workers maybe getting more while others less than or equal to ₹ 1,500. Further, when value of a variable is estimated by using some explanatory variable, the estimated value represents the value on the average for a particular value of the explanatory variable. In a similar way, all the laws of statistics are true only on the average.

- ***Statistical results are only approximately true:*** Most of the statistical studies are based on a sample taken from the population. Under certain circumstances, the estimated data are also used. Therefore, conclusions about a population based on such information are bound to be true only approximately. Further, if more observations are collected with a view to improve the accuracy of the results, these efforts are often offset by the errors of observation. In the words of Bowley, when observations are extended, many sources of inaccuracy are found to be present, and it is frequently impossible to remove them completely. Statistical results are, therefore, very general estimates rather than exact statements. Thus, whether statistical results are based on sample or census data, are bound to be true only approximately.

- ***Statistical methods constitute only one set of methods to study a problem:*** A given problem can often be studied in many ways. Statistical methods are used to simplify the mass of data and obtain quantitative results by its analysis. However, one should not depend entirely on statistical results. These results must invariably be supplemented by the results of alternative methods of analysing the problem. It should be kept in mind that statistics is only a means and not an end. According to D. Gregory and H. Ward, *"Statistics cannot run a business or a government. Nor can the study of statistics do more than provide a few suggestions or offer a few pointers as to firm's or government's future behaviour."*

  Statistical techniques, because of their flexibility and economy, have become popular and are used in numerous fields. But statistics is not a cure-all technique and has limitations. It cannot be applied to all kinds of situations and cannot be made to answer all queries.

- ***Statistics are liable to be misused:*** Statistical data are likely to be misused to draw any type of conclusion. If the attitude of the investigator is biased towards a particular aspect of the problem, he is likely to collect only such data which give more importance to that aspect. The conclusions drawn on the basis of such information are bound to be misleading. Suppose, for example, the attitude of the Government is biased and it wants to compute a price index which should show a smaller rise of prices than the actual one. In such a situation, the Government might use only those price quotations that are obtained from markets having lower prices.

- ***Statistics must be used only by experts:*** Statistics, being a technical subject, is very difficult for a common man to understand. Only the experts of statistics can use it correctly and derive right conclusions from the analysis. In the words of Yule and Kendall, *"Statistical methods are the most dangerous tools in the hands of inexperts."* Hence, this is the most important limitation of statistics.

1.6 DEVELOPMENT OF STATISTICS

The word statistics is derived from the Italian word 'Stato' which means 'state'; and 'Statista' refers to a person involved with the affairs of state. Thus, statistics originally was meant for collection of facts useful for affaires of the state, like taxes, land records, population demography, etc. There is an evidence of use of some of the principles of statistics by ancient Indian civilization. Some of the techniques find their mention in Vedic Mathematics. However, the modern statistical methods spread from Italy to France, Holland and Germany in 16^{th} century.

During ancients times even before 300BC, the rulers and kings, like Chandragupta Maurya used statistics to maintain the land and revenue records, collection of taxes and registration of births and deaths. During the seventeenth century, statistics was used in Europe for a variety of information like life expectancy and gambling. Theoretical development of modern statistics was during the mid-seventeenth centuries with the introduction of 'Theory of Probability' and 'Theory of Games and Chance'. Many famous problems like 'the problem of points' (posed by Chevalier de-Mere), 'the gambler's ruin', etc. posed by professional gamblers were solved by mathematicians. These solutions laid the foundation to the theory of probability and statistics. Some of the notable contributors in the development of statistics are: Pascal, Fermat, James Bernoulli, De-Moivre, Laplace, Gauss Euler, Lagrange, Bayes, Kolmogorov, Karl Pearson and so on. One of the most significant works in modern times is by Ronald A. Fisher (1890-1962), who is considered to be the 'Father of Statistics' by the community of statisticians all over. He applied statistics to diversified fields such as education, agriculture, genetics, biometry, psychology, etc. He also pioneered 'Estimation Theory', 'Exact Sampling Distribution', 'Analysis of Variance' and 'Experimental Design'.

Significant contribution has also been made by Indians in the field of statistics. Prof Prasant Chandra Mahalanobis, is the first to pioneer the study of statistical science in India. He founded the Indian Statistical Institute (ISI) in 1931. Mahalanobis viewed statistics as a tool in increasing the efficiency of all human efforts and also concentrated on sample surveys. Mahalanobis is known for his famous work on an important statistic known as D2 statistic, which is very popular among social scientists. Prof C.R. Rao is another Indian Statistician who made significant contribution in the field of statistical inference and multivariate analysis.

1.7 CLASSIFICATION OF STATISTICS

Statistical methods are broadly divided into five categories. These categories are not mutually exclusive. These are often found to be overlapping.

1.7.1 Descriptive Statistics

When statistical methods are used, a problem is always formulated in terms of 'population' or 'universe', which is defined as all the elements about which conclusions or decisions are to be made. In statistics, there is a specific meaning to the words population and universe. We shall discuss exact definitions subsequently. For example, if we want to find customer satisfaction, all our customers represent the population. If information or data is taken from each and every element of the population, we are dealing with 'Descriptive Statistics'. In research vocabulary, such a process is called 'Census'. This includes methods for collection, collation, tabulation, summarization and analysis of the data on entire population. Averages, trends, index numbers, dispersion and skewness help in summarizing and describing the main features of the statistical data. This is primarily to present the data in the form easily

understandable to the decision-maker. One example is the national census conducted every 10 years.

1.7.2 Analytical Statistics

This deals with establishing relationship between two or more variables. This includes methods like correlation and regression, association of attributes, multivariate analysis, etc., which help establishing relationship between variables. This facilitates comparison, interpolation, extrapolation and relationships. In these cases, we require multiple samples on different populations or same population, for example, sales of a product before and after launch of promotion campaign.

1.7.3 Inductive Statistics

Decision-making in most business situations requires estimates about future like trends and forecast. Inductive statistics include methods that help in generalizing the trends based on the random observations. This process provides estimation indirectly on the basis of partial data or method of forecasting based on past data for example, future share price of a share based on the inflow of funds by FII.

1.7.4 Inferential Statistics

Another way, in which conclusions or decisions are made, is using a portion of population or sample from the universe. The sample data is analyzed. Then based on the sample evidence, conclusions are generalized about the target population. Exit poll during elections is an example of sample survey. This method is referred to as 'Statistical Inference'. Hypotheses and significance tests form an important part of inferential statistics.

1.7.5 Applied Statistics

It is the application of statistical methods and techniques used for solving the real life problems. Quality control, sample surveys, inventory management, simulations, quantitative analysis for business decision-making, etc., form a part of this category.

1.8 ROLE OF STATISTICS IN DECISION-MAKING

Very often, people consider decision-making just as an act of selection among alternatives. However, there are two more phases in decision-making. Noble Laureate Sir Herbert A Simon identified the phases of decision-making as:

- ***Information gathering:*** Searching the environment for information, called the intelligence activity.
- ***Generation of alternatives:*** Inventing, developing and analyzing possible courses of action, called the design activity.
- ***Selection of alternatives:*** Selecting a particular course of action from those available, called the decision activity.

Most important task of a manager is to take decisions in a given situation that helps an organization to achieve its goals. Management is a process of converting information into action – this we call decision-making. Decision-making is a deliberate thought process based on available data developing alternatives to choose from so as to find the best solution to the problem at hand.

Statistics and statistical tools play very vital role during all these three phases of decisions. There are two basic approaches of decision-making, namely, quantitative (or mathematical) and qualitative (or rational, creative and judgmental). In the first

approach, statistics and mathematics play dominant role. Even in second approach, statistics plays a role for collection and presentation of data to help decision-maker's intuition. Extent to which statistical and mathematical tools can be used, depend upon the situations. These can be briefly classified as:

- ***Decision-making under certainty:*** These are deterministic situations amenable to mathematical tools to fullest extent.
- ***Decision-making under risk:*** These are stochastic situations amenable to statistical tools to a large extent with supplement of rational decision-making.
- ***Decision-making under uncertainty:*** These are amenable to judgmental and creative approaches.

It is observed that middle level and senior level managers primarily deal with decision-making under risk or in a few cases decision-making under uncertainty. Thus, knowledge of statistical and mathematical computational tools is necessary, if not mandatory, for efficient and effective decision-making. It is not required to apply all advanced statistical tools in every situation. Certain tools may not be applicable in some cases. Simple statistics like average, weighted average, percentage and standard deviation, index would reveal a great deal of information in many decision-making scenarios. Exploratory investigation may, however, require some advanced tools.

1.9 ROLE OF STATISTICS IN RESEARCH

Statistical analysis is a vital component in every aspect of research. Social surveys, laboratory experiment, clinical trials, marketing research, human resource planning, inventory management, quality management, etc., require statistical treatment before arriving at valid conclusions. Today, with availability of computers, we can very effectively apply statistical techniques in every field of knowledge. The findings of any research have to be justified in the light of statistical logic. In business situations, use of statistical tools in marketing research, operations research, forecasting, factor analysis, human resource development, etc., could immensely benefit managers to gain competitive advantage, improve productivity and reduce costs. Thus, every manager must be aware of statistical tools and should have knowledge to use them.

- ***Condensation:*** Statistics compresses mass of figures to small meaningful information, for example, average sales, BSE index (SENSEX), growth rate. It is impossible to get a precise idea about the profitability of a business from a record of income and expenditure transactions. The information of Return on Investment (ROI), Earnings Per Share (EPS), profit margins, etc., however, can be easily remembered, understood and used in decision-making.
- ***Comparison:*** Statistics facilitates comparing two related quantities for example, Price to Earning Ratio (PE Ratio) of Reliance Industries stood at 17.5 as compared to the industry figure of 13 showing the confidence of investors.
- ***Forecast:*** Statistics helps in forecast by looking at trends. These are essential for planning and decision-making. Predictions based on the gut feeling or hunch could be harmful for the business. For example, to decide the refining capacity for a petrochemical plant, we need to predict the demand of petrochemical product mix, supply of crude, cost of crude, substitution products, etc., over next 15 to 25 years, before committing an investment.
- ***Testing of hypotheses:*** Hypotheses are statements about the population parameters based on our past knowledge or information that we would like to check its validity in the light of current information. Inductive inference about the population based on the sample estimates involves an element of risk. However,

sampling keeps the costs of decision-making low. Statistics provides quantitative base for testing our beliefs about the population.

- ***Preciseness:*** Statistics present facts precisely in quantitative form. Statement of facts conveyed in exact quantitative terms are always more convincing than vague utterances. For example, 'increase in profit margin is less in year 2006 than in year 2005' does not convey a definite piece of information. On the other hand, statistics presents the information more definitely like "profit margin is 10% of the turnover in year 2006 against 12% in year 2005".
- ***Expectation:*** Statistics provides the basic building block for framing suitable policies. For example, how much raw material should be imported, how much capacity should be installed, or manpower recruited, etc., depends upon the expected value of outcome of our present decisions.

1.10 LAWS OF STATISTICS

There are two fundamental laws of statistics. These are given below:

1. ***The Law of Statistical Regularity:*** This law states, *"A moderately large number of items, chosen at random from a large group, are almost sure on an average to possess the characteristics of the large group."* For example, it is difficult to predict failure of an individual machine or an accident on express way but not difficult to indicate what percentage of large number of machines might suffer from a breakdown in given period. Similarly, average number of accident on expressway would remain stable over a fairly long period of time unless the conditions have changed drastically.
2. ***The Law of Inertia of Large Number:*** It states, 'Other things being equal, as the sample size increases the result tends to be more reliable and accurate.' As the sample size increases, the possibility of the effect of extreme values in data reduces due to the compensation on the both sides. Thus, as the sample size increases chances of stability of results enhance and confidence in our estimate of the population increases. In the limiting case, if the sample size reaches to the population size we can exactly describe the characteristics of the population.

   Many managers have doubts in using the result of statistical analysis for decision-making, particularly if the analysis goes against their intuition. Some of them also relate it to their past experience when statistical analysis has misled them. The problem of misleading could be due to the incorrect use of data. This happens due to lack of understanding of statistical principles or intentional fudging with the figures with ulterior motives.

1.10.1 Common Statistical Issues

There are different types of statistical issues faced by a researcher. These are broadly classified into the following groups:

- ***Data collection and recording stage:*** These include sampling plan, data collection and data representation.
- ***Computing basic statistics:*** These include proportions, computing central tendency, variation and skewness, measuring consistency of data, frequency distribution and cross tabulation.
- ***Statistical tests of hypotheses:*** These include comparison of means, comparison of proportions and comparison of variances.
- ***Associations and relationship:*** These include testing of dependence between attributes, correction and regression and non-parametric methods.

- ***Multivariate method:*** These include factor analysis, cluster analysis, discriminate analysis, probit and logit analysis, path analysis, profile analysis, multivariate ANOVA and analysis of factorial experiments.

Each of these requires a fundamental understanding of its statistical origin and purpose.

Check Your Progress

Fill in the blanks:

1. Statistics (or Data) implies _____ facts.
2. The word statistics is derived from the Italian word _____ which means 'state'; and 'Statista' refers to a person involved with the affairs of state.
3. Hypotheses and significance tests form an important part of _____ statistics.
4. Searching the environment for information called the _____ activity.
5. If information or data is taken from each and every element of the population, we are dealing with _____ Statistics.
6. The modern statistical methods spread from Italy to France, Holland and Germany in _____ century.

1.11 LET US SUM UP

- The word 'Statistics' can be used in both 'the plural' and 'the singular' sense.
- In plural sense, it implies a set of numerical figures, commonly known as statistical data.
- In singular sense, statistics implies a scientific method used for the collection, analysis and interpretation of data.
- Any set of numerical figures cannot be regarded as statistics or data. A set of numerical figures collected for the investigation of a given problem can be regarded as data only if these are comparable and affected by a multiplicity of factors.
- As a scientific method, statistics is used in almost every subject of natural and social sciences.
- Statistics as a method can be divided into two broad categories viz. Theoretical Statistics and Applied Statistics.
- Theoretical statistics can further be divided into Descriptive, Inductive and Inferential statistics.
- Statistics is used to collect, present and analyze numerical figures on a scientific basis.
- The use of various statistical methods help in presenting complex mass of data in a simplified form so as to facilitate the process of comparison of characteristics in two or more situations.

- Statistics also provide important techniques for the study of relationship between two or more characteristics (or variable), in forecasting, testing of hypothesis, quality control, decision-making, etc.
- Statistics as a scientific method has its importance in almost all subjects of natural and social sciences.
- Statistics is an indispensable tool for the modern government to ensure efficient running of its administration in addition to fulfillment of welfare objectives.
- It is rather impossible to think of planning in the absence of statistics.
- The importance of statistics is also increasing in modern business world.
- Every business, whether big or small, uses statistics for analysing various business situations, including the feasibility of launching a new business.
- The limitations of statistics must always be kept in mind.
- Statistical methods are applicable only if data can be expressed in terms of numerical figures.
- The results of analysis are applicable to groups of individuals or units and are true only on average, etc.

1.12 UNIT END ACTIVITY

Explain by giving reasons whether the following are data or not:

(a) Arun is more intelligent than Avinash.

(b) Arun got 75% marks in B.Sc. and Avinash got 70% marks in B.Com.

(c) Arun was born on August 25, 1974.

1.13 KEYWORDS

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

Applied Statistics: It consists of the application of statistical methods to practical problems.

Descriptive Statistics: All those methods which are used for the collection, classification, tabulation, diagrammatic presentation of data and the methods of calculating average, dispersion, correlation and regression, index numbers, etc., are included in descriptive statistics.

Inductive Statistics: It includes all those methods which are used to make generalisations about a population on the basis of a sample. The techniques of forecasting are also included in inductive statistics.

Inferential Statistics: It includes all those methods which are used to test certain hypotheses regarding characteristics of a population.

National Income Accounting: The system of keeping the accounts of income and expenditure of a country is known as national income accounting.

Numerical Facts: Quantitative facts are capable of being represented in the form of numerical figures and therefore, are also known as numerical facts.

Qualitative Facts: These facts represent only the qualitative characteristics like honesty, intelligence, colour of eyes, beauty, etc.

Quantitative Facts: The facts which are capable to be expressed in forms of quantity/amount are called.

1.14 QUESTIONS FOR DISCUSSION

1. Define the term statistics.
2. Distinguish between statistical methods and statistics.
3. Discuss the scope and significance of the study of statistics.
4. "Statistics are numerical statements of facts, but all facts stated numerically are not statistics". Clarify this statement and point out briefly which numerical statements of facts are statistics.
5. Discuss briefly the utility of statistics in economic analysis and business.
6. "Statistics are the straws out of which one like other economists have to make bricks". Discuss.
7. "Science without statistics bear no fruit, statistics without science have no roots". Explain the statement.
8. "It is usually said that statistics is science and art both". Do you agree with this statement? Discuss the scope of statistics.
9. "Statistics is not a science, it is a scientific method". Discuss it critically and explain the scope of statistics.
10. Explain clearly the three meanings of the word 'Statistics' contained in the following statement: "You compute statistics from statistics by statistics".
11. "Statistics is the backbone of decision-making". Comment.
12. Discuss the nature and scope of statistics.
13. What are the fields of investigation and research where statistical methods and techniques can be usefully employed?
14. Explain the importance of statistics in economic analysis and planning.

Check Your Progress: Model Answer

1. Numerical
2. Stato
3. Inferential
4. Intelligence
5. Descriptive
6. 16^{th}

1.15 REFERENCE & SUGGESTED READINGS

- Berenson, M. L., Levine, D. M., Szabat, K. A., & Stephan, D. F. (2022). Basic Business Statistics: Concepts and Applications(14th ed.). Pearson. ISBN: 9780134684896
- Black, K. (2019). **Business Statistics: For Contemporary Decision Making** (10th ed.). Wiley. ISBN: 9781119433518

- Keller, G. (2020). **Statistics for Management and Economics** (11th ed.). Cengage Learning. ISBN: 9780357108251
- Render, B., Stair, R. M., Hanna, M. E., & Hale, T. S. (2021). **Quantitative Analysis for Management** (13th ed.). Pearson. ISBN: 9780134543162
- McClave, J. T., Benson, P. G., &Sincich, T. (2018). **Statistics for Business and Economics** (13th ed.). Pearson. ISBN: 9780134506594
- Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). **Statistical Techniques in Business and Economics** (17th ed.). McGraw-Hill Education. ISBN: 9781259666360
- Siegel, A. F. (2019). **Practical Business Statistics** (7th ed.). Academic Press. ISBN: 9780128131350

UNIT - II

STATISTICAL INVESTIGATION

CONTENTS

Contd...

2.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Explain the Organisation of Statistical Investigation
- Explain collection, editing and classification of primary and secondary data
- Define tabulation and presentation of data

2.1 INTRODUCTION

By the term investigation (or enquiry), we mean the search for information or knowledge. Statistical investigation thus implies search for knowledge with the help of statistical devices like collection, classification, analysis and interpretation, etc. According to Griffin, *"Statistical enquiries have always required considerable skill on the part of the statistician rooted in a broad knowledge of the subject matter area and combined with considerable ingenuity in overcoming practical difficulties."* So to apply statistical methods to any problem it is necessary to collect the numerical facts since statistical analysis is not possible without them.

Once the researcher has decided the 'Research Design' the next job is of data collection. For data to be useful, our observations need to be organized so that we can get some patterns and come to logical conclusions. Statistical investigation requires systematic collection of data, so that all relevant groups are represented in the data. Depending upon the sources utilized, whether the data has come from actual observations or from records that are kept for normal purposes, statistical data can be classified into two categories – primary and secondary data.

2.2 ORGANISATION OF STATISTICAL INVESTIGATION

The search for knowledge, done by analysing numerical facts, is known as a statistical investigation. A statistical investigation is a process of collection and analysis of data. The relevance and accuracy of data obtained in an investigation depends directly upon the care with which it is planned. A properly planned investigation can give the best results with least cost and time. The investigation of the levels of living of the inhabitants of a particular area, the investigation of relationship between rainfall and the yield of a crop, etc., are some examples of statistical investigation.

A statistical investigation can be done by collecting numeral facts or data through the conduct of statistical surveys. The collected data are then analysed to get the results. Statistical investigation is a long and comprehensive process. It extends over various stages from initial planning to the final preparation of the report. The various stages are mentioned below:

1. Planning of statistical investigation
2. Collection of data
3. Editing of data
4. Presentation of data
5. Analysis of data
6. Interpretation of data
7. Preparation of the report

2.2.1 Planning of Statistical Investigation

A proper system is essential for conducting a statistical investigation. Planning must precede the execution. Careful planning is essential to get the best results at the minimum cost and time. It is essential to consider the following points while planning a statistical investigation:

1. Objective of the enquiry should be fully known.
2. Scope of the enquiry should be determined.
3. Nature of information to be collected should be decided.
4. Unit of data collection should be defined.
5. Source of data collection or type of data to be used, that is, primarily or secondary should be decided.
6. Method of data collection, that is, census or sampling method, should be decided beforehand.
7. Choice of frame should be made.
8. Reasonable standard of accuracy should be fixed.

2.2.2 Collection of Data

Collection data is the first step in a statistical investigation. The person who conducts the enquiry is known as 'investigator'. The persons who help the investigator in collecting the information are called 'enumerators'. The persons from whom the information is collected are known as 'respondents'.

The primary task in any statistical enquiry is to determine its aims and objectives. Once these objectives have been determined, the next task is to collect the data. The data to be used can be of two types namely (1) Primary data and (2) Secondary data.

2.2.3 Editing of Data

When the researcher collects the data it is in raw form and it needs to be edited, organized and analyzed. The raw data needs to be transformed into a comprehensible form of data. The first steps in this process are to edit the data. The edited data is then coded and inferences are drawn. The editing of the data is not a complex task but it requires an experienced, talented and knowledgeable person to do so. With editing the data the researcher makes sure that all responses are now very clear to understand. Bringing clarity is important otherwise the researcher can draw wrong inferences from the data. Sometimes the respondents make some spelling and grammatical mistakes the editor needs to correct them. The respondents might not be able to express their opinion in proper wording. The editor can rephrase the response, but he needs to be very careful in doing so. Any bias can be introduced by taking the wrong meanings of the respondents' point of view.

2.2.4 Presentation of Data

Presenting the data includes the pictorial representation of the data by using graphs, charts, maps and other methods. These methods help in adding visual aspect to data which makes it much easier and quick to understand. A great presentation can be a deal maker or deal breaker. Some people make extremely effective presentation with the same set of facts and figures which are available with others. At times people who did all the hard work but failed to present it present it properly have lost important contracts, the work which they did is unable to impress the decision makers. You can have variety of data which can be used in presentations. Some of these types include:

- Time Series Data
- Bar Charts
- Combo Charts
- Pie Charts
- Tables
- Geo Map
- Scorecard
- Scatter Charts
- Bullet Charts
- Area Chart
- Text & Images

2.2.5 Analysis of Data

Data analysis helps in interpretation of data and takes a decision or answer the research question. Data analysis starts with the collection of data followed by sorting

and processing it. Processed data helps in obtaining information from it as the raw data is non-comprehensive in nature. Data analysis helps people in understanding the results of surveys conducted, makes use of already existing studies to obtain new results. Analysis of data helps in validating the existing study or to add/expand existing study.

2.2.6 Interpretation of Data

The collection of the data is followed by the analysation of the data, which further is followed by the interpretation of the data. This step enables the researcher to interpret the results which have been obtained from the analysation of the data.

According to C. William Emory, *"Interpretation has two major aspects namely establishing continuity in the research through linking the results of a given study with those of another and the establishment of some relationship with the collected data. Interpretation can be defined as the device through which the factors, which seem to explain what has been observed by the researcher in the course of the study, can be better understood. Interpretation provides a theoretical conception which can serve as a guide for the further research work".*

Interpretation of the data has become a very important and essential process, mainly because of some of the following factors:

- Enables the researcher to have an in – depth knowledge about the abstract principle behind his own findings.
- The researcher is able to understand his findings and the reasons behind their existence.
- More understanding and knowledge can be obtained with the help of the further research.
- Provides a very good guidance in the studies relating to the research work.
- Sometimes may result in the formation of the hypothesis.

2.2.7 Preparation of the Report

A report is the formal writing up of a project or a research investigation. A report has clearly defined sections presented in a standard format, which are used to tell the reader what you did, why and how you did it and what you found. Reports differ from essays because they require an objective writing style which conveys information clearly and concisely.

2.3 COLLECTION OF DATA

The collection and analysis of data constitute the main stages of execution of any statistical investigation. The procedure for collection of data depends upon various considerations such as objective, scope, nature of investigation, etc. Availability of resources like money, time, manpower, etc., also affects the choice of a procedure. Data may be collected either from a primary or from a secondary source. They are described below.

2.3.1 Types of Data – Primary and Secondary

Data used in statistical study is termed either 'primary' or 'secondary' depending upon whether it was collected specifically for the study undertaken or for some other purposes.

When the data used in a statistical study was collected under the control and supervision of the investigator, such type of data is referred to as 'primary data'.

Primary data are collected afresh and for the first time, and thus, happen to be original in character. On the other hand, when the data is not collected for this purpose, but is derived from other sources then such data is referred to as 'secondary data'. Generally speaking, secondary data are collected by some other organization to satisfy their need but being used by someone else for entirely different reasons.

The difference between primary and secondary data is only in terms of degree. For example, data, which are primary in the hands of one, becomes secondary in the hands of another. Suppose an investigator wants to study the working conditions of labourers in an industry. If the investigator or his agent collects the data directly, then it is called a 'primary data'. But if subsequently someone else uses this collected data for some other purpose, then this data becomes a 'secondary data'.

2.3.2 Methods of Collecting Primary Data

Generally, for managerial decision-making, it is necessary to analyze information regarding a large number of characteristics. Collection of primary data may thus be time consuming, expensive, and hence requires a great deal of deliberation. According to the nature of information required, one of the following methods or their combination could be selected.

- ***Observation Method:*** In this method, investigator collects the data through his/her personal observations. This method is very useful if data is created in the system through capturing transactions. Computerized transaction processing could be modified to generate necessary data or information. An investigator well versed with the system or a part of the system is ideally suited for collecting this kind of data. Since the investigator is solely involved in collecting the data, his/her training, skill and knowledge plays an important role as far as the quality of the data is concerned. Sometimes, audio/video aids could also be used to record the observations.

- ***Indirect Investigation:*** In this case, data is collected from a person, who is likely to have information about the problem under study. The information collected by oral or written interrogation forms a primary data. Usually enquiry commissions, board of investigations, investigation teams and committees collect data in this manner. Quality of the data largely depends upon the person interviewed, his/her motives, memory and co-operation, and interviewer's repute and rapport with the person being interviewed. We should be careful while collecting data by this method.

- ***Questionnaire with Personal Interview:*** This is by far the most common and popular method. In this method, individuals are personally interviewed and answers recorded to collect the data. Questionnaire is structured and followed in specific sequence. Occasionally, a part of the questionnaire may be unstructured to motivate the interviewee to give additional information or information on intimate matters. Accuracy of the data depends on the ability, sincerity and tactfulness of the interviewer to conduct the interview in friendly and professional environment.

- ***Mailed Questionnaire:*** In this case, structured questionnaire is mailed to selected persons with request to fill them and return. Supplementary information clarifying terms, explaining process, etc., is also attached with the questions. In a few cases, inducements for filling and returning the questionnaire are also given. Covering letter with a questionnaire is necessary for developing rapport, explaining the reason for collecting the data, and alleviating fears of the respondent if any. It is assumed that the respondents are literate and can answer the questions without any ambiguity. This is a less expensive and faster method to collect large volume of

data, over a wide geographic area, in standard form, and at the convenience of the respondent. This method is, therefore, most popular and extensively used. However, we must guard against two disadvantages of this method viz. absence of interviewer, resulting in large proportion of non-response and possibility of lowering of the reliability of the responses if the respondent is not motivated enough. These shortcomings could be overcome by increasing sample size and comprehensive design of questionnaire.

- ***Telephonic Interview:*** This method is less expensive but limited in scope as the respondent must possess a telephone and has it listed. Further, the respondent must be available and in the frame of mind to provide correct answers. This method is comparatively less reliable for public surveys. However, for industrial survey, in developed regions, and with known customers, this method could be the best suited. Obviously, in this method, there is a limit to the number of questions that the interviewee could answer in three to four minutes. If there are just three to five yes/no type questions and two to three short questions, this method is very efficient.
- ***Internet Surveys:*** Of late, Internet surveys have become popular. These are less expensive, fast and could be interactive. However, its scope is limited to those who have regular Internet access. With rapid growth in personal computers and Internet connectivity it would be one of the main methods of collecting primary data. With its interactivity and multimedia facilities it combines the advantages of other methods.

2.3.3 Merits and Demerits of Collecting Primary Data

Type of research, its purpose, conditions under which the data are obtained will determine the method of collecting the data. If relatively few items of information are required quickly, and funds are limited telephonic interviews are recommended. If respondents are industrial clients Internet could also be used. If depth interviews and probing techniques are to be used, it is necessary to employ investigators to collect data. Thus, each method has its utility and none is superior in all situations. We could combine two methods to improve the quality of data collected. For example, when a wide geographical area is being covered, the mail questionnaires supplemented by personnel interviews will yield more reliable results.

Merits

- Original data are collected.
- Collected data are more accurate and reliable.
- The investigator can modify or put indirect questions in order to extract satisfactory information.
- The collected data are often homogeneous and comparable.
- Some additional information may also get collected, along with the regular information, which may prove to be helpful in future investigations.
- Misinterpretations or misgivings, if any, on the part of the respondents can be avoided by the investigators.
- Since the information is collected from the persons who are well aware of the situation, it is likely to be unbiased and reliable.
- This method is particularly suitable for the collection of confidential information. For example, a person may not like to reveal his habit of drinking, smoking, gambling, etc., which may be revealed by others.

Demerits

- This method is expensive and time consuming, particularly when the field of investigation is large.
- It is not possible to properly train a large team of investigators.
- The bias or prejudice of investigators can affect the accuracy of data to a large extent.
- Data are collected as per the convenience and willingness of the respondents.
- The persons, providing the information, may be prejudiced or biased.
- Since the interest of the person, providing the information, is not at stake, the collected information is often vague and unreliable.
- The information collected from different persons may not be homogeneous and comparable.

2.3.4 Methods of Collecting Secondary Data

Secondary data is one that has been collected/analyzed by some other agency for another purpose.

Sources of secondary data could be:

- Various publications of central, state and local governments. This is an important and reliable source to get unbiased data.
- Various publications of foreign governments or of international bodies. Although it is a good source, context under which it is collected needs to be verified before using this data. For international situations, this data could be very useful and authentic.
- Journals of trade, commerce, economics, scientific, engineering, medicine, etc. This data could be very reliable for a specific purpose.
- Other published sources like books, magazines, newspapers, reports, etc.

Unpublished data, based on internal records and documents of an organization could provide most authentic and much cheaper information provided we could identify the source. Diaries, letters, etc. could also provide a secondary data. The problem with the unpublished data is that it's difficult to locate and get access.

2.4 DESIGNING QUESTIONNAIRE

The success of collecting data through a questionnaire depends mainly on how skillfully and imaginatively the questionnaire has been designed. A badly designed questionnaire will never be able to gather the relevant data.

In designing the questionnaire, some of the important points to be kept in mind are:

- ***Covering letter:*** Every questionnaire should contain a covering letter. The covering letter should highlight the purpose of study and assure the respondent that all responses will be kept confidential. It is desirable that some inducement or motivation is provided to the respondent for better response. The objectives of the study and questionnaire design should be such that the respondent derives a sense of satisfaction through his involvement.
- ***Number of questions should be kept to the minimum:*** The fewer the questions, the greater the chances of getting a better response and of having all the questions answered. Otherwise the respondent may feel disinterested and provide inaccurate

answers particularly towards the end of the questionnaire. As a rough indication, the number of questions should be between 10 to 20. If number of questions have to be more than 25, it is desirable that the questionnaire be divided into various parts to ensure clarity.

- ***Questions should be simple, short and unambiguous:*** The questions should be simple, short, and easy to understand and such that their answers are unambiguous. For example, if the question is, "Are you literate?" the respondent may have doubts about the meaning of literacy. To some, literacy may mean a university degree whereas to others even the capacity to read and write may mean literacy. Hence, it is desirable to specify "Have passed (a) high school, (b) graduation and (c) post-graduation".
- ***Type of questions:*** Questions can be of Yes/No type, or of multiple choices depending on the requirement of the investigator. Open-ended questions should generally be avoided.
- ***Questions of sensitive or personal nature should be avoided:*** The questions should not require the respondent to disclose any private, personal or confidential information. For example, questions relating to sales, profits, marital happiness, tax liability, etc., should be avoided as far as possible. If such questions are necessary in the survey, an assurance should be given to the respondent that the information provided shall be kept strictly confidential and shall not be used at any cost to respondent's disadvantage.
- ***Answers to questions should not require calculations:*** The questions should be framed in such a way that their answers do not require any calculations.
- ***Logical arrangement:*** The questions should be logically arranged so that there is a continuity of responses and the respondent does not feel the need to refer back to the previous questions. It is desirable that the questionnaire should begin with some introductory questions followed by vital questions crucial to the survey and ending with some light questions so that the overall impression of the respondent is a happy one.
- ***Crosscheck and footnotes:*** The questionnaire should contain some such questions, which act as a crosscheck to the reliability of the information provided. For example, when a question relating to income is asked, it is desirable to include a question: "Are you an income tax payer?" Certain questions might create a doubt in the mind of respondents. For the purpose of clarity, it is desirable to give footnotes. The purpose of footnotes is to clarify all possible doubts, which may emerge from the questions and cannot be removed while framing them. For example, if a question relates to income limits like 1000-2000, 2000-3000, etc., a person getting exactly ₹ 2000 should know in which income class he has to place himself.
- ***Pre-test the questionnaire:*** Once the questionnaire has been designed, it is important to pre-test it. The pre-testing is also known as pilot survey because it precedes the main survey work. Pre-testing allows rectification of problems, inconsistencies, repetition, etc. Proper testing, revisiting, and re-testing, yields high dividends.

2.4.1 Importance of Questionnaire in Research

To study:

- Behaviour, past and present
- Demographic characteristics such as age, sex, income and occupation

- Attitudes and opinions
- Level of knowledge

2.4.2 Developing a Good Questionnaire

- It must be simple. The respondents should be able to understand the questions.
- It must generate replies that can be easily be recorded by the interviewer.
- It should be specific, so as to allow the interviewer to keep the interview to the point.
- It should be well arranged, to facilitate analysis and interpretation.
- It must keep the respondent interested throughout.

2.4.3 Merits and Demerits of Questionnaire Method

Merits

- This method is useful for the collection of information from an extensive area of investigation.
- This method is economical as it requires less time, money and labour.
- The collected information is original and more reliable.
- It is free from the bias of the investigator.

Demerits

- Very often, there is problem of 'non-response' as the respondents are not willing to provide answers to certain questions.
- The respondents may provide wrong information if the questions are not properly understood.
- It is not possible to collect information if the respondents are not educated.
- It is not possible to ask supplementary questions, the method is not flexible.
- The results of an investigation are likely to be misleading if the attitude of the respondents is biased.
- The process is time consuming, particularly when the information is to be obtained by post.

2.5 TYPES OF QUESTIONNAIRE

Some of the types of questionnaire are given below:

1. Structured and non-disguised
2. Structured and disguised
3. Non-structured and Disguised
4. Non-structured and Non-disguised

2.5.1 Structured and Non-disguised Questionnaire

Here, questions are structured so as to obtain the facts. The interviewer will ask the questions strictly in accordance with the pre-arranged order. For example, what are the strengths of soap A in comparison with soap B?

- Cost is less

- Lasts longer
- Better fragrance
- Produces more lather
- Available in more convenient sizes

Structured and non-disguised questionnaire is widely used in market research. Questions are presented with exactly the same wording and same order to all respondents. The reason for standardizing the question is to ensure that all respondents reply the same question. The purpose of the question is clear. The researcher wants the respondent to choose one of the five options given above. This type of questionnaire is easy to administer. The respondents have no difficulty in answering, because it is structured, the frame of reference is obvious.

In a non-disguised type, the purpose of the questionnaire is known to the respondent.

Example: "Subjects attitude towards Cyber laws and the need for government legislation to regulate it".

- Certainly, not needed at present
- Certainly not needed
- I can't say
- Very urgently needed
- Not urgently needed

2.5.2 Structured and Disguised Questionnaire

This type of questionnaire is least used in marketing research. This type of questionnaire is used to know the peoples' attitude, when a direct undisguised question produces a bias. In this type of questionnaire, what comes out is "what does the respondent know" rather than what he feels. Therefore, the endeavour in this method is to know the respondent's attitude.

Currently, the "Office of Profit" Bill is:

- In the Lok Sabha for approval.
- Approved by the Lok Sabha and pending in the Rajya Sabha.
- Passed by both the Houses, pending the presidential approval.
- The bill is being passed by the President.

Depending on which answer the respondent chooses, his knowledge on the subject is classified.

In a disguised type, the respondent is not informed of the purpose of the questionnaire. Here the purpose is to hide "what is expected from the respondent?"

Examples:

- "Tell me your opinion about Mr. Ram's healing effect show conducted at Bangalore?"
- "What do you think about the Babri Masjid demolition?"

2.5.3 Non-structured and Disguised Questionnaire

The main objective is to conceal the topic of enquiry by using a disguised stimulus. Though the stimulus is standardized by the researcher, the respondent is allowed to answer in an unstructured manner. The assumption made here is that individual's

reaction is an indication of respondent's basic perception. Projective techniques are examples of non-structured disguised technique. The techniques involve the use of a vague stimulus, which an individual is asked to expand or describe or build a story, three common types under this category are (a) Word association, (b) Sentence completion and (c) Story telling.

2.5.4 Non-structured and Non-disguised Questionnaire

Here the purpose of the study is clear, but the responses to the question are open-ended.

Example: "How do you feel about the Cyber law currently in practice and its need for further modification"? The initial part of the question is consistent. After presenting the initial question, the interview becomes much unstructured as the interviewer probes more deeply. Subsequent answers by the respondents determine the direction the interviewer takes next. The question asked by the interviewer varies from person to person. This method is called "the depth interview". The major advantage of this method is the freedom permitted to the interviewer. By not restricting the respondents to a set of replies, the experienced interviewers will be above to get the information from the respondent fairly and accurately. The main disadvantage of this method of interviewing is that it takes time, and the respondents may not co-operate. Another disadvantage is that coding of open-ended questions may pose a challenge.

For example, when a researcher asks the respondent "Tell me something about your experience in this hospital". The answer may be "Well, the nurses are slow to attend and the doctor is rude. 'Slow' and 'rude' are different qualities needing separate coding. This type of interviewing is extremely helpful in exploratory studies.

2.6 PREPARATION OF QUESTIONNAIRE

The following are the seven steps:

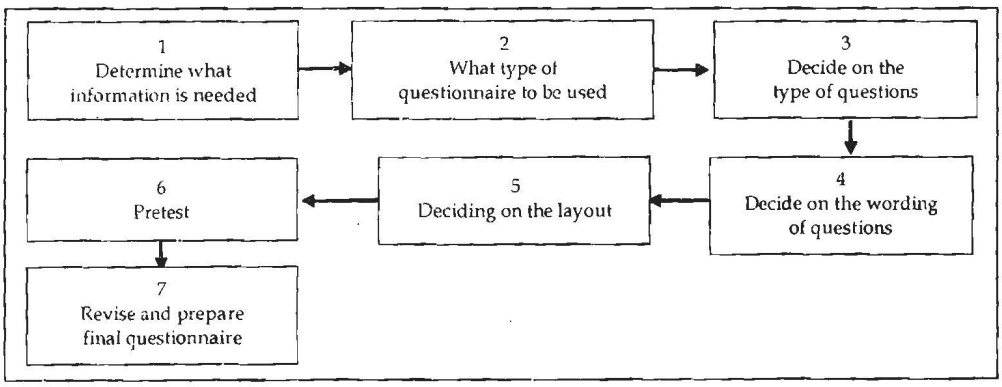


Figure 2.1: Seven Steps in Questionnaire

2.6.1 Determine What Information is Needed

The first question to be asked by the market researcher is "what type of information does he need from the survey?" This is valid because if he omits some information on relevant and vital aspects, his research is not likely to be successful. On the other hand, if he collects information which is not relevant, he is wasting his time and money.

At this stage, information required, and the scope of research should be clear. Therefore, the steps to be followed at the planning stage are:

- Decide on the topic for research.

- Get additional information on the research issue, from secondary data and exploratory research. The exploratory research will suggest "what are the relevant variables?"
- Gather what has been the experience with similar study.
- The type of information required. There are several types of information such as

  (a) awareness, (b) facts, (c) opinions, (d) attitudes,

  (e) future plans, and (f) reasons.

Facts are usually sought out in marketing research.

Example: Which television programme did you see last Saturday? This requires a reasonably good memory and the respondent may not remember. This is known as recall loss. Therefore questioning the distant past should be avoided. Memory of events depends on (1) Importance of the events and (2) Whether it is necessary for the respondent to remember. In the above case, both the factors are not fulfilled. Therefore, the respondent does not remember. On the contrary, a birthday or wedding anniversary of individuals is remembered without effort since the event is important. Therefore, the researcher should be careful while asking questions about the past. First, he must make sure that the respondent has the answer.

Example: Do you go to the club? He may answer 'yes', though it is untrue. This may be because the respondent wants to impress upon the interviewer that he belongs to a well-to-do family and can afford to spend money on clubs. To obtain facts, the respondents must be conditioned (by good support) to part with the correct facts.

2.6.2 Mode of Collecting the Data

The questionnaire can be used to collect information either through personal interview, mail or telephone. The method chosen depends on the information required and also the type of respondent. If the information is to be collected from illiterate individuals, a questionnaire would be the wrong choice.

2.6.3 Types of Questions

The types of questions are given below:

Open-ended Questions

These are questions where respondents are free to answer in their own words.

Example: "What factor do you consider while buying a suit"? If multiple choices are given, it could be colour, price, style, brand, etc., but some respondents may mention attributes which may not occur to the researcher.

Therefore, open-ended questions are useful in exploratory research, where all possible alternatives are explored. The greatest disadvantage of open-ended questions is that the researcher has to note down the answer of the respondents verbatim. Therefore, there is a likelihood of the researcher failing to record some information.

Another problem with open-ended question is that the respondents may not use the same frame of reference.

Example: "What is the most important attribute in a job?"

Answer: Pay

The respondent may have meant "basic pay" but interviewer may think that the respondent is talking about "total pay including dearness allowance and incentive". Since both of them refer to pay, it is impossible to separate two different frames.

Dichotomous Question

These questions have only two answers, 'Yes' or 'no', 'true' or 'false' 'use' or 'don't use'.

Do you use toothpaste? Yes ..... No .....

There is no third answer. However sometimes, there can be a third answer.

Example: "Do you like to watch movies?"

Answer: Neither like nor dislike

Dichotomous questions are most convenient and easy to answer.

Close-ended Questions

There are two basic formats in this type:

1. Make one or more choices among the alternatives
2. Rate the alternatives

(a) ***Choice among Alternatives:*** Which of the following words or phrases best describes the kind of person you feel would be most likely to use this product, based on what you have seen in the commercial?

Young ..... old .....

Single ..... Married .....

Modern ..... Old fashioned .....

(b) ***Rating Scale***

(i) Please tell us your overall reaction to this commercial?

- A great commercial would like to see again.
- Just so-so, like other commercials.
- Another bad commercial.
- Pretty good commercial.

(ii) Based on what you saw in the commercial, how interested do you feel, you would be buying the products?

- Definitely
- Probably I would buy
- I may or may not buy
- Probably I would not buy
- Definitely I would not buy

Closed-ended questionnaires are easy to answer. It requires less effort on the part of the interviewer. Tabulation and analysis is easier. There are lesser errors, since the same questions are asked to everyone. The time taken to respond is lesser. We can compare the answer of one respondent to another respondent.

One basic criticism of closed-ended questionnaires is that middle alternatives are not included in this, such as "don't know". This will force the respondents to choose among the given alternative.

2.6.4 Wordings of Questions

Wordings of particular questions could have a large impact on how the respondent interprets them. Even a small shift in the wording could alter the respondent's answer.

Examples:

- "Don't you think that Brazil played poorly in the FIFA cup?" The answer will be 'yes'. Many of them, who do not have any idea about the game, will also most likely say 'yes'. If the question is worded in a slightly different manner, the response will be different.
- "Do you think that, Brazil played poorly in the FIFA cup?" This is a straightforward question. The answer could be 'yes', 'no' or 'don't know' depending on the knowledge the respondents have about the game.
- "Do you think anything should be done to make it easier for people to pay their phone bill, electricity bill and water bill under one roof"?
- "Don't you think something might be done to make it easier for people to pay their phone bill, electricity bill, water bill under one roof"?

A change of just one word as above can generate different responses by respondents.

Guidelines towards the use of correct wording:

Is the vocabulary simple and familiar to the respondents?

Examples:

- Instead of using the word 'reasonably', 'usually', 'occasionally', 'generally', 'on the whole'.
- "How often do you go to a movie?" "Often, may be once a week, once a month, once in two months or even more."

2.6.5 Avoid Double-Barrelled Questions

These are questions, in which the respondent can agree with one part of the question, but not agree with the other or cannot answer without making a particular assumption.

Examples:

- "Do you feel that firms today are employee-oriented and customer-oriented?" There are two separate issues here – [yes] [no]
- "Are you happy with the price and quality of branded shampoo?" [yes] [no]

2.6.6 Avoid Leading and Loading Questions

Leading Questions

A leading question is one that suggests the answer to the respondent. The question itself will influence the answer, when respondents get an idea that the data is being collected by a company. The respondents have a tendency to respond positively.

Examples:

1. "How do you like the programme on 'Radio Mirchy'? The answer is likely to be 'yes'. The unbiased way of asking is "which is your favourite F.M. Radio station? The answer could be any one of the four stations namely:

   (a) Radio City

   (b) Mirchi

   (c) Rainbow

   (d) Radio-One

2. Do you think that offshore drilling for oil is environmentally unsound? The most probable response is 'yes'. The same question can be modified to eliminate the leading factor.

What is your feeling about the environmental impact of offshore drilling for oil? Give choices as follows:

- Offshore drilling is environmentally sound.
- Offshore drilling is environmentally unsound.
- No opinion.

Loaded Questions

A leading question is also known as a loaded question. In a loaded question, special emphasis is given to a word or a phrase, which acts as a lead to respondent.

Examples:

- "Do you own a Kelvinator refrigerator?"
- A better question would be "what brand of refrigerator do you own?"
- "Don't you think the civic body is 'incompetent'?"
- Here the word incompetent is 'loaded'.

Are the Questions Confusing?

If there is a question unclear or is confusing, then the respondent becomes more biased rather that getting enlightened.

Example: "Do you think that the government publications are distributed effectively"?

This is not the correct way, since respondent does not know what the meaning of the word effective distribution is. This is confusing. The correct way of asking questions is "Do you think that the government publications are readily available when you want to buy?"

Example: "Do you think whether value price equation is attractive"? Here, respondents may not know the meaning of value price equation.

Applicability

"Is the question applicable to all respondents?" Respondents may try to answer a question even though they don't qualify to do so or may lack from any meaningful opinion.

Examples:

- "What is your present education level"
- "Where are you working" (assuming he is employed)?
- "From which bank have you taken a housing loan" (assuming he has taken a loan).

Avoid Implicit Assumptions

An implicit alternative is one that is not expressed in the options. Consider following two questions:

1. Would you like to have a job, if available?
2. Would you prefer to have a job, or do you prefer to do just domestic work?

Even though, we may say that these two questions look similar, they vary widely. The difference is that Q-2 makes explicit the alternative implied in Q-1.

2.6.7 Split Ballot Technique

This is a procedure used wherein (1) The question is split into two halves and (2) Different sequencing of questions is administered to each half. There are occasions when a single version of questions may not derive the correct answer and the choice is not obvious to the respondent.

Example: "Why do you use Ayurvedic soap"? One respondent might say "Ayurvedic soap is better for skin care". Another may say "Because the dermatologist has recommended". A third might say "It is a soap used by my entire family for several years". The first respondent answers the reason for using it at present. The second respondent answers how he started using. The third respondent says "the family tradition for using". As can be seen, different reference frames are used. The question may be balanced and rephrased.

Complex Questions

In which of the following do you like to park your liquid funds?

- Debenture
- Preferential share
- Equity linked M.F
- I.P.O
- Fixed deposit

If this question is posed to the general public, they may not know the meaning of liquid fund. Most of the respondents will guess and tick one of them.

Are the Questions Too Long?

Generally as a thumb rule, it is advisable to keep the number of words in a question not exceeding 20. The question given below is too long for the respondent to comprehend, leave alone answer.

Example: Do you accept that the people whom you know, and associate yourself have been receiving ESI and P.F benefits from the government accept a reduction in those benefits, with a view to cut down government expenditure, to provide more resources for infrastructural development?

Yes _____ No_____ Can't say_____

Participation at the Expense of Accuracy

Sometimes the respondent may not have the information that is needed by the researcher.

Examples:

- The husband is asked a question "How much does your family spend on groceries in a week"? Unless the respondent does the grocery shopping himself, he will not know how much has been spent. In a situation like this, it will be helpful to ask a 'filtered question'. An example of a filtered question can be, "Who buys the groceries in your family"?
- "Do you have the information of Mr. Ben's visit to Bangalore"? Not only should the individual have the information but also (he) should remember the same. The inability to remember the information is known as "recall loss".

2.6.8 Sequence and Layout

Some guidelines for sequencing the questionnaire are as follows:

Dividing the Questionnaire

Divide the questionnaire into three parts:

1. Basic information
2. Classification
3. Identification information

Items such as age, sex, income, education, etc. are questioned in the classification section. The identification part involves body of the questionnaire. Always move from general to specific questions on the topic. This is known as funnel sequence. Sequencing of questions is illustrated below:

1. Which TV shows do you watch?

   Sports _____ News_____

2. Which among the following are you most interested in?

   Sports _____ News_____

   Music _____ Cartoon_____

3. Which show did you watch last week?

   World Cup Football _____

   Bournvita Quiz Contest _____

   War News in the Middle East _____

   Tom and Jerry cartoon show_____

The above three questions follow a funnel sequence. If we reverse the order of question and ask "which show was watched last week?" the answer may be biased. This example shows the importance of sequencing.

Layout

How the questionnaire looks or appears.

Example: Clear instructions, gaps between questions, answers and spaces are part of layout. Two different layouts are shown below:

Layout 1: *How old is your bike?*

_____ Less than 1 year

_____ 1 to 2 years

_____ 2 to 4 years

_____ more than 4 years

Layout 2: *How old is your bike?*

_____ Less than 1 year

_____ 1 to 2 years

_____ 2 to 4 years

_____ More than 4 years

From the above example, it is clear that layout – 2 is better. This is because likely respondent error due to confusion is minimised.

Therefore, while preparing a questionnaire start with a general question. This is followed by a direct and simple question. This is followed by more focused questions. This will elicit maximum information.

Forced and Unforced Scales

Suppose the questionnaire is not provided with 'don't know' or 'no option', then the respondent is forced to choose one side or the other. "Don't know" is not a neutral response. This may be due to genuine lack of knowledge.

Balanced and Unbalanced Scales

In a balanced scale, the numbers of favourable responses are equal to the number of unfavourable responses. If the researcher knows that there is a possibility of a favourable response, it is best to use unbalanced scale.

Use Funnel Approach

Funnel sequencing gets the name from its shape, starting with broad questions and progressively narrowing down the scope. Move from general to specific examples.

- How do you think this country is getting along in its relations with other countries?
- How do you think we are doing in our relations with the US?
- Do you think we ought to be dealing with US?
- If yes, what should be done differently?
- Some say we are very weak on the nuclear deal with the US, while, some say we are OK. What do you feel?

The first question introduces the general subject. In the next question, a specific country is mentioned. The third and fourth questions are asked to seek views. The fifth question is to seek a specific opinion.

Pre-testing of Questionnaire

Pre-testing of a questionnaire is done to detect any flaws that might be present.

Example: The word used by researcher must convey the same meaning to the respondents. Are instructions clear skip questions clear?

One of the prime conditions for pre-testing is that the sample chosen for pre-testing should be similar to the respondents who are ultimately going to participate. Just because a few chosen respondents fill in all the questions going does not mean that the questionnaire is sound.

How Many Questions to be Asked?

The questionnaire should not be too long as the response will be poor. There is no rule to decide this. However, the researcher should consider that if he were the respondent, how he would react to a lengthy questionnaire. One way of deciding the length of the questionnaire is to calculate the time taken to complete the questionnaire. He can give the questionnaire to a few known people to seek their opinion.

2.7 MAIL QUESTIONNAIRE

Mail questionnaires can be explained as the questionnaires that are mailed to the respondents who can complete them at their convenience in their homes and at their own pace. They are expected to meet with a better response rate when respondents are notified in advance about the forthcoming survey and a reputed research organisation administers them with its own introductory cover letter.

2.7.1 Advantages of Mail Questionnaire

- Easier to reach a larger number of respondents throughout the country.
- Since the interviewer is not present face to face, the influence of interviewer on the respondent is eliminated.
- Where the questions asked are such that they cannot be answered immediately, and needs some thinking on the part of the respondent, the respondent can think over leisurely and give the answer.
- Saves cost (cheaper than interview).
- No need to train interviewers.
- Personal and sensitive questions are well answered.

2.7.2 Limitations of Mail Questionnaire

- It is not suitable when questions are difficult and complicated.

  Example: "Do you believe in value price relationship"?

- When the researcher is interested in a spontaneous response, this method is unsuitable. Because thinking time allowed to the respondent will influence the answer.

  Example: "Tell me spontaneously, what comes to your mind if I ask you about cigarette smoking".

- In case of a mail questionnaire, it is not possible to verify whether the respondent himself/herself has filled the questionnaire. If the questionnaire is directed towards the housewife, say, to know her expenditure on kitchen items, she alone is supposed to answer it. Instead, if her husband answers the questionnaire, the answer may not be correct.

- Any clarification required by the respondent regarding questions is not possible.

  Example: Prorated discount, product profile, marginal rate, etc., may not be understood by the respondents.

- If the answers are not correct, the researcher cannot probe further.
- Poor response (30%) - Not all reply.

2.7.3 Additional Consideration for the Preparation of Mail Questionnaire

- It should be shorter than the questionnaire used for a personal interview.
- The wording should be extremely simple.
- If a lengthy questionnaire has to be made, first write a letter requesting the cooperation of the respondents.
- Provide clear guidance, wherever necessary.
- Send a pre-addressed and stamped envelope to receive the reply.

2.8 EDITING AND CODING OF DATA

Between the two stages of collection of data and analysis of data there is always an intermediate stage, known as the editing of data.

The process of editing refines the collected data by checking inconsistencies, inaccuracies, illegible writings and other types of deficiencies or errors present in the collected information.

2.8.1 Editing Primary Data

Once the questionnaires have been filled and the data collected, it is necessary to edit this data to ensure completeness, consistency, accuracy and homogeneity.

- ***Completeness:*** Each questionnaire should be complete in all respects, i.e. the respondent should have answered each and every question. If some important questions have been left unanswered, attempts should be made to contact the respondent and get the response. If despite all efforts, answers to vital questions are not given, such questionnaires should be dropped from final analysis.
- ***Consistency:*** Questionnaire should be checked to see that there are no contradictory answers. Contradictory responses may arise due to wrong answers filled up by the respondent or because of carelessness on the part of the investigator in recording the data.
- ***Accuracy:*** The questionnaire should be checked for the accuracy of information provided by the respondent. This is the most difficult job of the investigator and at the same time the most important one. If inaccuracies were permitted, this would lead to misleading results. Inaccuracies may be randomly crosschecked by supervisor.
- ***Homogeneity:*** It is important to check whether all the respondents have understood the questions in the same sense. For instance, if there is a question on income, it should be very clearly stated whether it refers to weekly, monthly or yearly income and checked that the respondents have answered in the same way.

2.8.2 Editing Secondary Data

The editing of the data is a process of examining the raw data to detect errors and omissions and to correct them, if possible, so as to ensure completeness, consistency, accuracy and homogeneity. Editing can be done at two stages:

1. ***Field editing:*** The field editing consists of reviewing the interviewer's report for completeness and translating what the interviewer has written in abbreviated form at the time of interviewing the respondent. This sort of editing should be done as

soon as possible after the interview, as memory recall diminishes with time. Care should be taken that the interviewer does not complete the information by simply guessing.

2. ***Central editing:*** When all forms are filled up completely and returned to the headquarters, central editing is carried out. The editor may correct the obvious errors. If necessary, the respondent may be contacted for clarification. All the incorrect replies, which are obvious, must be deleted.

2.8.3 Coding of Data

Coding is the process of assigning some symbols either alphabetical or numeral or both to the answers so that the responses can be recorded into a limited number of classes or categories.

The classes should be appropriate to the research problem being studied. They must be exhaustive and must be mutually exclusive, so that the answer can be placed in one and only one cell in a given category. Further, every class must be defined in terms of only one concept. The coding is necessary for the efficient analysis of data. The coding decisions should usually be taken at the designing stage of the questionnaire so that the likely responses to questions are pre-coded. This simplifies computer tabulation of the data for further analysis.

2.9 CLASSIFICATION OF DATA

Classification refers to the grouping of data into homogeneous classes and categories. It is the process of arranging things in groups or classes according to their resemblances and affinities.

2.9.1 Rules of Classification

The principal rules of classifying data are:

- To condense the mass of data in such a way that salient features can be readily noticed; for example, household incomes can be grouped as higher income group, middle-income group and lower income group based on certain criterion.
- To facilitate comparison between attributes of variables; for example, comparison between education and income, income and expenditure on consumer durables, etc.
- To prepare data for tabulation.
- To highlight the significant features; for example, data is concentrated on one side, or one particular value may be dominant.
- To enable grasp of data.
- To study the relationship.

2.9.2 Bases of Classification

Some common types of bases of classification are given below:

- ***Geographical classification:*** In this type, the data is classified according to area or region; for example, state wise industrial production, city wise consumer behaviour, area wise sales figures, etc.
- ***Chronological classification:*** In this type, the data is classified according to the time of its occurrence; for example, monthly sales, yearly production, daily demands, etc.

- ***Qualitative classification:*** When the data is classified according to some attributes, which are not capable measurement, is known as qualitative classification. In dichotomous classification, an attribute is divided into two classes, one possessing the attribute and other not possessing it; for example, sex, smoker, non-smoker, employed, unemployed, etc. In many-fold classification, attribute is divided so as to form several classes; for example, education level, religion, mother tongue, etc.
- ***Classification of data according to some characteristics:*** It refers to the classification of data according to some characteristics that can be measured; for example, salary, age, height, etc. Quantitative data may be further classified into one or two types, discrete and continuous. In case of discrete type, values the variable can take are countable (could be infinitely large also for example, integers). Examples of these are number of accidents, number of defectives, etc. In case of continuous quantities, data can take any real values; for example, weight, distance, volume, etc.

2.9.3 Frequency Distribution

Classification of data, showing the different values of a variable and their respective frequency of occurrence is called a frequency distribution of the values.

There are two kinds of frequency distributions, namely, discrete frequency distribution (or simple, or ungrouped frequency distribution) and continuous frequency distribution (or condensed or grouped frequency distribution).

Discrete Frequency Distribution

The process of preparing discrete frequency distribution is simple. First, all possible values of variables are arranged in ascending order in a column. Then, another column of 'Tally' mark is prepared to count the number of times a particular value of the variable is repeated. To facilitate counting, a block of five 'Tally' marks is prepared. The last column contains frequency. To illustrate this let us consider one example.

Example: Construct frequency distribution table for the following data of number of family members in 30 families:

4 3 2 3 4 5 5 7 3 2

3 4 2 1 1 6 3 4 5 4

2 7 3 4 5 6 2 1 5 3

Solution: The discrete frequency distribution with the help of tally mark is shown below:

Number of Family Members	'Tally Marks'	Frequency
1	𝍸	3
2	𝍸	5
3	𝍸 ||	7
4	𝍸 |	6
5	𝍸	5
6	||	2
7	||	2
		Total N = 30

Continuous Frequency Distribution

For continuous data, a 'grouped frequency distribution' is necessary. For discrete data, discrete frequency distribution is better than array, but this does not condense the data. 'Grouped frequency distribution' is useful for condensing discrete data by putting them into smaller groups or classes called class-intervals. Some important terms used in case of continuous frequency distribution are as follows:

- ***Class limits:*** Class limits denote the lowest and highest value that can be included in the class. The two boundaries of class are known as the lower limit and upper limit of the class. For example, 10-19.5, 20-29.5, where 10 and 19.5 are limits of the first class; 20 and 29.5 are limits of second class, etc.
- ***Class intervals:*** The class interval represents the width (span or size) of a class. The width may be determined by subtracting the lower limit of one class from the lower limit of the following class. For example, classes 10-20, 20-30, etc. have class interval $20 - 10 = 10$.
- ***Class frequency:*** The number of observation falling within a particular class is called its class frequency. Total frequency indicates the total number of observations $N = \Sigma f$.
- ***Class mark or class mid-point:*** Mid-point of a class is defined as sum of two successive lower limits divided by 2. Thus class mark is the value lying halfway between lower and upper class limits. For example, classes 10-20, 20-30, etc. have class marks 15, 25, etc.
- ***Types of class intervals:*** There are different ways in which limits of class intervals can be shown.
  - *Exclusive method:* The class intervals are so arranged that upper limit of one class is the lower limit of next class. This method always presumes that the upper limit is excluded from the class, for example, with class limits 20-25, 25-30 observation with value 25 is included in class 25-30.
  - *Inclusive method:* In this method, the upper limit of the class is included in that class itself. In such case, there is no overlap of upper limit of former class and lower limit of successive class. For example, with class limits 20-29.5, 30-39.5, 40-49.5, etc. there is no ambiguity but values from 29.5 to 30 or 39.5 to 40, etc. are not allowed.
  - *Open end:* In an open-end distribution, the lower limit of the very first class and/or upper limit of the last class is not given. For example, while stating the distribution of monthly salary of managers in rupees, one may specify class limits as, below 15000, 15000-25000, 25000-35000, 35000-45000, above 45000. Similarly, while recording weights of college students in kg as grouped data the class intervals could be less than 50, 50 to 60, 60 to 70, 70 to 80, 80 to 90 and greater than 90.
  - *Unequal class interval:* This is another method to limit the class intervals where the width of the classes is not equal for all classes. This method is of practical use when there are large gaps in the data, or distribution of the data is uneven. It is used for explaining, visualizing and plotting data with unequal class interval. However, we must adjust formulae for calculations accordingly.

Guideline for Choosing the Class

- Number of classes should not be too small or too large, preferably between 5 and 15.
- If possible, the widths of the intervals should be numerically simple like 5, 10, 15, etc.

- It is desirable to have classes of equal width.
- Starting point of class should begin with 0, 5, 10 or multiple thereof.
- Class interval should be determined based on maximum values and number of classes to be formed.

All the above points can be explained with the help of the following example.

Example: Ages of 50 employees are given:

22 21 37 33 28 42 56 33 32 59

40 47 29 65 45 48 55 43 42 40

37 39 56 54 38 49 60 37 28 27

32 33 47 36 35 42 43 55 53 48

29 30 32 37 43 54 55 47 38 62

Prepare a frequency distribution table.

Solution: A frequency distribution table is prepared as follows:

- First, find the highest and lowest values. These are 65 and 21 respectively. Thus, the difference is 44.
- Since the total observations are 50 we decide to select 5 classes.
- The approximate class interval works out to be (65-21)/5 = 8.8. Hence, we select class interval as 10.
- As our lowest value is 21, we start from the lower class limit of the first class as 20. We use exclusive method of class interval.
- We then decide class intervals as 20-30, 30-40, 40-50, 50-60 and 60-70.
- Then, each observation is checked for the class interval in which it lies. For each observation, we make a tally mark against the corresponding class interval. As per the convention, every fifth tally is put horizontally across. This helps quick counting.

The frequency distribution is given below:

Age (Years)

Class Interval	Class Mark	Tally	Frequency
20-30	25	𝍸 ||	7
30-40	35	𝍸 𝍸 𝍸 |	16
40-50	45	𝍸 𝍸 𝍸	15
50-60	55	𝍸 |||	9
60-70	65	|||	3
			Total = 50

Cumulative and Relative Frequency

In many situations rather than listing the actual frequency opposite each class, it may be appropriate to list either cumulative frequencies or relative frequencies or both.

Cumulative Frequencies

The cumulative frequency of a given class interval thus, represents the total of all the previous class frequencies including the class against which it is written.

Relative Frequencies

Relative frequency is obtained by dividing the frequency of each class by the total number of observations (total frequency).

If we multiply relative frequency by 100, we get percentage frequency.

There are two important advantages in looking at relative frequencies (percentages) instead of the absolute frequencies in a frequency distribution. These are as follows:

1. Relative frequencies facilitate the comparison of two or more than sets of data.
2. Relative frequencies constitute the basis of understanding the concept of probability.

To explain the cumulative and relative frequencies we work these on our earlier problem.

Example: Ages of 50 employees are given:

22 21 37 33 28 42 56 33 32 59
40 47 29 65 45 48 55 43 42 40
37 39 56 54 38 49 60 37 28 27
32 33 47 36 35 42 43 55 53 48
29 30 32 37 43 54 55 47 38 62

Find cumulative frequency, relative frequency and percentage frequency.

Solution:

Class interval	Class Frequency	Cumulative Frequency	Relative Frequency	Percentage Frequency
20-30	7	(0+7) = 7	7/50 = 0.14	14
30-40	16	(7+16) = 23	16/50 = 0.32	32
40-50	15	(23+15) = 38	15/50 = 0.30	30
50-60	9	(38+9) = 47	9/50 = 0.18	18
60-70	3	(47+3) = 50	3/50 = 0.06	6
	$N = \sum f = 50$		**Total = 1**	**Total = 100**

A frequency distribution is constructed to satisfy three objectives: (i) to facilitate the analysis of data, (ii) to estimate frequencies of the unknown population distribution from the distribution of sample data, and (iii) to facilitate the computation of various statistical measures.

Frequency distribution can be of two types: (1) Univariate Frequency Distribution and (2) Bivariate Frequency Distribution.

2.10 TABULATION OF DATA

Once the raw data is collected, it needs to be summarized and presented to the decision-maker in a form that is easy to comprehend. The manager must be able to look at the data so as to decide what further analysis is required. Tabulation helps this process through effective presentation. Tabulation is arranging the data in flat table (two dimensional arrays) format by grouping the observations. Table is a spreadsheet with rows and columns with headings and stubs indicating class of the data. Tabulation not only condenses the data, but also makes it easy to understand.

Tabulation is the fastest way to extract information from the mass of data and hence popular even among those not exposed to the statistical method. The report card of a school is the most common example.

2.10.1 Objectives of Tabulation

The main objectives of tabulation are:

- To simplify complex data.
- To highlight chief characteristics of the data.
- To clarify objective of investigation.
- To present data in a minimum space.
- To detect errors and omissions in the data.
- To facilitate comparison of data.
- To facilitate reference.
- To identify trend and tendencies of the given data.
- To facilitate statistical analysis.

2.10.2 Main Parts of a Table

The main parts of a table are given below:

- ***Table Number:*** This number is helpful in the identification of a table. This is often indicated at the top of the table.
- ***Title:*** Each table should have a title to indicate the scope, nature of contents of the table in an unambiguous and concise form.
- ***Captions and stubs:*** A table is made up of rows and columns. Headings or subheadings used to designate columns are called captions while those used to designate rows are called stubs. A caption or a stub should be self-explanatory. A provision of totals of each row or column should always be made in every table by providing an additional column or row respectively.
- ***Main Body of the Table:*** This is the most important part of the table as it contains numerical information. The size and shape of the main body should be planned in view of the nature of figures and the objective of investigation. The arrangement of numerical data in main body is done from top to bottom in columns and from left to right in rows.
- ***Ruling and Spacing:*** Proper ruling and spacing is very important in the construction of a table. Vertical lines are drawn to separate various columns with the exception of sides of a table. Horizontal lines are normally not drawn in the body of a table; however, the totals are always separated from the main body by horizontal lines. Further, the horizontal lines are drawn at the top and the bottom of a table.

  Spacing of various horizontal and vertical lines should be done depending on the available space. Major and minor items should be given space according to their relative importance.
- ***Head-note:*** A head-note is often given below the title of a table to indicate the units of measurement of the data. This is often enclosed in brackets.
- ***Foot note:*** Abbreviations, if any, used in the table or some other explanatory notes are given just below the last horizontal line in the form of footnotes.

- ***Source-Note:*** This note is often required when secondary data are being tabulated. This note indicates the source from where the information has been obtained. Source note is also given as a footnote.

Example: The main parts of a table can also be understood by looking at its broad structure given below:

Structure of a table

Table No: .....

Title: .....

Stub Heading	Captions		Captions			Total
	Captions	Captions	Captions	Captions	Captions	
↑ Stub Enteries ↓	MAIN BODY					
Total						

Foot Note:

Source:

2.10.3 Rules for Tabulation

Now, let us learn about the general rules of tabulation.

- The table should be simple and compact which is contains simple details.
- Tabulation should be in accordance with the objective of investigation.
- The unit of measurements must always be indicated in the table.
- The captions and stubs must be arranged in a systematic manner so that it is easy to grasp the table.
- A table should be complete and self-explanatory.
- As far as possible the interpretative figures like totals, ratios and percentages must also be provided in a table.
- The entries in a table should be accurate.
- Table should be attractive to draw the attention of readers.

2.10.4 Types of Tabulation

Statistical tables can be classified into various categories depending upon the basis of their classification. Broadly speaking, the basis of classification can be any of the following:

- Purpose of investigation
- Nature of presented figures
- Construction

Different types of tables, thus, obtained are shown in the following chart.

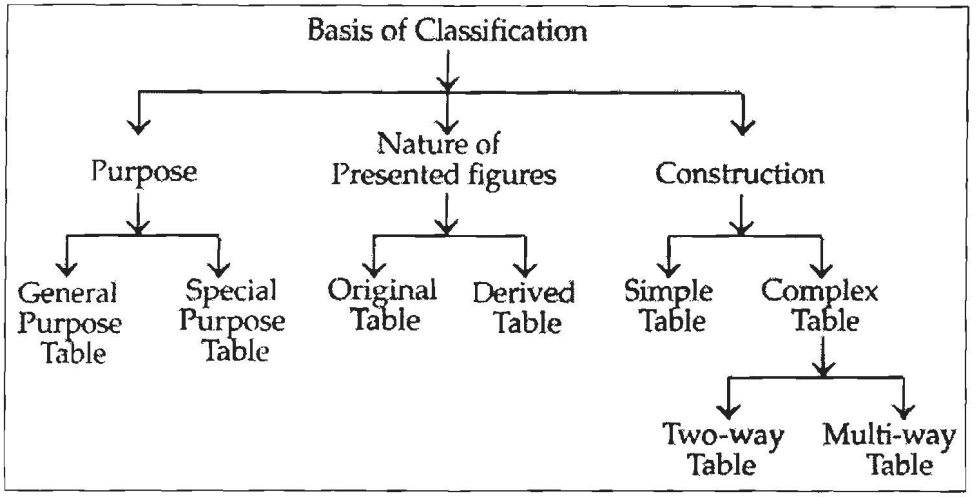


Figure 2.2: Classification of Table

- ***Classification on the basis of purpose of investigation:*** These tables are of two types viz. General purpose table and Special purpose table.

  - *General purpose table:* A general purpose table is also called as a reference table. This table facilitates easy reference to the collected data. In the words of Croxton and Cowden, "The primary and usually the sole purpose of a reference table are to present the data in such a manner that the individual items may be readily found by a reader." A general purpose table is formed without any specific objective, but can be used for a number of specific purposes. Such a table usually contains a large mass of data and is generally given in the appendix of a report.

  An example of general purpose table is as follows:

Table 2.1: Report Forms Name Codes for General Purpose Reports of Gint

Position	Description	Values
1.	Type	LOG (or L) - Log FNC (or F) - Fence GRF (or G) - Graph GTB (or T) - Graphical Table GTD (or X) - Graphical Text Document HST (or H) - Histogram TTB (or A) - Text Table TXD (or D) - Text Document SMP - Site Map
2.	Paper Size	A - US Letter 4 - ISO A4 B - US 11x17 3 - ISO A3 L - US Legal

Contd...

3.	Content	G - Predominantly Geotechnical E - Predominantly Environmental R - Predominantly Rock Core N - Not Applicable or can be generally used
4.	Well	W - Has well (applies to logs & fences) N - No well or Not Applicable
5.	Graph	G - Has graph (applies to logs & fences) N - No graph or Not Applicable
6.	Legend	L - Has legend N - No legend or Not applicable
7.	Counter	

- *Special purpose table:* A special purpose table is also called a text table or a summary table or an analytical table. Such a table presents data relating to a specific problem. According to H. Secrist, "These tables are those in which are recorded, not the detailed data which have been analyzed, but rather the results of analysis." Such tables are usually of smaller size than the size of reference tables and are generally found to highlight relationship between various characteristics or to facilitate their comparisons.

- ***Classification on the basis of the nature of presented figures:*** Tables, when classified on the basis of the nature of presented figures can be Primary table and Derivative table.

  - *Primary Table:* Primary table is also known as original table and it contains data in the form in which it were originally collected.

  - *Derivative Table:* A table which presents figures like totals, averages, percentages, ratios, coefficients, etc., derived from original data. A table of time series data is an original table but a table of trend values computed from the time series data is known as a derivative table.

- ***Classification on the basis of construction:*** Tables when classified on the basis of construction can be Simple table, Complex table and Cross-classified table.

  - *Simple Table:* In this table, the data are presented according to one characteristic only. This is the simplest form of a table and is also known as table of first order.

    The following blank table, for showing the number of workers in each shift of a company, is an example of a simple table.

Shifts	No. of Workers
I	
II	
III	
Total	

  - *Complex Table:* A complex table is used to present data according to two or more characteristics. Such a table can be two-way, three-way or multi-way, etc.

- Two-way table: Such a table presents data that is classified according to two characteristics. In such a table, the columns of a table are further divided into sub-columns.

  The example of such a table is given below.

Shifts	No. of Workers		Total
	Males	Females	
I			
II			
III			
Total			

- Three-way table: When three characteristics of data are shown simultaneously, we get a three-way table as shown below in the example.

Shifts	No. of Workers						Total No. of Workers
	Males		Total	Females		Total	
	Skilled	Unskilled		Skilled	Unskilled		
I							
II							
III							

- Multi-way table: If each shift is further classified into three departments, say, manufacturing, packing and transportation, we shall get a four-way table, etc.

❖ *Cross-classified Table:* Tables that classify entries in both directions, i.e., row-wise and column-wise, are called cross-classified tables. The two ways of classification are such that each category of one classification can occur with any category of the other. The cross-classified tables can also be constructed for more than two characteristics also. A cross-classification can also be used for analytical purpose, e.g., it is possible to make certain comparisons while keeping the effect of other factors as constant.

Example: Draw a blank table to show the population of a city according to age, sex and unemployment in various years.

Years	Age Sex	Population (in thousands)							
		Employed				Unemployed			
		Below 20	20-60	60 and above	Total	Below 20	20-60	60 and above	Total
1991	Males								
	Females								
	Total								
1992	Males								
	Females								
	Total								

(Note: The table can be extended for the years 1993, 94, 95, 96, etc.)

Example: In a sample study about coffee habit in two towns; the following information were received:

Town A: Females were 40%; total coffee drinkers were 45%; and male non-coffee drinkers were 20%.

Town B: Males were 55%; male non-coffee drinkers were 30%; and female coffee drinkers were 15%.

Represent the above data in a tabular form.

Solution:

The figures are in percentage

Habit	Town A			Town B		
	Males	**Females**	**Total**	**Males**	**Females**	**Total**
Coffee Drinkers	40	5	45	25	15	40
Non-Coffee Drinkers	20	35	55	30	30	60
Total	**60**	**40**	**100**	**55**	**45**	**100**

2.10.5 One-way Tabulation

Tabulation is primarily counting how many observations are in a particular category. Tabulation is like an in-process inventory. Tabulation in itself may not be the end of statistical processing. It may be noted that once we tabulate the data, we usually do not go back to the raw data. Any improper tabulation would definitely mislead decision-maker for further processing. Hence, before tabulation manager must give sufficient thought to decide what kind of tabulation is required for decision-making. We need to first decide characteristics, their values and ranges, title of the table, stubs for the rows, headings for the columns, scale and dimensions used, foot notes, pivots if we need, etc. Table must suit the purpose for which the data is being processed. We also need to decide on the size of the table, clarity, approximations, boundaries, appearance, order, readability, etc. A meaningful title not only helps the manager to focus on the purpose, and thus, group the data properly but also others who refer the table later. The next step is to decide appropriate column headings; row stubs units and dimensions of the quantities used, labels for summary figures, etc. to improve the readability of the table. Many times the requirement of statistical analysis is to count the frequency of the distinct value of a variable. When we arrange the range of values (or just values) and their frequencies the tabulation is known as one way. Variable could be either quantitative or normative.

For example, examination result of MBA could be tabulated as,

Class	Number of Students (Frequency f)
Distinction (≥75%)	26
First Class (60-75%)	72
Second Class (50-60%)	94
Pass Class (40-50%)	42
Fail	16
Total Students Appeared	**250**

Foot Note:

- Each class includes its lower limit.
- Fail indicates failure in any one or more subjects irrespective of the percentage marks.

Example: Represent the following information in a neat table:

The number of students in a college in the year 1961 was 1100; of those 980 were boys and rest girls. In 1971, the number of boys increased by 100% and that of girls increased by 300% as compared to their strength in 1961. In 1981, the total number of students in a college was 3600, the number of boys being double the number of girls.

Solution:

Years	Number of Boys	Number of Girls	Total Students
1961	980	120	1100
1971	1960	480	2440
1981	2400	1200	3600

2.10.6 Two-way Tabulation

There are occasions that we want to summaries the frequency table against two attributes (categories) and want the count of the same population belonging to all possible combinations of these two attributes. For example, we want to know the frequency of personnel with different combinations of salary earned category and education qualification category for a given company. Since there are two variables, we call it a two-way tabulation (also referred to as cross-tabulation). We prepare the table with one of the category varied along the rows and other along the columns. For counting the frequency, a pair of combinations of categories one from each direction is considered. Thus, we get a table in m × n matrix form, with each cell containing data for one combination. This is also knows as contingency table. With m rows and n columns, we get m categories of one variable varying along column and n categories of another variable varying along row. There are obviously mn cells containing distinct mutually exclusive and collectively exhaustive data. It may be noted that, a two way table can be converted to one way table with mn distinct values of a combination variable. This is called a normalized table or a flat table in data base management.

Example: In a survey conducted in a city about preference of Coke or Pepsi or Maza, the sample consisted of 400 people that included 150 women and 250 men. It was observed that 50 women preferred Coke and 40 preferred Pepsi. In case of men, the preference was 100, 80 and 70 respectively. Present the information in two way table and answer the following:

(a) What is the percentage of men in Coke preferring population?

(b) What is the proportion of population preferring Pepsi?

(c) What is the proportion of women preferring Maza in total population?

Solution:

	Men	Women	Total
Coke Preferring People	100	50	150
Pepsi Preferring People	80	40	120
Maza Preferring People	70	60	130
Total	**250**	**150**	**400**

(a) Percentage of men in Coke preferring population $= \frac{100}{150} \times 100 = 66.67\%$

(b) Proportion of population preferring Pepsi $= \frac{120}{400} = 0.3$

(c) Proportion of women preferring Maza in total population = 0.15

2.10.7 Multi-way Tabulation

We can carry out cross tabulation with more than two variables. It is called a nested table. In fact, in most of the business situations the tabulation may have more than two variables (usually 10 to 15). Up to about 3 to 4 variables could be shown on two dimensional papers. These can also be represented as flat tables by taking one composite variable of dimension $n_1 \times n_2 \times n_3 \times n_4 \times n_5 \times \ldots$, where $n_1, n_2, n_3, n_4, n_5 \ldots$ are dimensions of each variable (attribute). Obviously, the number grows so rapidly, that it becomes too voluminous and complex to get any meaningful information for decision-making. However, that does not mean such multidimensional data is not tabulated. It is tabulated using computer database like MS Access, FOXPRO, Oracle, etc. We cannot view it together but definitely use it for the decision-making through 'query language'. Data base management systems and query languages are beyond the scope of this book. One simple, three-dimensional, tabulation is shown in the following example.

Example: A mutual fund wants to compare the performance of shares on NSE over past three years. It wants to categorize the shares as below average, average and above average as compared to the benchmark. It also wants to group the shares as large cap, mid-cap and small cap. The data obtained is as follows: In 40 large cap shares studied 27 performed average and 11 above average in year 2004. Similar, figures for year 2005 and 2006 were 34 and 8 out of 50, and 32 and 16 out of 50 respectively. In mid-cap segment, the number of shares below average, average and above average was 22, 35 and 23 in year 2004. These were 17, 40, 23 for year 2005 and 13, 38 and 29 for year 2006 respectively. In case of small cap shares, the performance figures for year 2004, 2005 and 2006 in categories below average, average and above average were 26, 32, 42; 25, 36, 39; and 12, 40, 48 respectively. Present the data as multi-way table.

Solution:

Year		2004	2005	2006
Large Cap	Below Average	12	8	2
	Average	27	34	32
	Above Average	11	8	16
	Total	**50**	**40**	**40**
Mid Cap	Below Average	22	17	13
	Average	35	40	38
	Above Average	23	23	29
	Total	**80**	**80**	**80**
Small Cap	Below Average	26	25	12
	Average	32	36	40
	Above Average	42	39	48
	Total	**100**	**100**	**100**

2.10.8 Advantages of Tabulation

Tabulation helps to achieve the following:

- It presents the data in easy to understand format.
- It reduces the voluminous size of data so as to view it in comprehensive way.
- It simplifies the data through grouping.
- It tries to highlight common features, salient points, characteristics, etc. from the data.
- Reveals underlying trends.
- It allows easy comparison within the data or with other tabulated data.
- Data storage, reference and retrieval at later stage are very easy.
- Processing the data through spreadsheet packages like MS Excel can be done.
- Charting of graphs and diagrams is easy with tabulated data.

Check Your Progress

Fill in the blanks:

1. When the data is not collected for this purpose, but is derived from other sources then such data is referred to as _____.
2. _____ are questions where respondents are free to answer in their own words.
3. Hypotheses and significance tests form an important part of _____ statistics.
4. _____ is obtained by dividing the frequency of each class by the total number of observations.
5. A _____ table is used to present data according to two or more characteristics.
6. A _____ table presents data relating to a specific problem.

2.11 LET US SUM UP

- There are two major divisions of the field of statistics, namely descriptive and inferential statistics. Both the segments of statistics are important and accomplish different objectives.
- Data can be obtained through primary source or secondary source according to need, situation, convenience, time, resources and availability. The most important method for primary data collection is through questionnaire. Data must be objective and fact-based so that it helps a decision-maker to arrive at a better decision.
- Statistical data is a set of facts expressed in quantitative form. Data is collected through various methods. Sometimes our data set consists of the entire population we are interested in. In other situations, data may constitute a sample from some population.
- Type of research, its purpose, conditions under which the data are obtained will determine the method of collecting the data. If relatively few items of information are required quickly, and funds are limited telephonic interviews are recommended. If respondents are industrial clients Internet could also be used.

If depth interviews and probing techniques are to be used, it is necessary to employ investigators to collect data.

- The quality of information collected through the filling of a questionnaire depends, to a large extent, upon the drafting of its questions. Hence, it is extremely important that the questions be designed or drafted very carefully and in a tactful manner.
- Before any processing of the data, editing and coding of data is necessary to ensure the correctness of data. In any research studies, the voluminous data can be handled only after classification. Data can be presented through tables and charts.
- Classification refers to the grouping of data into homogeneous classes and categories. It is the process of arranging things in groups or classes according to their resemblances and affinities.
- A frequency distribution is the principle tabular summary of either discrete data or continuous data. The frequency distribution may show actual, relative or cumulative frequencies. Actual and relative frequencies may be charted as either histogram (a bar chart) or a frequency polygon. Two commonly used graphs of cumulative frequencies are less than ogive or more than ogive.
- Once the raw data is collected, it needs to be summarized and presented to the decision-maker in a form that is easy to comprehend. Tabulation not only condenses the data, but also makes it easy to understand. Tabulation is the fastest way to extract information from the mass of data and hence popular even among those not exposed to the statistical method.

2.12 UNIT END ACTIVITY

In any organization of your choice, identify a problem and collect a data internally through questionnaire from randomly selected people of your organization. Using the collected data, present it in tabulated form and finds a solution to the problem.

2.13 KEYWORDS

Primary Data: Primary data are collected afresh and for the first time, and thus, happen to be original in character.

Secondary Data: When the data are not collected for this purpose, but is derived from other sources then such data is referred to as 'secondary data'.

Frequency Distribution: A tabular summary of data showing the number (frequency) of observations in each of several non-overlapping classes.

Tabulation: Tabulation is arranging the data in flat table (two dimensional arrays) format by grouping the observations.

2.14 QUESTIONS FOR DISCUSSION

1. Differentiate between descriptive and inferential statistics with examples.
2. Describe the various methods of collecting primary data and comment on their relative advantages and disadvantages.
3. Discuss the methods or sources of collecting secondary data.
4. How do you design a questionnaire? What are the important points to be kept in mind?

5. How is editing of primary and secondary data done? Also, describe coding of data.
6. Describe the classification of data. What are the rules and bases of classification of data?
7. What is frequency distribution? Differentiate between discrete and continuous frequency distribution with examples.
8. Discuss the concept of tabulation. What are objectives and main parts of table?
9. Differentiate among one-way tabulation, two-way tabulation and multi-way tabulation with examples.
10. Discuss the various stages of statistical investigation.

Check Your Progress: Model Answer

1. Secondary data
2. Open-ended Questions
3. Classification
4. Relative frequency
5. Complex
6. Special purpose

2.15 REFERENCE & SUGGESTED READINGS

- Groebner, D. F., Shannon, P. W., Fry, P. C., & Smith, K. D. (2022). **Business Statistics: A Decision-Making Approach** (11th ed.). Pearson. ISBN: 9780136681503
- Albright, S. C., Winston, W. L., & Zappe, C. (2021). **Data Analysis and Decision Making** (6th ed.). Cengage Learning. ISBN: 9780357131785
- Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2020). **Statistics for Business and Economics** (14th ed.). Cengage Learning. ISBN: 9780357114474
- Jaggia, S., Kelly, A., &Lertwachara, K. (2021). **Essentials of Business Statistics** (2nd ed.). McGraw-Hill Education. ISBN: 9781260205799

UNIT - III

SAMPLING METHODS AND STATISTICAL SERIES

CONTENTS

Contd...

3.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Understand the meaning, concepts and types of sampling design
- Discuss the concept of probability and non-probability sampling methods
- Understand sampling and sampling techniques
- Discuss the method of determination of sample size
- Explain the sampling distributions
- Describe the occurrence of sampling error
- Understand the preparation of statistical series and its types

3.1 INTRODUCTION

The formula for the sampling distribution depends on the distribution of the population, the statistic being considered and the sample size used. A more precise formulation would speak of the distribution of the statistic as that for all possible samples of a given size, not just "under repeated sampling". Suppose that we draw all possible samples of size n from a given population. Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. The probability distribution of this statistic is called a sampling distribution. Estimation is a procedure by which sample information is used to estimate the numerical magnitude of one or more parameters of the population. A function of sample values is called an estimator (or statistic) while its numerical value is called an estimate. For example, an estimator of population mean is m. On the other hand, if for a sample, the estimate of population mean is said to be 50.

3.2 SAMPLING DESIGN

A sample is a part of a target population, which is carefully selected to represent the population.

Sampling frame is the list of elements from which the sample is actually drawn. Actually, sampling frame is nothing but the correct list of population.

Example: Telephone directory, Product finder and Yellow pages.

3.2.1 Distinction between Census and Sampling

Census refers to complete inclusion of all elements in the population. A sample is a sub-group of the population.

When is a Census Appropriate?

- A census is appropriate if the size of population is small. For example, a researcher may be interested in contacting firms in iron and steel or petroleum products industry. These industries are limited in number, so a census will be suitable.
- Sometimes, the researcher is interested in gathering information from every individual.

  Example: Quality of food served in a mess.

3.2.2 Meaning

It is not possible, nor it is necessary, to collect information from the total population. Instead, a smaller sub-group of the target population or a sample is selected for the purpose of study. Sampling is the strategy of selecting a smaller section of the population that will accurately represent the patterns of the target population at large.

When is Sample Appropriate?

- When the size of population is large.
- When time and cost are the main considerations in research.
- If the population is homogeneous.
- Also, there are circumstances when a census is not possible.

  Example: Reactions to global advertising by a company.

3.2.3 Concepts

In carrying out a survey relating to the research, we should first select the problem and study its implications in different areas. Selection of the research problem, as has already been stated, should be in line with the researcher's interest, chain of thinking and existing research in the same area and should have some direct utility. What is most important in selecting a research problem is that the research topic should be within manageable limits.

Secondly, the topic should have practical feasibility. To study feasibility, what is important is to prepare a preliminary abstract on the research topic. Since this lesson is intended to acquaint the readers with survey procedure, we are not concentrating on the aspects of research in great detail which have in fact already been covered in our earlier discussions.

The first and foremost task in carrying out a survey is to select the sample. The difference between the population and sample has already been discussed earlier. Sample selection is undertaken for practical impossibility to survey the population. By applying rationality in selection of samples, we generalize the findings of our research. There are different types of sampling. We may categorize those in three major heads as follows:

1. Random Sampling
2. Purposive Sampling
3. Stratified Sampling

Random sampling is not a mere chance selection. Instead, it ensures inclusion of each and every sample of the population. The conventional way of selection of samples using random sampling methods are:

- Lottery method
- Tippet's number
- Selection from a sequential list
- Use of Grid System

Under lottery method, numbers or names of various units of population are noted on chits and put in a container. After thorough mixing, chits are drawn from the container and survey of drawn chits is carried out. Since this method of random sampling has some amount of chance in it, this is often described as a back-dated one.

Tippet's number which lists 10,400 four digit numbers written at random is constructed out of 41,600 digits taken from census reports by combining them in to fours. The method of drawing a sample from Tippet's number is very easy. If we want to draw a sample of 20 persons from a list of 6000 persons, for this purpose we shall first number each unit from 0 to 6000 using Tippet's four digit codes. Then we open any page of Tippet's numbers and select the first 20 numbers that are below 6000. Tippet's numbers are widely used in sampling techniques and are found to be quite reliable in regard to accuracy and representativeness.

Selection of sample from sequential list requires arrangement of names under the intended plan according to some order which may be alphabetical, geographical or simply serial. Thereafter, out of the list, every 10th or any other number of cases may be taken up. If every 10th unit is to be selected, the selection might begin from 7th, 17th, 27th, 37th, 47th, 57th, etc., or from 5th, 15th, 35th, 45th, 55th, etc.

Grid system is applied for selection of sample from a particular area. Under this method, a map of the entire area is prepared, and then a screen of squares is placed on the map. The areas falling within selected squares are taken as samples.

Purposive selection of samples, as the name goes, depends more on the researcher's deliberate choice. Thus, such a selection of samples, in its true spirit defeats the purpose of research as the samples suffer from the character of representativeness.

Stratified sampling combines the characteristics of random sampling and purposive sampling. Initially, the population is defined in different numbers of strata or groups. Then from each group certain number of items is taken on a random basis.

Apart from the above sampling procedures, there are other types of sampling like:

- Quota sampling (a special type of stratified sampling).
- Multi-stage sampling (where samples are selected from a very large area).
- Convenience sampling (where population is not clearly defined and complete source of list is not available).
- Self-selected sampling, etc.

After deciding over the samples to be surveyed, the next task is to go ahead with the survey matter.

Survey may be carried out either by directly interviewing the samples or by sending questionnaire to the samples or by mere observation of the characteristics of samples.

3.2.4 Types of Sampling Design

Sampling is divided into two types:

1. ***Probability sampling:*** In a probability sample, every unit in the population has equal chances for being selected as a sample unit.
2. ***Non-probability sampling:*** In the non-probability sampling, the units in the population have unequal or negligible, almost no chances for being selected as a sample unit.

3.3 IMPORTANT TERMS IN SAMPLING

Firstly, let us understand few basic definitions of sampling.

Sampling: It indicates the selection of a part of the whole with a view to obtain information about the whole.

Population: The aggregate or totality of all members is known as 'population'; for example, all items produced in a day at a factory is a population; few items selected for testing or performance measurement by quality control inspector is a sample.

Sample: The selected small part of the whole, which is used to ascertain the characteristics of the population is called sample.

Member (element): Individual units that form the population are called members or elements. Some of the members are included in the sample.

Population size (N): The total number of members of the population is called population size and denoted by '*N*'.

Sample size ('n'): The number of members selected in the sample is the sample size and denoted by 'n'.

Random sample: A random sample of 'n' elements is a sample selected from the entire population N such that each element (or member) has an equal chance (or probability) of getting selected. Thus, in random sampling, probability p of an element getting selected is (1/N).

Sampling frame: It is a list of all the units of the population. The preparation of a sampling frame is many times, a major formidable practical problem. The frame should be up-to-date and free from omission and duplication.

Measures of characteristic: In inferential statistics we are primarily concerned with population. The samples are of least interest to us in their own right. We are interested to know presence or absence of certain characteristic in the population. We measure the characteristic in a sample in the form of summary statistics, only to draw inference about population characteristic unknown to us. Some of these measures are the mean, the standard deviation, the proportion, etc., about certain characteristic the members possess.

Parameters: A numerical measure (or value) of characteristic of the population is called a 'parameter'.

Statistic: A numerical measure of characteristic of the sample is called a 'statistic'.

3.4 SAMPLING PROCESS

Sampling process consists of seven steps. They are as follows:

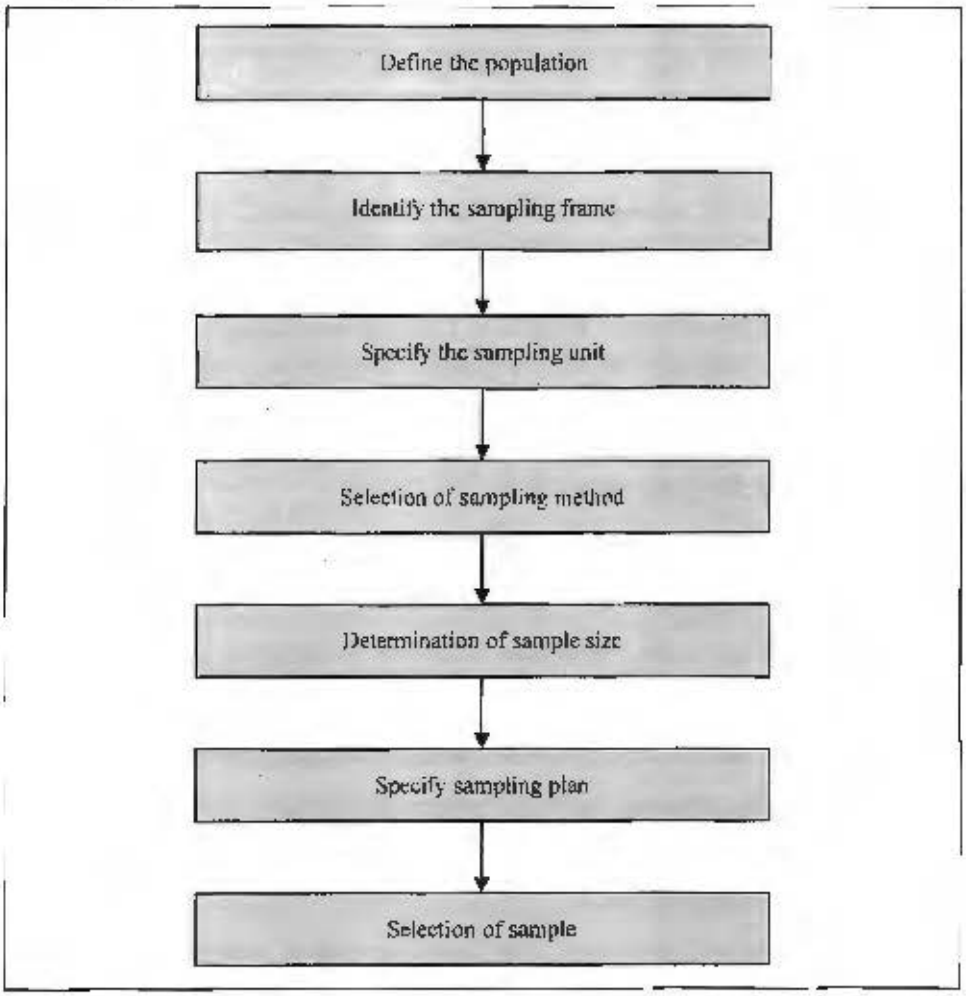


Figure 3.1: Sampling Process

Step 1: Define the Population

Population is defined in terms of:

1. Elements
2. Sampling units
3. Extent
4. Time

Example: If we are monitoring the sale of a new product recently introduced by a company, say (shampoo sachet) the population will be:

1. Element - Company's product
2. Sampling unit - Retail outlet, super market
3. Extent - Hyderabad and Secunderabad
4. Time - April 10 to May 10, 2006

Step 2: Identify the Sampling Frame

Sampling frame could be (a) Telephone Directory, (b) Localities of a city using the municipal corporation listing and (c) Any other list consisting of all sampling units.

Example: You want to learn about scooter owners in a city. The RTO will be the frame, which provides you names, addresses and the types of vehicles possessed.

Step 3: Specify the Sampling Unit

Individuals who are to be contacted are the sampling units. If retailers are to be contacted in a locality, they are the sampling units.

Sampling unit may be husband or wife in a family. The selection of sampling unit is very important. If interviews are to be held during office timings, when the heads of families and other employed persons are away, interviewing would under-represent employed persons, and over-represent elderly persons, housewives and the unemployed.

Step 4: Selection of Sampling Method

This refers to whether (a) probability or (b) non-probability methods are used.

Step 5: Determine the Sample Size

This means we need to decide "how many elements of the target population are to be chosen?" The sample size depends upon the type of study that is being conducted. For example, if it is an exploratory research, the sample size will be generally small. For conclusive research, such as descriptive research, the sample size will be large.

The sample size also depends upon the resources available with the company. It depends on the accuracy required in the study and the permissible errors allowed.

Step 6: Specify the Sampling Plan

A sampling plan should clearly specify the target population. Improper defining would lead to wrong data collection.

Example: This means that, if a survey of a household is to be conducted, a sampling plan should define a "household" i.e., "Does the household consist of husband or wife or both", minors etc., "Who should be included or excluded". Instructions to the interviewer should include "How he should obtain a systematic sample of households, probability sampling non-probability sampling". Advise him on what he should do to the household, if no one is available.

Step 7: Select the Sample

This is the final step in the sampling process. Based on the above parameters sample respondents may be selected to collect the data for the purpose of research.

3.5 PROBABILITY SAMPLING METHODS

In probability sampling, the decision whether a particular element is included in the sample or not, is governed by chance alone. All probability-sampling designs ensure that each element in the population has same non-zero probability of getting included in the sample. This requires defining a procedure for picking up the sample based on chance. In this method, we must avoid changes in the sampling processes, except the predefined ones. The picking up of the sample is therefore totally insulated against the judgment, convenience or whims of any person involved with the study. When probability-sampling designs are used, it is possible to quantify the magnitude of the

likely error in inference made. This is a great help in many situations in building up confidence in the inference.

The following are the types of probability sampling methods:

1. Random sampling
2. Systematic sampling
3. Stratified random sampling
4. Cluster sampling
5. Multi-stage sampling

3.5.1 Random Sampling

Simple random sample is a process in which every item of the population has an equal probability of being chosen.

There are two methods used in the random sampling:

1. ***Lottery method:*** Take a population containing four departmental stores: A, B, C and D. Suppose we need to pick a sample of two stores from the population using a simple random procedure. We write down all possible samples of two. Six different combinations, each containing two stores from the population, are AB, AD, AC, BC, BD and CD. We can now write down six sample combination on six identical pieces of paper, fold the piece of paper so that they cannot be distinguished. Put them in a box. Mix them and pull one at random. This procedure is the lottery method of making a random selection.
2. ***Using random number table:*** A random number table consists of a group of digits that are arranged in random order, i.e., any row, column or diagonal in such a table contains digits that are not in any systematic order. There are three tables for random numbers:

   (a) Tippet's table

   (b) Fisher and Yates's table

   (c) Kendall and Raington table

   The table for random number is as follows:

40743	39672
80833	18496
10743	39431
88103	23016
53946	43761
31230	41212
24323	18054

   Example: Taking the earlier example of stores. We first number the stores.

   1 A 2 B 3 C 4 D

   The stores A, B, C and D have been numbered as 1, 2, 3 and 4.

We proceed as follows, in order to select two shops out of four randomly:

Suppose, we start with the second row in the first column of the table and decide to read diagonally. The starting digit is 8. There are no departmental stores with the

number 8 in the population. There are only four stores. Move to the next digit on the diagonal, which is 0. Ignore it, since it does not correspond to any of the stores in the population. The next digit on the diagonal is 1 which corresponds to store A. Pick A and proceed until we get two samples. In this case, the two departmental stores are 1 and 4. The sample derived from this consists of departmental stores A and D.

In random sampling, there are two possibilities (1) Equal probability and (2) Varying probability.

- ***Equal Probability:*** This is also called as the random sampling with replacement. For example, put 100 chits in a box numbered 1 to 100. Pick one number at random. Now the population has 99 chits. Now, when a second number is being picked, there are 99 chits. In order to provide equal probability, the sample selected is being replaced in the population.
- ***Varying Probability:*** This is also called random sampling without replacement. Once a number is picked, it is not included again. Therefore, the probability of selecting a unit varies from the other. In our example, it is 1/100, 1/99, 1/98, 1/97 if we select four samples out of 100.

3.5.2 Systematic Random Sampling

There are three steps:

1. Sampling interval K is determined by the following formula:

$$K = \frac{\text{No. of units in the population}}{\text{No. of units desired in the sample}}$$

2. One unit between the first and K^{th} unit in the population list is randomly chosen.
3. Add K^{th} unit to the randomly chosen number.

Example: Consider 1,000 households from which we want to select 50 units.

$$\text{Calculate } K = \frac{1000}{50} = 20$$

To select the first unit, we randomly pick one number between 1 and 20, say 17. So our sample begins with 17, 37, 57..... Please note that only the first item was randomly selected. The rest are systematically selected. This is a very popular method because we need only one random number.

3.5.3 Stratified Random Sampling

A probability sampling procedure is in which simple random sub-samples are drawn from within different strata that are, more or less equal on some characteristics. Stratified sampling is of two types:

1. ***Proportionate stratified sampling:*** The number of sampling units drawn from each stratum is in proportion to the population size of that stratum.
2. ***Disproportionate stratified sampling:*** The number of sampling units drawn from each stratum is based on the analytical consideration, but not in proportion to the size of the population of that stratum.

Sampling process is as follows:

1. The population to be sampled is divided into groups (stratified).
2. A simple random sample is chosen.

Reason for Stratified Sampling

Sometimes, marketing professionals want information about the component part of the population. Assume there are three stores. Each store forms a strata and the sampling from within each strata is being selected. The resultant might be used to plan different promotional activities for each store strata.

Suppose a researcher wishes to study the retail sales of products, such as tea in a universe of 1,000 grocery stores (Kirana shops included). The researcher can first divide this universe into three strata based on the size of the store. This benchmark for size could be only one of the following (a) floor space, (b) volume of sales, (c) variety displayed, etc.

Size of Stores	No. of Stores	Percentage of Stores
Large stores	2,000	20
Medium stores	3,000	30
Small stores	5,000	50
	10,000	**100**

Suppose we need 12 stores. Now, choose four from each stratum, at random. If there was no stratification, simple random sampling from the population would be expected to choose two large stores (20% of 12) about four medium stores (30% of 12) and about six small stores (50% of 12).

As can be seen, each store can be studied separately using the stratified sample.

Selection by Proportionate Stratified Sample

Assume that there are 60 students in a class of a management school, of this, 10 has to be selected to take part in a Business quiz competition. Assume that the class has students specializing in marketing, finance and HR stream.

The first step is to subdivide the students of the class into 3 homogeneous groups or stratify the student population, by the area in which they are specializing.

Marketing Streaming			Finance Stream		HR Stream
1	32	8	11	33	34
2	36	12	13	35	37
3	40	15	17	38	39
4	43	18	20	41	42
5	46	19	21	44	45
7	47	22	24	49	48
14	50	27	29	52	51
16	53	31	30	55	54

Second step is to calculate the sampling fraction $f = n/N$

n = Sample size required

N = Population size

Third step - Determine how many are to be selected from marketing stream (say n_1)

$n_1 = 30 \times 1/10 = 30 \times 1/10$

Sample to be selected from marketing strata $n_1 = 30 \times 1/10 = 3$

Now we can select 3 numbers from among 30 numbers at random say 7, 60 and 22

Similarly we can select n_2 n_3

$n_2 = 20 \times 1/10 = 2$

The 2 numbers selected at random from finance stream are 13, 59

$N_3 = 10 \times 1/10 = 1$

Stratified sampling can be carried out with:

1. Same proportion across the strata proportionate stratified sample.
2. Varying proportion across the strata disproportionate stratified sample.

Example: Estimation of universe mean with a stratified sample

Size of Stores	No. of Stores (Population)	Sample Proportionate	Sample Disproportionate
Large	2,000	20	25
Medium	3,000	30	35
Small	5,000	50	40
	10,000	**100**	**100**

Solution:

Size of Stores	Sample Mean Sales per Store	No. of Stores	Percent of Stores
Large	200	2000	20
Medium	80	3000	30
Small	40	5000	50
		10,000	**100**

The population mean of monthly sales is calculated by multiplying the sample mean by its relative weight.

$200 \times 0.2 + 80 \times 0.3 + 40 \times 0.5 = 84$

Sample Proportionate

If N is the size of the population, n is the size of the sample.

i represents 1, 2, 3, ...*k* [number of strata in the population]

∴ Proportionate sampling

$$p = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots\ldots\ldots\ldots = \frac{n_k}{N_k} = \frac{n}{N}$$

$$\frac{n_1}{N_1} = \frac{n}{N} = n_1 = \frac{n}{N} \times n_1$$ And so on

n_1 is the sample size to be drawn from stratum 1

$n_1 + n_2 + \ldots\ldots\ldots\ldots n_k = n$ [Total sample size of the all strata]

Example: A survey is planned to analyse the perception of people towards their own religious practices. The population consists of various religions, viz., Hindu, Muslim, Christian, Sikh and Jain, assuming a total of 10,000. Hindu, Muslim, Christian, Sikh and Jains consist of 6,000, 2,000, 1,000, 500 and 500 respectively. Determine the sample size of each stratum by applying proportionate stratified sampling, if the sample size required is 200.

Solution:

Total population, $N = 10{,}000$

Population in the strata of Hindus $N_1 = 6{,}000$

Population in the strata of Muslims $N_2 = 2{,}000$

Population in the strata of Christians $N_3 = 1{,}000$

Population in the strata of Sikhs $N_4 = 500$

Population in the strata of Jains $N_5 = 500$

Proportionate Stratified Sampling

$$P = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \frac{n_4}{N_4} = \frac{n_5}{N_5} = \frac{n}{N}$$

Let us determine the sample size of strata N_1

$$\frac{n_1}{N_1} = \frac{n}{N} \times N_1 = \frac{200}{10{,}000} \times 6{,}000$$

$$= 20 \times 6$$

$$= 120$$

$$n_2 = \frac{n}{N} \times N_2 = \frac{200}{10{,}000} \times 2{,}000$$

$$= 40$$

$$n_3 = \frac{n}{N} \times N_3 = \frac{200}{10{,}000} \times 1{,}000$$

$$= 20$$

$$n_4 = \frac{n}{N} \times N_4 = \frac{200}{10{,}000} \times 500$$

$$= 10$$

$$n_5 = \frac{n}{N} \times N_5 = 10$$

$$n = n_1 + n_2 + n_3 + n_4 + n_5$$

$$= 120 + 40 + 20 + 10 + 10$$

$$= 200$$

Sample Disproportion

Let is the variance of the stratum i,

Where $i = 1, 2, 3.....k$.

The formula to compute the sample size of the stratum i is the variance of the stratum i,

Where size of stratum i

r_i = Sample size of stratum i

$$r_i = \frac{N_i}{N}$$

r_i = Ratio of the size of the stratum i with that of the population.

N_i = Population of stratum i

N = Total population.

Example: The Government of India wants to study the performance of Women Self Help Groups (WSHGs) in three regions viz. North, South and West. The total number of WSHGs is 1,500. The number of groups in North, South and West are 600, 500 and 400 respectively. The Government found more variation between WSHGs in the North, South and West regions. The variance of performance of WSHGs in these regions is 64, 25 and 16 respectively. If the disappropriate stratified sampling is to be used with the sample size of 100, determine the number of sampling units for each region.

Solutions:

Total Population $N = 1{,}500$

Size of the stratum 1, $N_1 = 600$

Size of the stratum 2, $N_2 = 500$

Size of the stratum 3, $N_3 = 400$

Variance of stratum 1, $\sigma = 1^2 = 64$

Variance of stratum 2, $\sigma = 2^2 = 25$

Variance of stratum 3, $\sigma = 3^2 = 16$

Sample size $n = 100$

Stratum Number	Size of the stratum N_i	$r_i = \frac{N_i}{N}$	σ_i	$r_i\sigma_{in}$	$r_i\sigma_{in} = \frac{r_i\sigma_{in}}{\sum_1^3 r_i\sigma_i}$
1	600	0.4	8	3.2	54
2	500	0.33	5	1.65	28
3	400	0.26	4	1.04	18
Total					**100**

3.5.4 Cluster Sampling

The following steps are followed:

1. The population is divided into clusters.
2. A simple random sample of few clusters is selected.
3. All the units in the selected cluster are studied.

Step 1: The above mentioned cluster sampling is similar to the first step of stratified random sampling. But the two sampling methods are different. The key to cluster sampling is decided by how homogeneous or heterogeneous the clusters are.

A major advantage of simple cluster sampling is the case of sample selection. Suppose, we have a population of 20,000 units from which we wish to select 500 units. Choosing a sample of that size is a very time-consuming process, if we use Random Numbers table. Suppose, the entire population is divided into 80 clusters of 250 units each, we can choose two sample clusters ($2 \times 250 = 500$) easily by using cluster sampling. The most difficult job is to form clusters. In marketing, the researcher forms clusters so that he can deal with each cluster differently.

Example: Assume there are 20 households in a locality.

Cross	Houses			
1	X_1	X_2	X_3	X_4
2	X_5	X_6	X_7	X_8
3	X_9	X_{10}	X_{11}	X_{12}
4	X_{13}	X_{14}	X_{15}	X_{16}

We need to select eight houses. We can choose eight houses at random. Alternatively, two clusters, each containing four houses can be chosen. In this method, every possible sample of eight houses would have a known probability of being chosen – i.e. chance of one in two. We must remember that in the cluster, each house has the same characteristics. With cluster sampling, it is impossible for certain random sample to be selected. For example, in the cluster sampling process described above, the following combination of houses could not occur: $X_1 X_2 X_5 X_6 X_9 X_{10} X_{13} X_{14}$. This is because the original universe of 16 houses has been redefined as a universe of four clusters. So, only clusters can be chosen as a sample.

3.5.5 Multi-stage Sampling

The name implies that sampling is done in several stages. This is used with stratified/ cluster designs.

An illustration of double sampling is as follows:

The management of a newly-opened club is solicits new membership. During the first rounds, all corporates were sent details so that those who are interested may enrol. Having enrolled, the second round concentrates on how many are interested to enrol for various entertainment activities that club offers such as billiards, indoor sports, swimming, gym, etc. After obtaining this information, you might stratify the interested respondents. This will also tell you the reaction of new members to various activities. This technique is considered to be scientific, since there is no possibility of ignoring the characteristics of the universe.

Advantage: May reduce cost, if first stage results are enough to stratify or cluster.

Disadvantage: Costs increase as more and more stages are included.

3.5.6 Area Sampling

This is a probability sampling, a special form of cluster sampling.

Example: If someone wants to measure the sales of toffee in retail stores, one might choose a city locality and then audit toffee sales in retail outlets in those localities.

The main problem in area sampling is the non-availability of lists of shops selling toffee in a particular area. Therefore, it would be impossible to choose a probability

sample from these outlets directly. Thus, the first job is to choose a geographical area and then list out outlets selling toffee. Then the probability sample for shops among the list prepared follows.

Example: You may like to choose shops which sell the brand—Cadbury dairy milk. The disadvantage of the area sampling is that it is expensive and time-consuming.

3.5.7 Advantages v/s Disadvantages of Probability Sampling

Advantages

- It is unbiased.
- Quantification is possible in probability sampling.
- Less knowledge of universe is sufficient.

Disadvantages

- It takes time.
- It is costly.
- More resources are required to design and execute than in non-probability design.

3.6 NON-PROBABILITY SAMPLING METHODS

In non-probability sampling, samples may be picked up based on the judgment or convenience of the enumerator. Usually, the complete sample is not decided at the beginning of the study but it evolves as the study progresses. However, the very same factors which govern the selection of a sample for example, judgment or convenience, can also introduce biases in the study. Moreover, there is no way that the magnitude of errors can be quantified when non-probability sampling designs are used.

The following are the types of non-probability sampling methods:

1. Deliberate sampling
2. Shopping mall intercept sampling
3. Sequential sampling
4. Quota sampling
5. Snowball sampling
6. Panel samples

3.6.1 Deliberate or Purposive Sampling

This is also known as the judgment sampling. The investigator uses his discretion in selecting sample observations from the universe. As a result, there is an element of bias in the selection. From the point of view of the investigator, the sample thus chosen may be a true representative of the universe. However, the units in the universe do not enjoy an equal chance of getting included in the sample. Therefore, it cannot be considered a probability sampling.

Example: Test market cities are being selected, based on the judgment sampling, because these cities are viewed as typical cities matching with certain demographical characteristics. Judgment sample is also frequently used to select stores for the purpose of introducing a new display.

3.6.2 Shopping Mall Intercept Sampling

This is a non-probability sampling method. In this method, the respondents are recruited for individual interviews at fixed locations in shopping malls. (*Example:* Shopper's Shoppe, Food World, Sunday to Monday). This type of study would include several malls, each serving different socio-economic population.

Example: The researcher may wish to compare the responses of two or more TV commercials for two or more products. Mall samples can be informative for this kind of studies. Mall samples should not be used under following circumstances i.e., if the difference in effectiveness of two commercials varies with the frequency of mall shopping, change in the demographic characteristic of mall shoppers, or any other characteristic. The success of this method depends on "How well the sample is chosen".

Merits

- It has a relatively small universe.
- In most cases, it is expected to give quick results. The purpose of deliberate sampling has become a practical method in dealing with economic or practical problems.
- In studies, where the level of accuracy can vary from the prescribed norms, this method can be used.

Demerits

- Fundamentally, this is not considered a scientific approach, as it allows for bias.
- The investigator may start with a preconceived idea and draw samples such that the units selected will be subjected to specific judgment of the enumerator.

3.6.3 Sequential Sampling

This is a method in which the sample is formed on the basis of a series of successive decisions. They aim at answering the research question on the basis of accumulated evidence. Sometimes, a researcher may want to take a modest sample and look at the results. Thereafter, he will decide if more information is required for which larger samples are considered. If the evidence is not conclusive after a small sample, more samples are required. If the position is still inconclusive, still larger samples are taken. At each stage, a decision is made about whether more information should be collected or the evidence is now sufficient to permit a conclusion.

Example: Assume that a product needs to be evaluated.

A small probability sample is taken from among the current user. Suppose it is found that average annual usage is between 200 to 300 units. It is known that the product is economically viable only if the average consumption is 400 units. This information is sufficient to take a decision to drop the product. On the other hand, if the initial sample shows a consumption level of 450 to 600 units, additional samples are needed for further study.

3.6.4 Quota Sampling

Quota sampling is quite frequently used in marketing research. It involves the fixation of certain quotas, which are to be fulfilled by the interviewers.

Example: Suppose, 2,00,000 students are appearing for a competitive examination. We need to select 1% of them based on quota sampling. The classification of quota may be as follows:

Classification of Samples

Category	Quota
General merit	1,000
Sport	600
NRI	100
SC/ST	300
Total	**2,000**

Quota sampling involves the following steps:

1. The population is divided into segments on the basis of certain characteristics. Here, the segments are termed as cells.
2. A quota of unit is selected from each cell.

Advantages

1. Quota sampling does not require prior knowledge about the cell to which each population unit belongs. Therefore, this sampling has a distinct advantage over stratified random sampling, where every population unit must be placed in the appropriate stratum before the actual sample selection.
2. It is simple to administer. Sampling can be done very quickly.
3. The necessity of the researcher going to various geographical locations is avoided and thus cost is reduced.

Limitations

1. It may not be possible to get a "representative" sample within the quota as the selection depends entirely on the mood and convenience of the interviewer.
2. Since too much liberty is being allowed to the interviewer, the quality of work suffers if they are not competent.

3.6.5 Snowball Sampling

This is a non-probability sampling. In this method, the initial group of respondents is selected randomly. Subsequent respondents are being selected based on the opinion or referrals provided by the initial respondents. Further referrals will lead to more referrals, thus leading to a snowball sampling. The referrals will have demographic and psychographic characteristics that are relatively similar to the person referring them.

Example: College students bring in more students on the consumption of Pepsi. The major advantage of snowball sampling is that it monitors the desired characteristics in the population.

3.6.6 Panel Samples

Panel samples are frequently used in marketing research. To give an example, suppose that one is interested in knowing the change in the consumption pattern of households. A sample of households is drawn. These households are contacted to gather information on the pattern of consumption. Subsequently, say after a period of six

months, the same households are approached once again and the necessary information on their consumption is collected.

3.7 DETERMINATION OF SAMPLE SIZE

One of the questions most frequently asked by marketing researchers to statisticians is "How large should my sample be?" This is also a commonly asked question by the students while undertaking a market research project. The best answer is "Get as large a sample as you can afford". If possible, 'sample' the entire population, provided you can ensure quality and control costs. This is better than any estimate. This, however, is unrealistic or impractical in most situations due to economic constraints, time constraints, and other limitations. Therefore, the best answer is "Get as large a sample as you can afford". Larger the sample, smaller is the standard error of our statistic.

When the sampling budget is limited, the question often is how to find the minimum sample size that will satisfy some precision requirements. In such cases, you should first give the answers of the following three questions:

1. How close do you want your sample estimate to be to the unknown parameter? The answer to this question is "bound", denoted by 'B'.
2. What do you want the confidence level to be so that the distance between the estimate and the parameter is less than or equal to B?
3. The last, and often misunderstood, question that must be answered is, "What is your estimate of the variance (or standard deviation) of the population in question?"

Only after you have answers to all three questions you can specify the minimum required sample size. If the population is approximately normal and you can get 95% bounds on the values in the population, by dividing the difference between the upper and lower bounds by 4; this will give you a rough guess of σ. Or you may take a small, inexpensive pilot survey and estimate σ from the sample standard deviation. B is called as half width of the required interval in which you need your estimate to fall with desired probability. Once you have obtained the three required pieces of information, all you need to do is to substitute the values into the appropriate formula as follows:

Minimum required sample size in estimating the population mean μ is,

$$n = \frac{z^2_{\alpha/2} \times \sigma^2}{B^2}$$

Minimum required sample size in estimating the population proportion p is,

$$n = \frac{z^2_{\alpha/2} \times p \times (1-p)}{B^2}$$

Note, that B is the margin of error. We are solving for the minimum sample size for a given margin of error. We need to guess the unknown population mean σ or the unknown population proportion p. Any prior estimate of the parameter can be used. When it is not available, we may take a pilot sample, or in the absence of any information, we use the value $p = 0.5$. This value maximizes $p \times (1-p)$ and thus ensures us a minimum required sample size that will work for any value of p.

This approach does not work if the population standard deviation is not known. The sample standard deviation is known only after the sample has been analyzed whereas the sample size decision is required before the sample is picked up.

Example: Suppose we know that the weight of cement in filed bags is distributed normally with a standard deviation σ of 0.2 Kg. We want to know how large a sample should be taken so that the mean weight of cement in filled bag can be estimated within plus or minus 0.05 kg of the true value with a confidence level of 90%.

Solution:

As the confidence level of 90% is specified, the corresponding interval for normal distribution is $\left(\bar{X} - \frac{1.645 \times \sigma}{\sqrt{n}}\right)$ to $\left(\bar{X} + \frac{1.645 \times \sigma}{\sqrt{n}}\right)$ that contained the true value of the population mean 90% of the time. We also want that the interval as $(\bar{X} - 0.05)$ to $(\bar{X} + 0.05)$ should give us a 90% confidence level. Hence,

$$\frac{1.645 \times \sigma}{\sqrt{n}} = 0.05 \Rightarrow \frac{1.645 \times 0.2}{\sqrt{n}} = 0.05 \Rightarrow n = \left(\frac{1.645 \times 0.2}{0.05}\right)^2$$

$n = 43.2964$

Thus, we must have a sample size of at least 44 so that the mean weight of cement in a filled bag can be estimated within plus or minus 0.05 kg of the true value with a 90% confidence level.

Example: Suppose we want to estimate the proportion of consumers in the population who prefer our product to the next competing brand. How large a sample should be taken so that the population proportion can be estimated within plus or minus 0.05 with a 90% confidence level?

Solution:

We shall use the sample proportion $\bar{p}$. If n is sufficiently large, the distribution of $\bar{p}$ can be approximated by a normal distribution with mean p and variance $\frac{p \times (1-p)}{n}$.

From the normal tables, we can say that the probability that $\bar{p}$ will lie between $[p \pm 1.645 \times \sqrt{\frac{p \times (1-p)}{n}}]$ is 0.90. In other words, the interval $[p \pm 1.645 \times \sqrt{\frac{p \times (1-p)}{n}}]$ will contain p, 90% of the time.

However, we want that the interval $[\bar{p} \pm 0.05]$ should contain p, 90% of the time.

Therefore, $1.645 \times \sqrt{\frac{p \times (1-p)}{n}} = 0.05 \Rightarrow \sqrt{\frac{p \times (1-p)}{n}} = \frac{0.05}{1.645} = 0.3039$

Or, $n = p \times (1-p) \times 1082.41$

But we do not know the value of p, so n cannot be calculated directly. However, whatever be the value of p, the highest value for the expression $p \times (1-p)$ is 0.25, when $p = 0.5$. Hence, considering the worst case,

$n = p \times (1-p) \times 1082.41 = 0.25 \times 1082.41 = 270.6025$

Therefore, if we take a sample of size 271, then we are sure that our estimate of the population proportion would be within ±0.05 of the true value with a confidence level of 90%, irrespective of the value of p.

Selecting Optimum Sample Size

As we have seen above, we need three things to decide the optimum sample size. Firstly, we take population standard deviation or sample standard deviation as estimators. In case of population proportion is parameter we take $p \times (1-p)$ as standard deviation if population proportion is known or 0.25 if population proportion is unknown. Secondly, we need to find z value for given confidence level if population is known to be normally distributed, or population proportion is parameter. If the population is not normally distributed, we need t value for given confidence level. Thirdly, we need to specify the half width B. This is called margin of error or bound we are ready to accept for our population estimate. Usual practice is to take 95 % or 99% confidence level. The corresponding values of level of significance α are 0.05 and 0.01 respectively. Thus, to select the optimum sample size the procedure is as follows:

(i) For given α, find $z_{\alpha/2}$ or $t_{\alpha/2}$ according to the population distribution is normal or not.

(ii) Find population S.D. σ or use its estimate as s of a pilot survey if population is known to be approximately normal or use value of $p \times (1-p)$ or 0.25 if proportion is being estimated.

(iii) Decide the value of half width desired for our estimate.

(iv) Then use the formula $n = \dfrac{z_{\alpha/2}^2 \times \sigma^2}{B^2}$, or $n = \dfrac{t_{\alpha/2}^2 \times \sigma^2}{B^2}$, or $n = \dfrac{z_{\alpha/2}^2 \times p \times (1-p)}{B^2}$

If we don't know population standard deviation σ or population proportion p, but know their estimates from a pilot survey as s or $\bar{p}$, we can find sample size as,

$$n = \frac{z_{\alpha/2}^2 \times s^2}{B^2}, \text{ or } n = \frac{t_{\alpha/2}^2 \times s^2}{B^2}, \text{ or } \quad n = \frac{z_{\alpha/2}^2 \times \bar{p} \times (1-\bar{p})}{B^2}$$

Example: Suppose an HR manager of a company wants to find the average time of travel to work for its employees within the ± 0.5 minute's interval accuracy. From the past data, a population standard deviation is known to be 2.5 minutes.

(i) Find the optimum sample size the company must choose that will give the average travel time of all employees with confidence level of 95%.

(ii) If after sampling we find the sample mean is 30 minutes, what should we conclude?

Solution:

(i) Now, we know that the value of alpha $\alpha = 0.05$ and the population S.D. is given as $\sigma = 2.5$

For $\alpha = 0.05$, from standard normal tables, $z_{\alpha/2} = 1.96$. Also, given B = 0.5. Now, sample size is,

$$n = \frac{z_{\alpha/2}^2 \times \sigma^2}{B^2} = \frac{(1.96)^2 (2.5)^2}{(0.5)^2} = 96.04 \approx 97$$

Thus, the HR manager needs to select sample size of 97.

(ii) If after sampling we find the sample mean is 30 minutes, we should conclude with confidence level of 95% that average travel time of all employees lies in interval [29.5, 30.5].

3.8 SAMPLING DISTRIBUTION

If we take only one sample from a population, calculate the sample statistics and wants to estimate about population parameters, the accuracy of estimation can be commented upon only if we know how the sample statistics behave statistically. Suppose we take large number of random samples of same size from the same population with replacement, the variations in the sample statistics of these samples would tell us about the variations due to chance. We take the random samples with replacement to remove any systematic errors. Now we can study the frequency distribution of the sample statistics obtained for all these samples. This gives us probability distribution of various sample statistics for example, probability distribution of sample means. It is a probability distribution because the variation in sample statistic of each sample is due to random probabilistic nature of sampling. A probability distribution of all possible statistics is called as sampling distribution.

For example, a probability distribution of all the possible sample means is called as sampling distribution of the mean. Similarly, we can also find the sampling distribution of variance, proportion, etc. We have seen earlier that any distribution can be described by its distribution parameters like mean, standard deviation, moments, skewness, kurtosis, characteristic function, etc.

In general, mean and standard deviation approximately describe the distribution. In case of normal distribution, mean and standard deviation describe the distribution completely. Thus, any sampling distribution can at least partially be described by its mean and standard deviation. Standard deviation of the sampling distribution is called as standard error. Standard error represents the variability of the sample statistic like sample mean, sample proportion, etc. When we draw a sample from population, obviously it is unlikely that the sample mean is exactly equal to the population mean because, in a sample we have taken only few of the items. Similarly, it is unlikely that means of all different samples are equal since the sample is randomly drawn from the population. The errors, i.e. difference between the populations mean and sample means is due to the chance solely due to the random selection of the samples. This error leads to the variability of the distribution of sample means.

Thus, in general, the standard deviation of the distribution of a sample statistic is known as the standard error of the statistic. Standard error is an indicator of the size of the error and accuracy we get in using sample statistics as an estimator of population parameters.

Standard error is applicable to all sample statistics like sample mean, sample median, sample range, sample proportion, etc. Standard error indicates how spread the distribution of sample statistics is. It gives manager a sort of confidence in his estimate. It also indicates the manager to what extent he should depend on the sample statistics. Obviously, if the sample is as large as the population itself, there would not be any difference between sample statistics and population parameters, thus, standard error would be zero. (Of course, then we don't need statistical inference.)

Hence, we conclude that as sample size increases, the standard error reduces. Although the sampling introduces random errors (chance errors), it gives us a mathematical measure for our errors and hence a mathematical measure for the level of confidence manager can have in using the sample statistic as an estimate of population.

Sampling process is a random experiment whose outcome is a sample. Thus, the sample statistic ($\bar{x}$, s, $\bar{p}$, etc.) are a random variables, and therefore, have a probability mass functions or a probability density functions according to whether the characteristic is discrete or continuous.

Thus, the sampling distribution of statistic is a probability distribution of a random variable defined as sample statistic and computed from random samples of the same size drawn from the population. Although, sampling distribution could be for any parameter of the sample, we usually refer to the sampling distribution of statistics like sample mean $\bar{x}$, sample variance s, sample proportion $\bar{p}$, sample median, etc, which are useful in drawing inference about the population. If we know the probability distribution of the sample static, then we can calculate the probability that the sample statistic assumes a particular value or has the value in a given interval. This ability to calculate the probability that the sample statistic lies in a particular interval is the most important factor in all statistical inferences.

For example, suppose we know that 40% of the soft drink consumers prefer Pepsi. Now we launch new advertisement campaign. After the campaign we take a random sample of 1000 soft drink consumers, and find 500 prefer Pepsi. What should we conclude? Here our sample distribution is going to help us. Now we would like to know the probability that the sample proportion in a sample size of 1000 is as large as 500 (50% of the sample) or higher when the true population proportion is only 40%. If this probability is large say 0.3, we may say that higher sample proportion of 500 (50% of the sample) is probably the result of sampling errors (including natural randomness) and not really due to new advertisement. On the other hand, if this probability works out to be very small, say 0.04, then we may conclude the preference for Pepsi has indeed increased above 40%. Though it may not be to the level of 50% as observed in the sample, but very likely to be above the original 40%. To calculate this probability, we need to know the probability distribution of sample proportion or the 'Sampling Distribution of the Proportion.'

3.8.1 Sampling Distribution of the Mean (Distribution of Sample Mean)

First, we study the concept of sample mean or its expected value and variance. Then we will discuss its distribution in general as well as in specific cases.

Because the sample is an outcome of a random experiment (random sampling), the sample mean itself is a random variable. The possible values of this random variable 'sample mean' depend on the possible values of elements in the random sample. Suppose we take a simple random sample of size 'n' picked up from a population. We measure the characteristics of interest of each sample member and denote the observations as $x_1, x_2, \ldots, x_n$ respectively. We then calculate mean $\bar{x}$ for the sample. This random sample and hence its mean in turn, depends on the distribution of the population from which it is drawn.

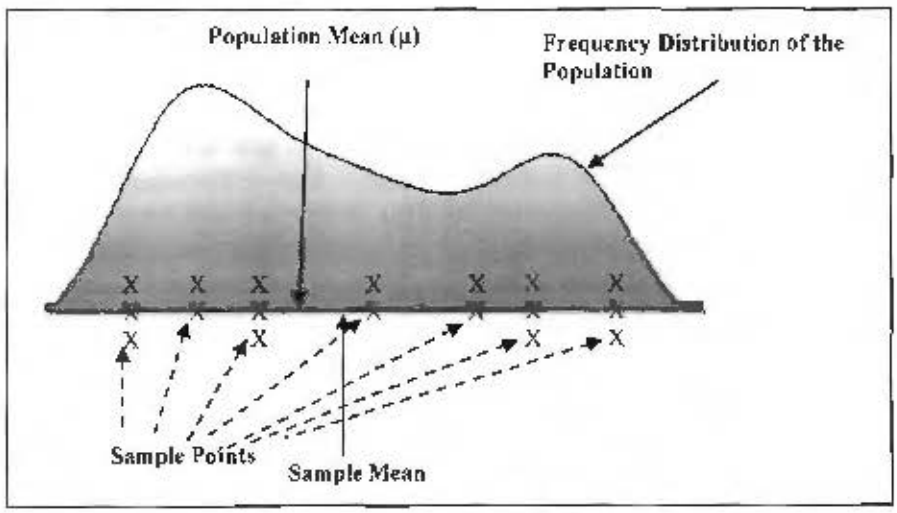


Figure 3.2: Sampling Distribution of Mean

If we draw another sample of size 'n' from the same population, we are most likely to end up with totally different sample mean. Thus, there are many (theoretically infinite) possible values of the sample mean and we got our sample mean only by a chance. The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the random variable $\bar{x}$ may take when a sample of size 'n' is taken from a specific population. Now let us discuss sampling distribution of a sample mean under various situations.

3.8.2 Sampling from Infinite Population

We consider the population is infinite when we take sample with replacement or the population size N is very large as compared to the sample size 'n' ($N >> n$). Usually as a rule of thumb, we consider sampling from infinite population if ratio $\frac{n}{N}$ is less than 0.05.

Suppose we assume that we have a population, which is infinitely large, and having a population mean of μ and a population variance of σ^2. This implies that if is a random variable X denoting the measurement of the characteristic that we are interested in, and one element of the population picked up randomly say x_i, then we get,

The expected value of X is $E(X) = \mu$

And the variance of X, Variance $(X) = E(X^2) - [E(X)]^2 = \sigma^2$

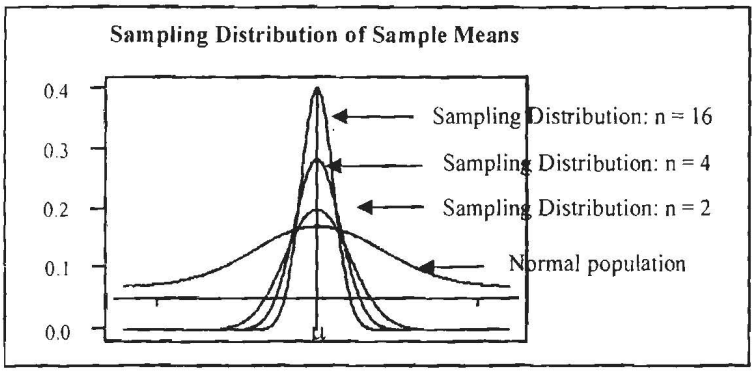


Figure 3.3: Sampling from Infinite Population

The sample mean, $\bar{x}$ can be looked at as the sum of 'n' values of random variables representing each sample element, viz. $x_1, x_2, ..., x_n$, divided by n. Here X_1 is a random variable representing the first pick in the sample and its value is x_1, X_2 is a random variable representing the second pick in the sample with its value as x_2 and so on. Now, when the population is infinitely large, whatever the value of x_1 may be, the distribution of x_2 is not affected by it. This is true of any other pair of random variables as well. In other words, $X_1, X_2, ...X_n$ are independent random variables and all are picked up from the same population. (Note that the random variable is denoted by a capital letter and values of the random variables are denoted by the small letters as per the convention.)

Therefore, the expected value of X_1 is $E(X_1) = \mu$

and the variance of X_1, Variance $(X_1) = E(X_1^2) - [E(X_1)]^2 = \sigma^2$

Similarly, $E(X_2) = \mu$, Variance $(X_2) = \sigma^2$ and so on.

Finally, $E(\bar{X}) = E\left(\frac{X_1 + X_2 + ... + X_n}{n}\right)$

$$= \frac{E(X_1) + E(X_2) + ... + E(X_n)}{n} = \frac{\mu + \mu + ... + \mu \quad n \ \ times}{n}$$

$$E(\bar{X}) = \mu \qquad \ldots (1)$$

And, $\text{Variance}(\bar{X}) = \text{Variance}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right)$

$$= \text{Variance}\left(\frac{X_1}{n}\right) + \text{Variance}\left(\frac{X_2}{n}\right) + ... + \text{Variance}\left(\frac{X_n}{n}\right)$$

$$= \frac{1}{n^2}\text{Variance}(X_1) + \frac{1}{n^2}\text{Variance}(X_2) + ... + \frac{1}{n^2}\text{Variance}(X_n)$$

$$= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + ... + \frac{1}{n^2}\sigma^2 = \frac{\sigma^2 + \sigma^2 + ... + \sigma^2 \quad n \ \ times}{n}$$

$$\text{Variance}(\bar{X}) = \frac{\sigma^2}{n} \qquad \ldots (2)$$

We have arrived at two very important results for the case when the population is infinitely large, which very commonly used for reasonably large population. The first result says that the expected value of the sample mean is the same as the population mean. The second result says that the variance of the sample mean is the variance of the population divided by the sample size. Thus, if we take a large number of samples of size n, then the average value of the sample means tends to the true population mean. On the other hand, if the sample size is increased then the variance of $\bar{X}$ can be made as small as desired.

The standard deviation of $\bar{X}$ is also called the standard error of the mean and denoted by $\sigma_{\bar{X}}$. Very often we use the sample mean as estimate the population mean. The standard error of the mean is a measure of the extent to which the observed value of sample mean can be away from the true value, due to sampling errors. For example, if the standard error of the mean is small, we are reasonably confident that the observed sample mean value is very close to the true value of the population mean.

3.8.3 Sampling with Replacement

The above results have been obtained under the assumption that the random variables $X_1, X_2, ..., X_n$ are independent. This assumption is valid when the population is infinitely large. It is also valid when the sampling is done with replacement, so that the population is back to the same form before the next sample member is picked up. Hence, if the sampling were done with replacement, we would again get,

$$E(\bar{X}) = \mu$$

and $\text{Variance}(\bar{X}) = \frac{\sigma^2}{n}$ i.e. $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

3.8.4 Sampling without Replacement from Finite Populations

When a sample is picked up without replacement from a finite population, the probability distribution of the second random variable depends on what has been the outcome of the first pick and so on. As the 'n' random variables representing the 'n'

sample members do not remain independent, the expression for the variance of $\bar{X}$ changes. The result we get is,

$E(\bar{X}) = \mu$ same as infinite population result.

And, Variance $(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \times \left[\frac{N-n}{N-1}\right]$

i.e. $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$ ...(3)

By comparing these expression with the ones derived earlier for infinite population, we note that the standard error of $\bar{X}$ is multiplied by a factor $\sqrt{\frac{N-n}{N-1}}$. This factor is known as the finite population multiplier. In practice, almost all the samples are picked up without replacement. Also, most populations are finite, although may be very large. So the standard error of the mean should theoretically be found by using expression (3) with the finite population multiplier. However, if the population size 'N' is large enough ($N >> n$) with the sampling ratio $\frac{n}{N}$ small, then the finite population multiplier is almost equal to 1 and hence need not be used. Thus, we can treat large finite population as if it was infinitely large. For example, if N = 100000 and n = 100, the finite population multiplier.

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{100000-100}{100000-1}} = \sqrt{\frac{99900}{99999}} = 0.999 \approx 1$$

Thus, the standard error of the mean would, for all practical purpose, be the same whether the population is treated as finite or infinite.

As a rule of thumb, the finite population multiplier need not be used if the sampling ratio $\frac{n}{N} < 0.05$.

3.9 SAMPLING ERROR

The only way to guarantee the minimisation of sampling error is to choose the appropriate sample size. As the sample keeps on increasing, the sampling error decreases. Sampling error is the gap between the sample mean and population mean.

Example: If a study is done amongst Maruti car-owners in a city to find the average monthly expenditure on the maintenance of car, it can be done by including all Maruti car-owners. It can also be done by choosing a sample without covering the entire population. There will be a difference between the two methods with regard to monthly expenditure.

3.9.1 Non-sampling Error

One way of distinguishing between the sampling and the non-sampling error is that, while sampling error relates to random variations which can be found out in the form of standard error, non-sampling error occurs in some systematic way which is difficult to estimate.

3.9.2 Sampling Frame Error

A sampling frame is a specific list of population units, from which the sample for a study being chosen.

Example: An MNC bank wants to pick up a sample among the credit card holders. They can readily get a complete list of credit card holders, which forms their data bank. From this frame, the desired individuals can be chosen. In this example, sample frame is identical to ideal population namely all credit card holders. There is no sampling error in this case.

Example: Assume that a bank wants to contact the people belonging to a particular profession over phone (doctors, lawyers) to market a home loan product. The sampling frame in this case is the telephone directory. This sampling frame may pose several problems: (1) People might have migrated. (2) Numbers have changed. (3) Many numbers were not yet listed. The question is "Are the residents who are included in the directory likely to differ from those who are not included"? The answer is yes. Thus in this case, there will be a sampling error.

3.9.3 Non-response Error

This occurs, because the planned sample and final sample vary significantly.

Example: Marketers want to know about the television viewing habits across the country. They choose 500 households and mail the questionnaire. Assume that only 200 respondents reply. This does not show a non-response error, which depends upon the discrepancy. If those 200 who replied did not differ from the chosen 500, there is no non-response error.

Consider an alternative. The people who responded are those who had plenty of leisure time. Therefore, it is implied that non-respondents do not have adequate leisure time. In this case, the final sample and the planned sample differ. If it was assumed that all the 500 chosen have leisure time, but in the final analysis only 200 have leisure time and not others. Therefore, a sample with respect to leisure time leads to response error.

Guidelines to Increase the Response Rate

Every researcher likes to get maximum possible response from the respondents, and will be most delighted if cent percent respondent unfortunately, this does not happen. The non-response error can be reduced by increasing the response rate. Higher the response rate, more accurate and reliable is the data. In order to achieve this, some useful hints could be as follows:

(a) Intimate the respondents in advance through a letter. This will improve the preparedness.

(b) Personalized questionnaire should be accompanied by a covering letter.

(c) Ensure/Assure that confidentiality will be maintained.

(d) Questionnaire length is to be restricted.

(e) Increase of personal interview, ID card is essential to prove the bonafide.

(f) Monetary incentives are gifts will act as motivator.

(g) Reminder/Revisits would help.

(h) Send self-addressed/stamped envelope to return the completed questionnaire.

3.9.4 Data Error

This occurs during the data collection, analysis of data or interpretation. Respondents sometimes give distorted answers unintentionally for questions which are difficult, or if the question is exceptionally long and the respondent may not have answer. Data errors can also occur depending on the physical and social characteristics of the interviewer and the respondent. Things such as the tone and voice can affect the responses. Therefore, we can say that the characteristics of the interviewer can also result in data error. Also, cheating on the part of the interviewer leads to data error. Data errors can also occur when answers to open-ended questions are being improperly recorded.

3.9.5 Failure of the Interviewer to Follow Instructions

The respondent must be briefed before beginning the interview, "What is expected"? "To what extent he should answer"? Also, the interviewer must make sure that respondent is familiar with the subject. If these are not made clear by the interviewer, errors will occur.

Editing mistakes made by the editors in transferring the data from questionnaire to computers are other causes for errors.

The respondent could terminate his/her participation in data gathering, because it may be felt that the questionnaire is too long and tedious.

How to Reduce Non-sampling Error?

1. For non-response – provide incentives such as a gift or cash. This enhances the possibility as well as incidence of response.
2. Data error: Don't ask question, which respondents cannot answer. Also, do not ask sensitive questions.
3. Train the interviewer to establish a good rapport with the respondents.
4. Avoid leading questions.
5. Pre-test the questionnaire.
6. Modify the sampling frame to make it a representative of the population.

3.10 STATISTICAL SERIES AND ITS TYPES

The classified data when arranged in some logical order, e.g., according to the size, according to the time of occurrence or according to some other measurable or non-measurable characteristics, is known as Statistical Series. H. Secrist defined a statistical series as, *"A series, as used statistically, may be defined as things or attributes of things arranged according to some logical order."* Another definition given by L. R. Connor as, *"If the two variable quantities can be arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other, the result is said to form a statistical series."*

A statistical series can be one of the following four types:

(i) Spatial Series,

(ii) Conditional Series,

(iii) Time Series and

(iv) Qualitative or Quantitative Series.

1. ***Spatial Series:*** The series formed by the geographical or spatial classification is termed as spatial series. A method of analyzing time series, called the spatial analysis. The analysis consists mainly of the statistical inference on the distribution given by the expected local time, which we define to be the spatial distribution, of a given time series. The spatial distribution is introduced primarily for the analysis of non-stationary time series whose distributions change over time. However, it is well defined for both stationary and non-stationary time series, and reduces to the time invariant stationary distribution if the underlying time series is indeed stationary. The spatial analysis may therefore be regarded as an extension of the usual inference on the distribution of a stationary time series to accommodate for non-stationary time series.

2. ***Conditional Series:*** Similarly, a series formed by the conditional classification is known as the conditional series.

3. ***Time Series:*** A time series is the result of chronological classification of data. In this case, various figures are arranged with reference to the time of their occurrence. For example, the data on exports of India in various years is a time series.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988
Exports (in ₹ cr.)	6591	7242	8309	8810	9981	10427	11490	15741	20295

4. ***Qualitative or Quantitative Series:*** This type of series is obtained when the classification of data is done on the basis of qualitative or quantitative characteristics. Accordingly, we can have two types of series, namely, qualitative and quantitative series.

   (a) *Qualitative Series:* In case of qualitative series, the numbers of items in each group are shown against that group. These groups are either expressed in ascending order or in descending order of the number of items in each group. The example of such a series is given below.

   Distribution of Students of a College according to Sex

Sex	Males	Females	**Total**
No. of Students	1700	500	**2200**

   (b) *Quantitative Series:* In case of quantitative series, the number of items possessing a particular value is shown against that value.

   A quantitative series can be of two types:

   (i) Individual series: In an individual series, the names of the individuals are written against their corresponding values. For example, the list of employees of a firm and their respective salary in a particular month.

   (ii) Frequency Distribution: A table in which the frequencies and the associated values of a variable are written side by side, is known as a frequency distribution. According to Croxton and Cowden, "Frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in a group with their corresponding frequencies side by side." A frequency distribution can be discrete or continuous depending upon whether the variable is discrete or continuous.

Check Your Progress

Fill in the blanks:

1. In _____ sampling, samples may be picked up based on the judgment or convenience of the enumerator.
2. Purposive sampling is also known as the _____ sampling.
3. _____ is the gap between the sample mean and population mean. Searching the environment for information called the _____ activity.
4. _____ is the list of elements from which the sample is actually drawn.
5. A _____ is appropriate if the size of population is small.
6. Under _____ method, numbers or names of various units of population are noted on chits and put in a container.

3.11 LET US SUM UP

- A sample is a part of a total aggregate selected with view to obtain information about the whole group. The whole group is known as population.
- Sampling is less expensive, less time consuming and more accurate as compared to completing enumeration (census). In case of destructive tests, there is no alternative to the sampling. Sampling is used in many of the practical situations like quality control, market research, medical research, experimental analysis, inventory control, surveys and so on.
- Sample is a representative of population. Census represents cent percent of population. The most important factors distinguishing whether to choose sample or census is cost and time. There are seven steps involved in selecting the sample.
- There are two types of sample (a) Probability sampling and (b) Non-probability sample. Probability sampling includes random sampling, stratified random sampling systematic sampling, cluster sampling and multistage sampling. Random sampling can be chosen by lottery method or using random number table. Samples can be chosen either with equal probability or varying probability. Random sampling can be systematic or stratified. In systematic random sampling, only the first number is randomly selected. Then by adding a constant "K" remaining numbers are generated. In stratified sampling, random samples are drawn from several strata, which have more or less same characteristics. In multistage sampling, sampling is drawn in several stages.

3.12 UNIT END ACTIVITY

Suppose you have to compare the fashion sense of different age groups in Delhi. How will you select the optimum sample size?

3.13 KEYWORDS

Population: The aggregate or totality of all members is known as 'population'; for example, all items produced in a day at a factory is a population; few items selected for testing or performance measurement by quality control inspector is a sample.

Sample: Data sample is a set of data collected and/or selected from a statistical population by a defined procedure.

Sampling: It indicates the selection of a part of the whole with a view to obtain information about the whole.

Random Sample: A random sample of 'n' elements is a sample selected from the entire population N such that each element (or member) has an equal chance (or probability) of getting selected.

Sampling Frame: It is a list of all the units of the population. The preparation of a sampling frame is many times, a major formidable practical problem. The frame should be up-to-date and free from omission and duplication.

Standard Error: Standard error indicates how spread the distribution of sample statistics is. It gives manager a sort of confidence in his estimate.

Census: A census is appropriate if the size of population is small.

Simple Random Sampling: Simple Random Sampling is the simplest type of sampling, in which we draw a sample of size (n) in such a way that each of the 'N' members of the population has the same chance of being included in the sample.

Stratified Random Sampling: In stratified random sampling, the members of the population are first assigned to strata or groups, on the basis of some characteristic and a simple random sample is drawn from each stratum.

Judgment Sampling: In judgment sampling, the judgment or opinion of some experts forms the basis of the sampling method.

3.14 QUESTIONS FOR DISCUSSION

1. Define sample, sampling, sampling frame and parameters.
2. What are the advantages of sampling over census?
3. Discuss the various sampling techniques. Differentiate between probability and non-probability sampling.
4. A sample of 400 items is taken from a normal population whose mean as well as variance is 4. If the sample mean is 4.5, can the sample be regarded as a truly random sample?
5. Differentiate between simple random sampling and systematic sampling.
6. Write a short note on cluster vs. stratified sampling.
7. How are judgement and purposive sampling easy to use as compared to probability sampling methods?
8. What is the minimum required sample size in estimating the population mean and population proportion?
9. How do you select an optimum sample size? Explain with the help of an example.
10. What is a sampling frame? What are its different types?
11. Construct a sample design to find the information on following:
    (a) Satisfaction level of Maruti car users in Pune
    (b) Preference of Coke over Pepsi in your town
    (c) Proportion of PC owners in your organization
12. Define population and sampling unit for selecting a random sample in each of the following cases:
    (a) Hundred voters from a constituency
    (b) Twenty stocks of National Stock Exchange

(c) Fifty account holders of State Bank of India

(d) Twenty employees of Tata Motors

13. What are the main questions that you keep in mind while determining the sample size and how do you select optimum sample size?

14. Explain in detail concept of sampling distribution. Also explain the sampling distribution of the mean with diagram.

15. How is sampling with replacement different from sampling without replacement? Explain with diagram and examples.

16. Identify the appropriate target population and sampling frame for various situations listed below:

    (a) The regional marketing manager of a beverage company wants to test market three new flavours to gauge their acceptance.

    (b) A manufacturer wants to assess whether adequate inventories of spare parts are being maintained by the distributors to prevent shortages and loss of business.

    (c) A wholesaler dealing with audio/video equipments wants to evaluate the reaction of dealers towards a new promotion policy announced.

    (d) A TV channel wants to determine the viewing habits of housewives and their programme preferences.

    (e) A departmental chain such as Food World wants to determine the shopping behaviour of customers who use the credit cards.

17. The above TV manufacturer is in the market for the past eight years. A survey conducted in the past by an MR agency produced the following score using Likert Scale. The data for various years is as below:

    2000 - 18

    2001 - 16

    2002 - 17

    2003 - 18

    2004 - 20

    What do you conclude about the customers' attitude? Is it favourable or unfavourable?

18. What is statistical series?

19. Discuss four types of statistical series.

Check Your Progress: Model Answer

1. Non-probability
2. Judgment
3. Sampling error, intelligence
4. Sampling frame
5. Census
6. Lottery

3.15 REFERENCE & SUGGESTED READINGS

- Groebner, D. F., Shannon, P. W., Fry, P. C., & Smith, K. D. (2022). **Business Statistics: A Decision-Making Approach** (11th ed.). Pearson. ISBN: 9780136681503
- Albright, S. C., Winston, W. L., & Zappe, C. (2021). **Data Analysis and Decision Making** (6th ed.). Cengage Learning. ISBN: 9780357131785
- Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2020). **Statistics for Business and Economics** (14th ed.). Cengage Learning. ISBN: 9780357114474
- Jaggia, S., Kelly, A., &Lertwachara, K. (2021). **Essentials of Business Statistics** (2nd ed.). McGraw-Hill Education. ISBN: 9781260205799

BLOCK - 2

UNIT-IV

MEASUREMENT OF CENTRAL TENDENCY

CONTENTS

Contd...

4.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Define the term average and its functions and characteristics
- Write the uses, merits and demerits of mean, median and mode
- Tell about mathematical averages – AM, GM and HM
- Establish the relationship amongst mean, median and mode
- Establish the relationship amongst AM, GM and HM

4.1 INTRODUCTION

Summarization of the data is a necessary function of any statistical analysis. As a first step in this direction, the huge mass of unwieldy data is summarized in the form of tables and frequency distributions. In order to bring the characteristics of the data into sharp focus, these tables and frequency distributions need to be summarized further. A measure of central tendency or an average is very essential and an important summary measure in any statistical analysis.

4.2 AVERAGE

The average of a distribution has been defined in various ways. Some of the important definitions are mentioned below:

"An average is an attempt to find one single figure to describe the whole of figures".

— Clark and Sekkade

"Average is a value which is typical or representative of a set of data".

— Murray R. Spiegal

"An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data it is sometimes called a measure of central value".

— Croxton and Cowden

"A measure of central tendency is a typical value around which other figures congregate".

— Sipson and Kafka

4.2.1 Functions of an Average

- ***To present huge mass of data in a summarised form:*** It is very difficult for human mind to grasp a large body of numerical figures. A measure of average is used to summarise such data into a single figure which makes it easier to understand and remember.

- ***To facilitate comparison:*** Different sets of data can be compared by comparing their averages. For example, the level of wages of workers in two factories can be compared by mean (or average) wages of workers in each of them.
- ***To help in decision making:*** Most of the decisions to be taken in research, planning, etc., are based on the average value of certain variables.

  Example: If the average monthly sales of a company are falling, the sales manager may have to take certain decisions to improve it.

4.2.2 Characteristics of a Good Average

A good measure of average must possess the following characteristics:

- It should be rigidly defined, preferably by an algebraic formula, so that different persons obtain the same value for a given set of data.
- It should be easy to compute.
- It should be easy to understand.
- It should be based on all the observations.
- It should be capable of further algebraic treatment.
- It should not be unduly affected by extreme observations.
- It should not be much affected by the fluctuations of sampling.

4.2.3 Various Measures of Average

Various measures of average can be classified into the following three categories:

1. ***Mathematical Averages***
   (a) Arithmetic Mean or Mean
   (b) Geometric Mean
   (c) Harmonic Mean
   (d) Quadratic Mean
2. ***Positional Averages***
   (a) Median
   (b) Mode
3. ***Commercial Average***
   (a) Moving Average
   (b) Progressive Average
   (c) Composite Average

Out of above mentioned, we will discuss here only mathematical averages and positional averages.

An average is a single value which can be taken as representative of the whole distribution.

4.3 ARITHMETIC MEAN

Before the discussion of arithmetic mean, we shall introduce certain notations. It will be assumed that there are n observations whose values are denoted by $X_1, X_2, \ldots\ldots X_n$ respectively. The sum of these observations $X_1 + X_2 + \ldots\ldots + X_n$ will be denoted in

abbreviated form as $\sum_{i=1}^{n} X_i$, where Σ (called sigma) denotes summation sign. The subscript of X, i.e., 'i' is a positive integer, which indicates the serial number of the observation. Since there are n observations, variation in i will be from1 to n. This is indicated by writing it below and above Σ, as written earlier. When there is no ambiguity in range of summation, this indication can be skipped and we may simply write $X_1 + X_2 + ... + X_n = \Sigma X_i$.

Arithmetic mean is defined as the sum of observations divided by the number of observations. It can be computed in two ways:

1. Simple arithmetic mean
2. Weighted arithmetic mean

In case of simple arithmetic mean, equal importance is given to all the observations while in weighted arithmetic mean, the importance given to various observations is not same.

4.3.1 Simple Arithmetic Mean

When Individual Observations are Given

Let there be n observations X_1, X_2 ..... X_n. Their arithmetic mean can be calculated either by direct method or by short cut method. The arithmetic mean of these observations will be denoted by $\overline{X}$.

Direct Method: Under this method, X is obtained by dividing sum of observations by number of observations, i.e.,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Short-cut Method: This method is used when the magnitude of individual observations is large. The use of shortcut method is helpful in the simplification of calculation work.

Let A be any assumed mean. We subtract A from every observation. The difference between an observation and A, i.e., $X_i - A$ is called the deviation of ith observation from A and is denoted by d_i. Thus, we can write ; $d_1 = X_1 - A$, $d_2 = X_2 - A$, ... $d_n = X_n - A$. On adding these deviations and dividing by n we get

$$\frac{\sum d_i}{n} = \frac{\sum (X_i - A)}{n} = \frac{\sum X_i - nA}{n} = \frac{\sum X_i}{n} - A$$

$$\text{or } \overline{d} = \overline{X} - A \qquad (\text{Where } \overline{d} = \frac{\sum d_i}{n})$$

On rearranging, we get $\overline{X} = A + \overline{d} = A + \frac{\sum d_i}{n}$

This result can be used for the calculation of $\overline{X}$

Remarks: Theoretically we can select any value as assumed mean. However, for the purpose of simplification of calculation work, the selected value should be as nearer to the value of $\overline{X}$ as possible.

Example: The following figures relate to monthly output of cloth of a factory in a given year:

Months :	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Output :	80	88	92	84	96	92	96	100	92	94	98	86

(in '000 metres)

Calculate the average monthly output.

Solution:

1. Using Direct Method

$$\overline{X} = \frac{80+88+92+84+96+92+96+100+92+94+98+86}{12}$$

$= 91.5$ ('000mts)

2. Using Short-cut Method

Let $A = 90$

X_i	80	88	92	84	96	92	96	100	92	94	98	86	**Total**
$d_i = X_i - A$	–10	–2	2	–6	6	2	6	10	2	4	8	–4	$\Sigma d_i = 18$

$$\therefore \overline{X} = 90 + \frac{18}{12} = 90 + 1.5 = 91.5 \text{ thousand mtrs}$$

When Data are in the Form of an Ungrouped Frequency Distribution

Let there be n values $X_1, X_2, \ldots\ldots X_n$ out of which X_1 has occurred f_1 times, X_2 has occurred f_2 times, ... X_n has occurred f_n times. Let N be the total frequency, i.e., $N = \sum_{i=1}^{n} f_i$. Alternatively, this can be written as follows:

Values	X_1	X_2	- - -	X_n	**Total Frequency**
Frequency	f_1	f_2	- - -	f_n	**N**

Direct Method: The arithmetic mean of these observations using direct method is given by

$$X = \frac{\overbrace{X_1 + X_1 + \ldots + X_1}^{f_1 \text{ times}} + \overbrace{X_2 + \ldots + \ldots + X_2}^{f_2 \text{ times}} + \ldots + \overbrace{\ldots + \ldots + X_n}^{f_n \text{ times}}}{f_1 + f_2 + \ldots + f_n}$$

Since $X_1 + X_1 + \ldots\ldots + X_1$ added f_1 times can also be written f_1X_1. Similarly, by writing other observation in same manner, we have

$$\overline{X} = \frac{f_1X_1 + f_2X_2 + \ldots\ldots + f_nX_n}{f_1 + f_2 + \ldots\ldots + f_n} = \frac{\sum_{i=1}^{n} f_iX_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_iX_i}{N}$$

Short-cut Method: As before, we take the deviations of observations from an arbitrary value A. The deviation of i^{th} observation from A is $d_i = X_i - A$.

Multiplying both sides by f_i we have $f_i\, d_i = f_i\,(X_i - A)$

Taking sum over all the observations

$$\Sigma f_i d_i = \Sigma f_i (X_i - A) = \Sigma f_i X_i - A\Sigma f_i = \Sigma f_i X_i - A.N$$

Dividing both sides by N we have

$$\frac{\sum f_i d_i}{N} = \frac{\sum f_i X_i}{N} - A = \overline{X} - A \text{ or } \overline{X} = A + \frac{\sum f_i d_i}{N} = A + \overline{d}$$

Example: The following is the frequency distribution of age of 670 students of a school. Compute the arithmetic mean of the data:

X (in years)	5	6	7	8	9	10	11	12	13	14
Frequency	25	45	90	165	112	96	81	26	18	12

Solution:

Direct Method: The computations are shown in the following table:

X	5	6	7	8	9	10	11	12	13	14	Total
f	25	45	90	165	112	96	81	26	18	12	Σf = 670
fX	125	270	630	1320	1008	960	891	312	234	168	ΣfX = 5918

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{5918}{670} = 8.83 \text{ years.}$$

Short-cut Method: The method of computations is shown in the following table:

X	5	6	7	8	9	10	11	12	13	14	Total
f	25	45	90	165	112	96	81	26	18	12	670
D = X – 8	–3	–2	–1	0	1	2	3	4	5	6	
fd	–75	–90	–90	0	112	192	243	104	90	72	558

$$\therefore \quad \overline{X} = A + \frac{\sum fd}{N} = 8 + \frac{558}{670} = 8 + 0.83 = 8.83 \text{ years.}$$

When Data are in the Form of a Grouped Frequency Distribution

In a grouped frequency distribution, there are classes along with their respective frequencies.

Let l_i be the lower limit and u_i be the upper limit of i^{th} class. Further, let the number of classes be n, so that i = 1, 2, ... n. Also let f_i be the frequency of i^{th} class. This distribution can be written in tabular form, as shown.

Note: Here u_1 may or may not be equal to l_2, i.e., the upper limit of a class may or may not be equal to the lower limit of its following class.

It may be recalled here that, in a grouped frequency distribution, we only know the number of observations in a particular class interval and not their individual magnitudes. Therefore, to calculate mean, we have to make a fundamental assumption that the observations in a class are uniformly distributed.

Under this assumption, the mid-value of a class will be equal to the mean of observations in that class and hence can be taken as their representative. Therefore, if X_i is the mid-value of i^{th} class with frequency f_i, the above assumption implies that there are f_i observations each with magnitude X_i (i = 1 to n). Thus, the arithmetic mean

of a grouped frequency distribution can also be calculated by the use of the formula, given below.

Class Interval	Frequency (f)
l_1-u_1	f_1
l_1-u_1	f_2
⋮	⋮
l_n-u_n	f_n
Total Frequency	**$= \Sigma f_i = N$**

Remarks: The accuracy of arithmetic mean calculated for a grouped frequency distribution depends upon the validity of the fundamental assumption. This assumption is rarely met in practice. Therefore, we can only get an approximate value of the arithmetic mean of a grouped frequency distribution.

Example: Calculate arithmetic mean of the following distribution:

Class Intervals:	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency :	3	8	12	15	18	16	11	5

Solution:

Here only short-cut method will be used to calculate arithmetic mean but it can also been calculated by the use of direct-method.

Class Intervals	Mid Values (X)	Frequency (f)	D = X – 35	fd
0-10	5	3	–30	–90
10-20	15	8	–20	–160
20-30	25	12	–10	–120
30-40	35	15	0	0
40-50	45	18	10	180
50-60	55	16	20	320
60-70	65	11	30	330
70-80	75	5	40	200
Total		**88**		**660**

$$\therefore \ \overline{X} = A + \frac{\sum fd}{N} = 35 + \frac{660}{88} = 42.5$$

4.3.2 Weighted Arithmetic Mean

In the computation of simple arithmetic mean, equal importance is given to all the items. But this may not be so in all situations. If all the items are not of equal importance, then simple arithmetic mean will not be a good representative of the given data. Hence, weighing of different items becomes necessary. The weights are assigned to different items depending upon their importance, i.e., more important items are assigned more weight.

Example: To calculate mean wage of the workers of a factory, it would be wrong to compute simple arithmetic mean if there are a few workers (say managers) with very high wages while majority of the workers are at low level of wages. The simple arithmetic mean, in such a situation, will give a higher value that cannot be regarded

as representative wage for the group. In order that the mean wage gives a realistic picture of the distribution, the wages of managers should be given less importance in its computation. The mean calculated in this manner is called weighted arithmetic mean. The computation of weighted arithmetic is useful in many situations where different items are of unequal importance, e.g., the construction index numbers, computation of standardised death and birth rates, etc.

Formulae for Weighted Arithmetic Mean

Let X_1, X_2 ....., X_n be n values with their respective weights w_1, w_2 ....., w_n. Their weighted arithmetic mean denoted as $\bar{X}_w$ is given by,

1. Using direct method

$$\overline{X}_w = \frac{\sum w_i X_i}{\sum w_i}$$

2. Using short-cut method

$$\overline{X}_w = A + \frac{\sum w_i d_i}{\sum w_i} \text{ (where } d_i = X_i - A)$$

3. Using step-deviation method

$$\overline{X}_w = A + \frac{\sum w_i u_i}{\sum w_i} \times h \text{ (where } u_i = \frac{X_i - A}{h})$$

Remarks: If $\bar{X}$ denotes simple mean and $\bar{X}_w$ denotes the weighted mean of the same data, then

1. $\bar{X} = \bar{X}_w$, when equal weights are assigned to all the items.
2. $\bar{X} > \bar{X}_w$, when items of small magnitude are assigned greater weights and items of large magnitude are assigned lesser weights.
3. $\bar{X} < \bar{X}_w$, when items of small magnitude are assigned lesser weights and items of large magnitude are assigned greater weights.

4.3.3 Properties of Arithmetic Mean

Arithmetic mean of a given data possesses the following properties:

1. The sum of deviations of the observations from their arithmetic mean is always zero. According to this property, the arithmetic mean serves as a point of balance or a centre of gravity of the distribution; since sum of positive deviations (i.e., deviations of observations which are greater than $\bar{X}$) is equal to the sum of negative deviations (i.e., deviations of observations which are less than $\bar{X}$).
2. The sum of squares of deviations of observations is minimum when taken from their arithmetic mean. Because of this, the mean is sometimes termed as 'least square' measure of central tendency.
3. Arithmetic mean is capable of being treated algebraically.

   This property of arithmetic mean highlights the relationship between $\bar{X}$, $\sum f_i X_i$ and N. According to this property, if any two of the three values are known, the third can be easily computed.

4. If $\overline{X}_1$ and N_1 are the mean and number of observations of a series and $\overline{X}_2$ and N_2 are the corresponding magnitudes of another series, then the mean $\overline{X}$ of the combined series of $N_1 + N_2$ observations is given by

$$\overline{X} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

5. If a constant B is added (subtracted) from every observation, the mean of these observations also gets added (subtracted) by it.
6. If every observation is multiplied (divided) by a constant β, the mean of these observations also gets multiplied (divided) by it.
7. If some observations of a series are replaced by some other observations, then the mean of original observations will change by the average change in magnitude of the changed observations.

Example: Find out the missing item (x) of the following frequency distribution whose arithmetic mean is 11.37.

X	:	5	7	(*x*)	11	13	16	20
f	:	2	4	29	54	11	8	4

Solution:

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{(5\times 2) + (7\times 4) + 29x + (11\times 54) + (13\times 11) + (16\times 8) + (20\times 4)}{112}$$

$$11.37 = \frac{10 + 28 + 29x + 594 + 143 + 128 + 80}{112} \text{ or } 11.37\times 112 = 983 + 29x$$

$$\therefore x = \frac{290.44}{29} = 10.015 = 10 \text{ (approximately)}$$

Example: The arithmetic mean of 50 items of a series was calculated by a student as 20. However, it was later discovered that an item 25 was misread as 35. Find the correct value of mean.

Solution:

$N = 50$ and $\overline{X} = 20 \therefore \sum X_i = 50\times 20 = 1000$

Thus $\sum X_{i(corrected)} = 1000 + 25 - 35 = 990$ and $\overline{X}_{(corrected)} = \frac{990}{50} = 19.8$

Alternatively, using property 7:

$$\overline{X}_{new} = \overline{X}_{old} + \text{average change in magnitude} = 20 - \frac{10}{50} = 20 - 0.2 = 19.8$$

4.3.4 Merits and Demerits of Arithmetic Mean

Merits of Arithmetic Mean

Out of all averages arithmetic mean is the most popular average in statistics because of its merits given below:

- Arithmetic mean is rigidly defined by an algebraic formula.

- Calculation of arithmetic mean requires simple knowledge of addition, multiplication and division of numbers and hence, is easy to calculate. It is also simple to understand the meaning of arithmetic mean, e.g., the value per item or per unit, etc.
- Calculation of arithmetic mean is based on all the observations and hence, it can be regarded as representative of the given data.
- It is capable of being treated mathematically and hence, is widely used in statistical analysis.
- Arithmetic mean can be computed even if the detailed distribution is not known but sum of observations and number of observations are known.
- It is least affected by the fluctuations of sampling.
- It represents the centre of gravity of the distribution because it balances the magnitudes of observations which are greater and less than it.
- It provides a good basis for the comparison of two or more distributions.

Demerits of Arithmetic Mean

Although, arithmetic mean satisfies most of the properties of an ideal average, it has certain drawbacks and should be used with care. Some demerits of arithmetic mean are mentioned below:

- It can neither be determined by inspection nor by graphical location.
- Arithmetic mean cannot be computed for a qualitative data; like data on intelligence, honesty, smoking habit, etc.
- It is too much affected by extreme observations and hence, it does not adequately represent data consisting of some extreme observations.
- The value of mean obtained for a data may not be an observation of the data and as such it is called a fictitious average.
- Arithmetic mean cannot be computed when class intervals have open ends. To compute mean, some assumption regarding the width of class intervals is to be made.
- In the absence of a complete distribution of observations the arithmetic mean may lead to fallacious conclusions. For example, there may be two entirely different distributions with same value of arithmetic mean.
- Simple arithmetic mean gives greater importance to larger values and lesser importance to smaller values.

4.4 MEDIAN

Median of distribution is that value of the variate which divides it into two equal parts. In terms of frequency curve, the ordinate drawn at median divides the area under the curve into two equal parts. Median is a positional average because its value depends upon the position of an item and not on its magnitude.

4.4.1 Determination of Median

When Individual Observations are Given

The following steps are involved in the determination of median:

1. The given observations are arranged in either ascending or descending order of magnitude.

2. Given that there are n observations, the median is given by:

   (a) The size of $\left(\frac{n+1}{2}\right)$th observations, when n is odd.

   (b) The mean of the sizes of $\frac{n}{2}$th and $\left(\frac{n+1}{2}\right)$th observations, when n is even.

Example: Find median of the following observations:

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

Solution:

Writing the observations in ascending order, we get 15, 16, 18, 20, 25, 28, 30.

Since $n = 7$, i.e., odd, the median is the size of $\left(\frac{7+1}{2}\right)$ i.e., 4th observation.

Hence, median, denoted by $M_d = 20$.

Note: The same value of M_d will be obtained by arranging the observations in descending order of magnitude.

Example: Find median of the data : 245, 230, 265, 236, 220, 250

Solution:

Arranging these observations in ascending order of magnitude, we get

220, 230, 236, 245, 250, 265. Here n = 6, i.e., even.

Median will be arithmetic mean of the size of $\frac{6}{2}$th, i.e., 3rd and $\left(\frac{6}{2}+1\right)$th, i.e., 4th observations.

Hence, $M_d = \frac{236+245}{2} = 240.5$

When Ungrouped Frequency Distribution is Given

In this case, the data are already arranged in the order of magnitude. Here, cumulative frequency is computed and the median is determined in a manner similar to that of individual observations.

Example: Locate median of the following frequency distribution:

Variable (X) :	10	11	12	13	14	15	16
Frequency (f) :	8	15	25	20	12	10	5

Solution:

X :	10	11	12	13	14	15	16
f :	8	15	25	20	12	10	5
c. f. :	8	23	48	68	80	90	95

Here N = 95, which is odd. Thus, median is size of $\left[\frac{95+1}{2}\right]^{th}$ i.e., 48^{th} observation.

From the table 48^{th} observation is 12,

$M_d = \mathbf{12}$

Alternative Method

$\frac{N}{2} = \frac{95}{2} = 47.5$ Looking at the frequency distribution we note that there are 48 observations which are less than or equal to 12 and there are 72 (i.e., 95 – 23) observations which are greater than or equal to 12. Hence, median is 12.

Example: Locate median of the following frequency distribution:

X :	0	1	2	3	4	5	6	7
f :	7	14	18	36	51	54	52	20

Solution:

X	0	1	2	3	4	5	6	7
f	7	14	18	36	51	54	52	20
c.f.	7	21	39	75	126	180	232	252

Here N = 252, i.e., even.

Now $\frac{N}{2} = \frac{252}{2} = 126$ and $\frac{N}{2} + 1 = 127.$

Median is the mean of the size of 126^{th} and 127^{th} observation. From the table, we note that 126^{th} observation is 4 and 127^{th} observation is 5.

$$M_d = \frac{4+5}{2} = 4.5$$

Alternative Method

Looking at the frequency distribution we note that there are 126 observations which are less than or equal to 4 and there are 252 – 75 = 177 observations which are greater than or equal to 4. Similarly, observation 5 also satisfies this criterion. Therefore, median $\frac{4+5}{2} = 4.5$.

When Grouped Frequency Distribution is Given (Interpolation Formula)

The determination of median, in this case, will be explained with the help of the following example.

Example: The following table shows the daily sales of 230 footpath sellers of Chandni Chowk:

Sales (in ₹) :	0-500	500-1000	1000-1500	1500-2000
No. of Sellers :	12	18	35	42

Sales (in ₹) :	2000-2500	2500-3000	3000-3500	3500-4000
No. of Sellers :	50	45	20	8

Locate the median of the above data using

(1) only the less than type ogive, and

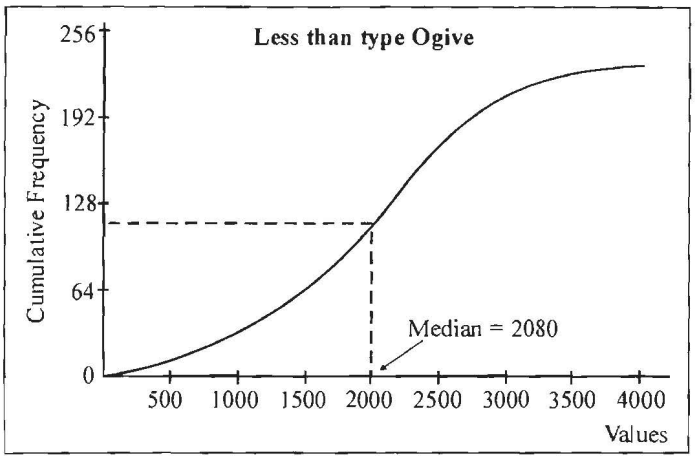
(2) both, the less than and the greater than type ogives.

Solution:

To draw ogives, we need to have a cumulative frequency distribution.

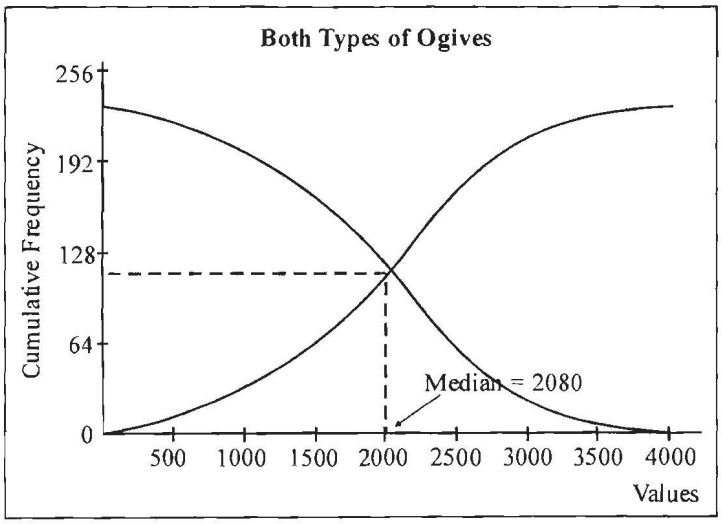
Class Intervals	Frequency	Less than c.f.	More than c.f.
0-500	12	12	230
500-1000	18	30	218
1000-1500	35	65	200
1500-2000	42	107	165
2000-2500	50	157	123
2500-3000	45	202	73
3000-3500	20	222	28
3500-4000	8	230	8

1. Using the less than type give



The value $\frac{N}{2}=115$ is marked on the vertical axis and a horizontal line is drawn from this point to meet the give at point S. Drop a perpendicular from *S*. The point at which this meets X-axis is the median.

2. Using both types of gives



A perpendicular is dropped from the point of intersection of the two ogives. The point at which it intersects the X-axis gives median. It is obvious from above figures that median = 2080.

4.4.2 Properties of Median

- It is a positional average.
- It can be shown that the sum of absolute deviations is the minimum when taken from median. This property implies that median is centrally located.

4.4.3 Merits, Demerits and Uses of Median

Merits

- It is easy to understand and easy to calculate, especially in series of individual observations and ungrouped frequency distributions. In such cases, it can even be located by inspection.
- Median can be determined even when class intervals have open ends or not of equal width.
- It is not much affected by extreme observations. It is also independent of range or dispersion of the data.
- Median can also be located graphically.
- It is centrally located measure of average since the sum of absolute deviation is the minimum when taken from median.
- It is the only suitable average when data are qualitative and it is possible to rank various items according to qualitative characteristics.
- Median conveys the idea of a typical observation.

Demerits

- In case of individual observations, the process of location of median requires their arrangement in the order of magnitude which may be a cumbersome task, particularly when the number of observations is very large.
- It, being a positional average, is not capable of being treated algebraically.
- In case of individual observations, when the number of observations is even, the median is estimated by taking mean of the two middle-most observations, which is not an actual observation of the given data.
- It is not based on the magnitudes of all the observations. There may be a situation where different sets of observations give same value of median. For example, the following two different sets of observations have median equal to 30.

Set I : 10, 20, 30, 40, 50 and Set II : 15, 25, 30, 60, 90.

- In comparison to arithmetic mean, it is much affected by the fluctuations of sampling.
- The formula for the computation of median, in case of grouped frequency distribution, is based on the assumption that the observations in the median class are uniformly distributed. This assumption is rarely met in practice.
- Since it is not possible to define weighted median like weighted arithmetic mean, this average is not suitable when different items are of unequal importance.

Uses

- It is an appropriate measure of central tendency when the characteristics are not measurable but different items are capable of being ranked.
- Median is used to convey the idea of a typical observation of the given data.
- Median is the most suitable measure of central tendency when the frequency distribution is skewed. For example, income distribution of the people is generally positively skewed and median is the most suitable measure of average in this case.
- Median is often computed when quick estimates of average are desired.
- When the given data has class intervals with open ends, median is preferred as a measure of central tendency since it is not possible to calculate mean in this case.

4.5 MODE

Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed. In the words of Croxton and Cowden, *"The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded the most typical of a series of values."* Further, according to A.M. Tuttle, *"Mode is the value which has the greatest frequency density in its immediate neighbourhood."*

If the frequency distribution is regular, then mode is determined by the value corresponding to maximum frequency. There may be a situation where concentration of observations around a value having maximum frequency is less than the concentration of observations around some other value. In such a situation, mode cannot be determined by the use of maximum frequency criterion. Further, there may be concentration of observations around more than one value of the variable and, accordingly, the distribution is said to be bi-model or multi-model depending upon whether it is around two or more than two values.

The concept of mode, as a measure of central tendency, is preferable to mean and median when it is desired to know the most typical value, e.g., the most common size of shoes, the most common size of a ready-made garment, the most common size of income, the most common size of pocket expenditure of a college student, the most common size of a family in a locality, the most common duration of cure of viral-fever, the most popular candidate in an election, etc.

4.5.1 Determination of Mode

When data are either in the form of individual observations or in the form of ungrouped frequency distribution

Given individual observations, these are first transformed into an ungrouped frequency distribution. The mode of an ungrouped frequency distribution can be determined in two ways, as given below:

- By inspection
- By method of grouping

By Inspection

When a frequency distribution is fairly regular, then mode is often determined by inspection. It is that value of the variate for which frequency is maximum. By a fairly regular frequency distribution, we mean that as the values of the variable increase the corresponding frequencies of these values first increase in a gradual manner and reach

a peak at certain value and, finally, start declining gradually in, approximately, the same manner as in case of increase.

Example: Compute mode of the following data:

3, 4, 5, 10, 15, 3, 6, 7, 9, 12, 10, 16, 18,
20, 10, 9, 8, 19, 11, 14, 10, 13, 17, 9, 11

Solution:

Writing this in the form of a frequency distribution, we get

Values	: 3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequency	: 2	1	1	1	1	1	3	4	2	1	1	1	1	1	1	1	1	1

Mode = 10

Remarks:

- If the frequency of each possible value of the variable is same, there is no mode.
- If there are two values having maximum frequency, the distribution is said to be bi-modal.

By method of Grouping

This method is used when the frequency distribution is not regular. Let us consider the following example to illustrate this method.

Example: Determine the mode of the following distribution:

X	:	10	11	12	13	14	15	16	17	18	19
f	:	8	15	20	100	98	95	90	75	50	30

Solution:

This distribution is not regular because there is sudden increase in frequency from 20 to 100. Therefore, mode cannot be located by inspection and hence the method of grouping is used. Various steps involved in this method are as follows:

- Prepare a table consisting of 6 columns in addition to a column for various values of X.
- In the first column, write the frequencies against various values of X as given in the question.
- In second column, the sum of frequencies, starting from the top and grouped in twos, are written.
- In third column, the sum of frequencies starting from the second and grouped in twos are written.
- In fourth column, the sum of frequencies, starting from the top and grouped in threes are written.
- In fifth column, the sum of frequencies, starting from the second and grouped in threes are written.
- In the sixth column, the sum of frequencies, starting from the third and grouped in threes are written.

The highest frequency total in each of the six columns is identified and analysed to determine mode. We apply this method for determining mode of the above example.

X	f (1)	(2)	(3)	(4)	(5)	(6)
10	8	23	35	43	135	218
11	15					
12	20	120	198	293	283	260
13	(100)					
14	98	193	185	215	155	
15	95					
16	90	165	125			
17	75					
18	50	80				
19	30					

Analysis Table

Columns	V	A	R	I	A	B	L	E	
	10	11	12	13	14	15	16	17	18	19
1				1						
2					1	1				
3				1	1					
4				1	1	1				
5					1	1	1			
6						1	1	1		
Total	**0**	**0**	**0**	**3**	**4**	**4**	**2**	**1**	**0**	**0**

Since the value 14 and 15 are both repeated maximum number of times in the analysis table, therefore, mode is ill defined. Mode in this case can be approximately located by the use of the following formula, which will be discussed later, in this lesson.

Mode = 3 Median – 2 mean

Calculation of Median and Mean

X	10	11	12	13	14	15	16	17	18	19	**Total**
f	8	15	20	100	98	95	90	75	50	30	**581**
c.f.	8	23	43	143	241	336	426	501	551	581	
fX	80	165	240	1300	1372	1425	1440	1275	900	570	**8767**

Median = Size of (581+1/2)th, i.e., 291st observation = 15.

Mean = 8767/581 = 15.09

Mode = 3 × 15 – 2 × 15.09 = 45 – 30.18 = 14.82

Remarks: If the most repeated values, in the above analysis table, were not adjacent, the distribution would have been bi-modal, i.e., having two modes.

Example: The monthly profits (in ₹) of 100 shops are distributed as follows:

Profit per Shop :	0-100	100-200	200-300	300-400	400-500	500-600
No. of Shops :	12	18	27	20	17	6

Determine the 'modal value' of the distribution graphically and verify the result by calculation.

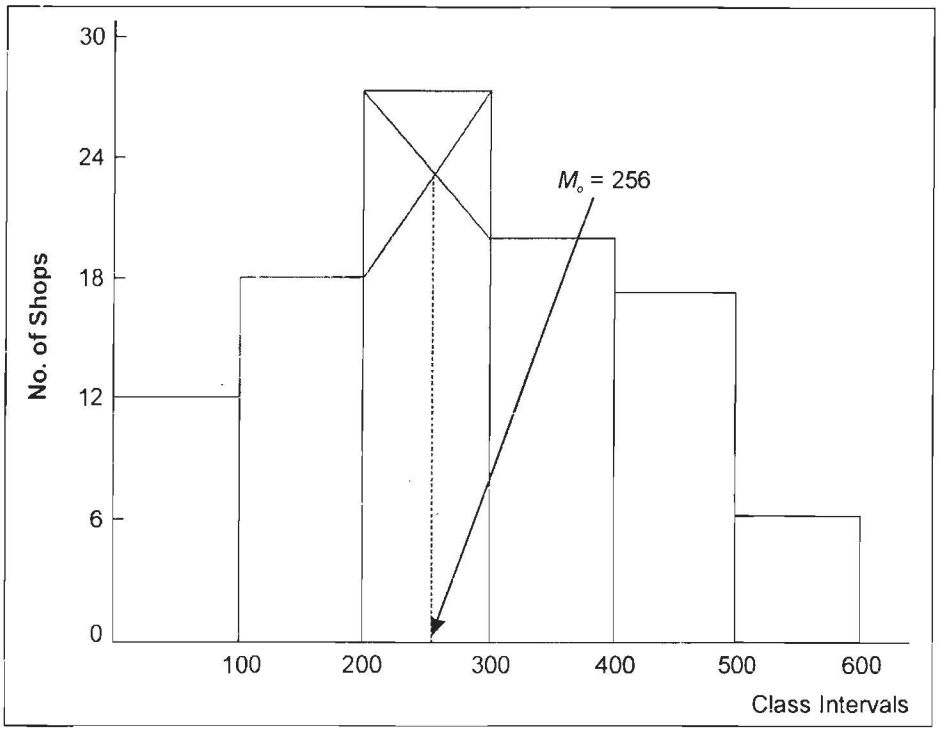
Solution:

Since the distribution is regular, the modal class would be a class having the highest frequency. The modal class, of the given distribution, is 200-300.

Graphical Location of Mode

To locate mode we draw a histogram of the given frequency distribution. The mode is located as shown in figure.

Graphical Location of Mode



From the figure, mode = ₹ 256.

Determination of Mode by Interpolation Formula

Since the modal class is 200-300, $L_m = 200$, $\Delta_1 = 27 - 18 = 9$, $\Delta_2 = 27 - 20 = 7$ and $h = 100$.

$\therefore$ $M_o = 200 + 9/9 + 7 \times 100 = ₹256.25$.

4.5.2 Merits and Demerits of Mode

Merits

- It is easy to understand and easy to calculate. In many cases, it can be located just by inspection.
- It can be located in situations where the variable is not measurable but categorisation or ranking of observations is possible.
- Like mean or median, it is not affected by extreme observations. It can be calculated even if these extreme observations are not known.
- It can be determined even if the distribution has open end classes.
- It can be located even when the class intervals are of unequal width provided that the width of modal and that of its preceding and following classes are equal.
- It is a value around which there is more concentration of observations and hence the best representative of the data.

Demerits

- It is not based on all the observations.
- It is not capable of further mathematical treatment.
- In certain cases mode is not rigidly defined and hence, the important requisite of a good measure of central tendency is not satisfied.
- It is much affected by the fluctuations of sampling.
- It is not easy to calculate unless the number of observations is sufficiently large and reveal a marked tendency of concentration around a particular value.
- It is not suitable when different items of the data are of unequal importance.
- It is an unstable average because, mode of a distribution, depends upon the choice of width of class intervals.

4.5.3 Relation among Mean, Median and Mode

The relationship between the above measures of central tendency will be interpreted in terms of a continuous frequency curve.

If the number of observations of a frequency distribution is increased gradually, then accordingly, we need to have more number of classes, for approximately the same range of values of the variable, and simultaneously, the width of the corresponding classes would decrease. Consequently, the histogram of the frequency distribution will get transformed into a smooth frequency curve, as shown in Figure 4.1.

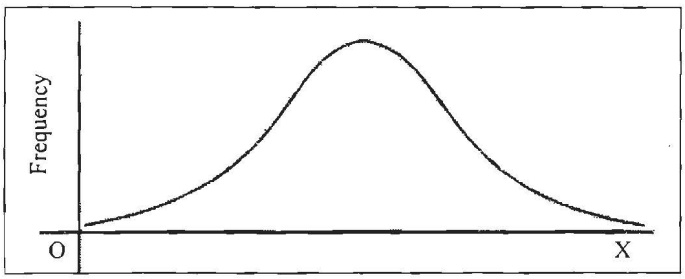


Figure 4.1: Frequency Curve

For a given distribution, the mean is the value of the variable which is the point of balance or centre of gravity of the distribution. The median is the value such that half of the observations are below it and remaining half are above it. In terms of the frequency curve, the total area under the curve is divided into two equal parts by the ordinate at median. Mode of a distribution is a value around which there is maximum concentration of observations and is given by the point at which peak of the curve occurs.

For a symmetrical distribution, all the three measures of central tendency are equal i.e. $\overline{X} = M_d = M_o$, as shown in Figure 4.2.

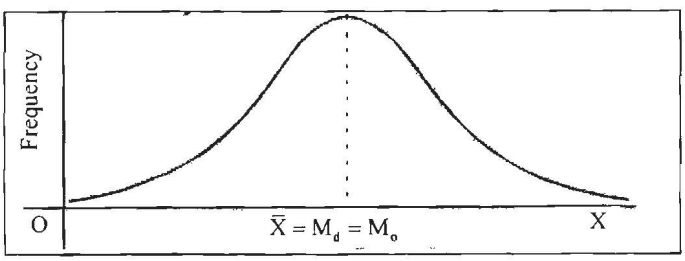


Figure 4.2: Symmetrical Distribution

Imagine a situation in which the above symmetrical distribution is made asymmetrical or positively (or negatively) skewed by adding some observations of very high (or very low) magnitudes, so that the right hand (or the left hand) tail of the frequency curve gets elongated. Consequently, the three measures will depart from each other. Since mean takes into account the magnitudes of observations, it would be highly affected. Further, since the total number of observations will also increase, the median would also be affected but to a lesser extent than mean. Finally, there would be no change in the position of mode. More specifically, we shall have $M_o < M_d < \bar{X}$, when skewness is positive and $\bar{X} < M_d < M_o$, when skewness is negative, as shown in Figure 4.3.

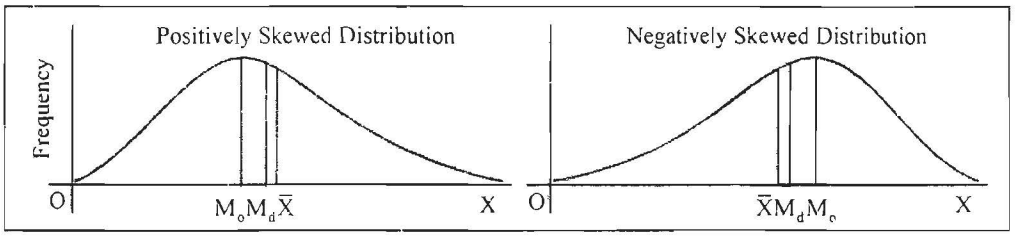


Figure 4.3: Positively and Negatively Skewed Distribution

4.5.4 Empirical Relation among Mean, Median and Mode

Empirically, it has been observed that for a moderately skewed distribution, the difference between mean and mode is approximately three times the difference between mean and median, i.e.,

$$\bar{X} - M_o = 3\,(X - M_d)$$

This relation can be used to estimate the value of one of the measures when the values of the other two are known.

Example:

(a) The mean and median of a moderately skewed distribution are 42.2 and 41.9 respectively. Find mode of the distribution.

(b) For a moderately skewed distribution, the median price of men's shoes is ₹ 380 and modal price is ₹ 350. Calculate mean price of shoes.

Solution:

(a) Here, mode will be determined by the use of empirical formula.

$$\bar{X} - M_o = 3(\bar{X} - M_d) \text{ or } M_o = 3M_d - 2\bar{X}$$

It is given that $\bar{X} = 42.2$ and $M_d = 41.9$

$M_o = 3 \times 41.9 - 2 \times 42.2 = 125.7 - 84.4 = 41.3$

(b) Using the empirical relation, we can write $\bar{X} = \dfrac{3M_d - M_o}{2}$

It is given that $M_d = ₹\,380$ and $M_o = ₹\,350$

$$\bar{X} = \frac{3 \times 380 - 350}{2} = ₹\,395$$

4.6 GEOMETRIC MEAN

The geometric mean of a series of n positive observations is defined as the n^{th} root of their product.

4.6.1 Calculation of Geometric Mean

Individual Series

If there are n observations, $X_1, X_2, \ldots\ldots X_n$, such that $X_i > 0$ for each i, their Geometric Mean (GM) is defined as:

$$GM = (X_1, X_2 \ldots X_n)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$$

where the symbol Π is used to denote the product of observations.

To evaluate GM, we have to use logarithms. Taking log of both sides, we have

$$\log (GM) = \frac{1}{n} \log (X_1, X_2 \ldots X_n)$$

$$= \frac{1}{n}\left[\log X_1 + \log X_2 + \ldots + \log X_n)\right] = \frac{\sum \log X_i}{n}$$

Taking antilog of both sides, we have

$$GM = \text{antilog}\left[\frac{\sum \log X_i}{n}\right]$$

This result shows that the GM of a set of observations is the antilog of the arithmetic mean of their logarithms.

Example: Calculate geometric mean of the following data:

1, 7, 29, 92, 115 and 375

Solution:

Calculation of Geometric Mean

X	1	7	29	92	115	375	**Σlog X**
log X	0.0000	0.8451	1.4624	1.9638	2.0607	2.5740	**8.9060**

$$GM = \text{antilog}\left[\frac{\sum \log X}{n}\right] = \text{antilog}\left[\frac{8.9060}{6}\right] = 30.50$$

$$GM = \text{antilog} = 30.50$$

Ungrouped Frequency Distribution

If the data consists of observations $X_1, X_2, \ldots\ldots X_n$ with respective frequencies $f_1, f_2, \ldots\ldots f_n$, where $\sum_{i=1}^{n} f_i = N$, the geometric mean is given by:

$$GM = \left[\underbrace{X_1, X_1, \ldots X_1}_{f_1 \text{ times}} \underbrace{X_2 \ldots X_2}_{f_2 \text{ times}} \ldots\ldots \underbrace{X_n, \ldots X_n}_{f_n \text{ times}}\right]^{\frac{1}{N}} = \left[X_1^{f_1} X_2^{f_2} \ldots\ldots X_n^{f_n}\right]^{\frac{1}{N}}$$

Taking log of both sides, we have

$$\log (GM) = \frac{1}{N}\left[\log X_1^{f_1} + \log X_2^{f_2} + ... + \log X_n^{f_n}\right]$$

$$= \frac{1}{N}\left[f_1 \log X_1 + f_2 \log X_2 + ... + f_n \log X_n\right]$$

$$= \frac{\sum_{i=1}^{n} f_i \log X_i}{N}$$

or GM = antilog $\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log X_i\right)$ which is again equal to the antilog of the arithmetic mean of the logarithm of observations.

Example: Calculate geometric mean of the following distribution:

X :	5	10	15	20	25	30
F :	13	18	50	40	10	6

Solution:

Calculation of GM

X	f	log X	f log X
5	13	0.6990	9.0870
10	18	1.0000	18.0000
15	50	1.1761	58.8050
20	40	1.3010	52.0400
25	10	1.3979	13.9790
30	6	1.4771	8.8626
Total	**137**		**160.7736**

∴ GM = antilog (160.7736/137) = antilog 1.1735 = 14.91.

4.6.2 Continuous Frequency Distribution

In case of a continuous frequency distribution, the class intervals are given. Let X_1, X_2, ...X_n denote the mid-values of the first, second ..... n^{th} class interval respectively with corresponding frequencies f_1, f_2, ... f_n, such that $\Sigma f_i = N$. The formula for calculation of GM is same as the formula used for an ungrouped frequency distribution

$$GM = \text{antilog}\left[\frac{\sum f_i \log X_i}{N}\right]$$

Example: Calculate geometric mean of the following distribution:

Class Intervals :	5-15	15-25	25-35	35-45	45-55
Frequencies :	10	22	25	20	8

Solution:

Calculation of GM

Class	f	Mid-value (X)	log X	f log X
5-15	10	10	1.0000	10.0000
15-25	22	20	1.3010	28.6227
25-35	25	30	1.4771	36.9280
35-45	20	40	1.6020	32.0412
45-55	8	50	1.6990	13.5918
Total	**85**			**121.1837**

GM = antilog 121.1837/85 = antilog 1.4257 = 26.65

4.6.3 Weighted Geometric Mean

If various observations, X_1, X_2, ...X_n, are not of equal importance in the data, weighted geometric mean is calculated. Weighted GM of the observations X_1, X_2, ...X_n with respective weights as w_1, w_2 .....w_n is given by:

$$GM = \text{antilog}\left[\frac{\sum w_i \log X_i}{\sum w_i}\right]$$

i.e., weighted geometric mean of observations is equal to the antilog of weighted arithmetic mean of their logarithms.

Example: Calculate weighted geometric mean of the following data:

Variable (X)	:	5	8	44	160	500
Weights (w)	:	10	9	3	2	1

How does it differ from simple geometric mean?

Solution:

Calculation of Weighted and Simple GM

X	Weight (w)	log (X)	w log X
5	10	0.6990	6.9900
8	9	0.9031	8.1278
44	3	1.6435	4.9304
160	2	2.2041	4.4082
500	1	2.6990	2.6990
Total	**25**	**8.1487**	**27.1554**

Weighted GM = antilog 27.1554/25 = antilog 1.0862 = 12.20

Simple GM = antilog 8.1487/ 5 (n = 5) = antilog 1.6297 = 42.63

Note that the simple GM is greater than the weighted GM because the given system of weights assigns more importance to values having smaller magnitude.

4.6.4 Geometric Mean of the Combined Group

If G_1, G_2, ... G_k are the geometric means of k groups having n_1, n_2, ... n_k observations respectively, the geometric mean G of the combined group consisting of $n_1 + n_2 + ... + n_k$ observations is given by

$$GM = \text{antilog}\left[\frac{n_1 \log G_1 + n_2 \log G_2 + ... + n_k \log G_k}{n_1 + n_2 + ... + n_k}\right] \text{antilog}\left[\frac{\sum n_i \log G_i}{\sum n_i}\right]$$

Example: If the geometric means of two groups consisting of 10 and 25 observations are 90.4 and 125.5 respectively, find the geometric mean of all the 35 observations combined into a single group.

Solution:

$$\text{Combined GM} = \text{antilog}\left[\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right]$$

Here $n_1 = 10$, $G_1 = 90.4$, $n_2 = 25$ and $G_2 = 125.5$

$$GM = \text{antilog}\left[\frac{10 \log 90.4 + 25 \log 125.5}{35}\right]$$

$$= \text{antilog}\left[\frac{10 \times 1.9562 + 25 \times 2.0986}{35}\right] = \text{antilog } 2.0579 = 114.27$$

To determine the average rate of change of price for the entire period when the rate of change of prices for different periods are given

Let P_0 be the price of a commodity in the beginning of the first year. If it increases by k_1 % in the first year, the price at the end of 1st year (or beginning of second year) is given by

$$P_1 = P_0 + P_0 \frac{k_1}{100} = P_0\left(1 + \frac{k_1}{100}\right) = P_0(1 + r_1)$$

where $r_1 = k_1/100$ denotes the rate of increase per rupee in first year. Similarly, if the price changes by k_2% in second year, the price at the end of second year is given by

$$P_2 = P_1 + P_1 \frac{k_2}{100} = P_1\left(1 + \frac{k_2}{100}\right) = P_1(1 + r_2)$$

Replacing the value of P_1 as $P_0(1 + r_1)$ we can write

$$P_2 = P_0(1 + r_1)(1 + r_2)$$

Proceeding in this way, if $100r_i$% is the rate of change of price in the ith year, the price at the end of nth period, P_n, is given by

$$P_n = P_0(1 + r_1)(1 + r_2) \ldots (1 + r_n) \qquad \ldots(1)$$

Further, let 100 r % per year be the average rate of increase of price that gives the price P_n at the end of n years. Therefore, we can write

$$P_n = P_0(1 + r)(1 + r) ..... (1 + r) = P_0(1 + r)^n \qquad \ldots(2)$$

Equating (1) and (2), we can write

$$(1 + r)^n = (1 + r_1)(1 + r_2) ..... (1 + r_n)$$

Or

$$(1+r)=\left[(1+r_1)(1+r_2).....(1+r_n)\right]^{\frac{1}{n}} \quad ...(3)$$

This shows that $(1 + r)$ is geometric mean of $(1 + r_1)$, $(1 + r_2)$, ... and $(1 + r_n)$.

From (3), we get

$$r=\left[(1+r_1)(1+r_2).....(1+r_n)\right]^{\frac{1}{n}}-1 \quad ...(4)$$

In the above equation, r denotes the per unit rate of change. This rate is termed as the rate of increase or the rate of growth if positive and the rate of decrease or the rate of decay if negative.

4.6.5 Average Rate of Growth of Population

The average rate of growth of price, denoted by r in the above section, can also be interpreted as the average rate of growth of population. If P_0 denotes the population in the beginning of the period and P_n the population after n years, using Equation (2), we can write the expression for the average rate of change of population per annum as

$$r=\left(\frac{P_n}{P_0}\right)^{\frac{1}{n}}-1$$

Similarly, Equation (4), given above, can be used to find the average rate of growth of population when its rates of growth in various years are given.

Remarks: The formulae of price and population changes, considered above, can also be extended to various other situations like growth of money, capital, output, etc.

Example: The population of a country increased from 2,00,000 to 2,40,000 within a period of 10 years. Find the average rate of growth of population per year.

Solution:

Let r be the average rate of growth of population per year for the period of 10 years. Let P_0 be initial and P_{10} be the final population for this period.

We are given $P_0 = 2,00,000$ and $P_{10} = 2,40,000$.

$$\therefore \quad r=\left(\frac{P_{10}}{P_0}\right)^{\frac{1}{10}}-1=\left(\frac{2,40,000}{2,00,000}\right)^{\frac{1}{10}}-1$$

$$\text{Now } \left(\frac{24}{20}\right)^{\frac{1}{10}}=\text{antilog}\left[\frac{1}{10}(\log 24-\log 20)\right]$$

$$=\text{antilog}\left[\frac{1}{10}(1.3802-1.3010)\right]=\text{antilog}(0.0079)=1.018$$

Thus, $r = 1.018 - 1 = 0.018$.

Hence, the percentage rate of growth $= 0.018 \times 100 = 1.8\%$ p.a.

4.6.6 Suitability of Geometric Mean for Averaging Ratios

It will be shown here that the geometric mean is more appropriate than arithmetic mean while averaging ratios.

Let there be two values of each of the variables x and y, as given below:

x	y	Ratio $\left(\frac{x}{y}\right)$	Ratio $\left(\frac{y}{x}\right)$
40	60	2/3	3/2
20	80	1/4	4

Now AM of (x/y) ratios = (2/3+1/4)/2 = 11/24 and the AM of (y/x) ratios = (3/2 +4)/2 = 11/4. We note that their product is not equal to unity.

However, the product of their respective geometric means, i.e., $\frac{1}{\sqrt{6}}$ and $\sqrt{6}$ is equal to unity.

Since it is desirable that a method of average should be independent of the way in which a ratio is expressed, it seems reasonable to regard geometric mean as more appropriate than arithmetic mean while averaging ratios.

4.6.7 Properties of Geometric Mean

1. As in case of arithmetic mean, the sum of deviations of logarithms of values from the log GM is equal to zero.

   This property implies that the product of the ratios of GM to each observation, that is less than it, is equal to the product the ratios of each observation to GM that is greater than it. For example, if the observations are 5, 25, 125 and 625, their GM = 55.9. The above property implies that

$$\frac{55.9}{5} \times \frac{55.9}{25} = \frac{125}{55.9} \times \frac{625}{55.9}$$

2. Similar to the arithmetic mean, where the sum of observations remains unaltered if each observation is replaced by their AM, the product of observations remains unaltered if each observation is replaced by their GM.

4.6.8 Merits, Demerits and Uses of Geometric Mean

Merits

- It is a rigidly defined average.
- It is based on all the observations.
- It is capable of mathematical treatment. If any two out of the three values, i.e., (i) product of observations, (ii) GM of observations and (iii) number of observations, are known, the third can be calculated.
- In contrast to AM, it is less affected by extreme observations.
- It gives more weights to smaller observations and vice-versa.

Demerits

- It is not very easy to calculate and hence not very popular.
- Like AM, it may be a value which does not exist in the set of given observations.

Uses

- It is most suitable for averaging ratios and exponential rates of changes.
- It is used in the construction of index numbers.
- It is often used to study certain social or economic phenomena.

4.7 HARMONIC MEAN

The harmonic mean of n observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals.

4.7.1 Calculation of Harmonic Mean

Individual Series

If there are n observations $X_1, X_2, \ldots\ldots X_n$, their harmonic mean is defined as

$$HM = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots\ldots + \frac{1}{X_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

Example: Obtain harmonic mean of 15, 18, 23, 25 and 30.

Solution:

$$HM = \frac{5}{\frac{1}{15} + \frac{1}{18} + \frac{1}{23} + \frac{1}{25} + \frac{1}{30}} = \frac{5}{0.239} = 20.92$$

Ungrouped Frequency Distribution

For ungrouped data, i.e., each $X_1, X_2, \ldots\ldots X_n$, occur with respective frequency $f_1, f_2 \ldots\ldots$ f_n, where $\Sigma f_i = N$ is total frequency, the arithmetic mean of the reciprocals of observations is given by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{f_i}{X_i}$$

Thus, $HM = \dfrac{N}{\sum \frac{f_i}{X_i}}$

Example: Draw a blank table to show the population of a city according to age, sex and unemployment in various years.

Solution:

Note: The table can be extended for the years 2012, 2013....., etc.

Example: Calculate harmonic mean of the following data:

X :	10	11	12	13	14
f :	5	8	10	9	6

Solution:

Calculation of Harmonic Mean

X	10	11	12	13	14	**Total**
Frequency (f)	5	8	10	9	6	**38**
$\mathbf{f \times \frac{1}{X}}$	0.5000	0.7273	0.8333	0.6923	0.4286	**3.1815**

$\therefore$ HM = 38/3.1815 = 11.94

Continuous Frequency Distribution

In case of a continuous frequency distribution, the class intervals are given. The mid-values of the first, second ..... nth classes are denoted by $X_1, X_2, \ldots\ldots X_n$. The formula for the harmonic mean is same.

Example: Find the harmonic mean of the following distribution:

Class Intervals :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency :	5	8	11	21	35	30	22	18

Solution:

Calculation of Harmonic Mean

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	**Total**
Frequency (f)	5	8	11	21	35	30	22	18	**150**
Mid-value (X)	5	15	25	35	45	55	65	75	
$\frac{f}{X}$	1.0000	0.5333	0.4400	0.6000	0.7778	0.5455	0.3385	0.2400	**4.4751**

$\therefore$ HM = 150/4. 4751 = 33.52

4.7.2 Weighted Harmonic Mean

If $X_1, X_2, \ldots\ldots X_n$ are n observations with weights $w_1, w_2, \ldots\ldots w_n$ respectively, their weighted harmonic mean is defined as follows:

$$HM = \frac{\sum w_i}{\sum \frac{w_i}{X_i}}$$

Example: A train travels 50 kms at a speed of 40 kms/hour, 60 kms at a speed of 50 kms/ hour and 40 kms at a speed of 60 kms/hour. Calculate the weighted harmonic mean of the speed of the train taking distances travelled as weights. Verify that this harmonic mean represents an appropriate average of the speed of train.

Solution:

$$HM = \frac{\sum w_i}{\sum \frac{w_i}{X_i}} = \frac{150}{\frac{50}{40} + \frac{60}{50} + \frac{40}{60}} = \frac{150}{1.25 + 1.20 + 0.67} \quad \ldots (1)$$

$= 48.13$ kms/hour

Verification: Average speed = Total distance travelled/Total time taken

We note that the numerator of Equation (1) gives the total distance travelled by train. Further, its denominator represents total time taken by the train in travelling 150 kms, since 50/40 is time taken by the train in travelling 50 kms at a speed of 40 kms/hour. Similarly, 60/50 and 40/60 are time taken by the train in travelling 60 kms and 40 kms at the speeds of 50 kms/hour and 60 kms/hour respectively. Hence, weighted harmonic mean is most appropriate average in this case.

Example: Ram goes from his house to office on a cycle at a speed of 12 kms/hour and returns at a speed of 14 kms/hour. Find his average speed.

Solution:

Since the distances of travel at various speeds are equal, the average speed of Ram will be given by the simple harmonic mean of the given speeds.

$$\text{Average speed} = \frac{2}{\frac{1}{12}+\frac{1}{14}} = \frac{2}{0.1547} = 12.92 \text{ kms/hour}$$

4.7.3 Merits and Demerits of Harmonic Mean

Merits

- It is a rigidly defined average.
- It is based on all the observations.
- It gives less weight to large items and vice-versa.
- It is capable of further mathematical treatment.
- It is suitable in computing average rate under certain conditions.

Demerits

- It is not easy to compute and is difficult to understand.
- It may not be an actual item of the given observations.
- It cannot be calculated if one or more observations are equal to zero.
- It may not be representative of the data if small observations are given correspondingly small weights.

Check Your Progress

Fill in the blanks:

1. _____ is defined as the sum of observations divided by the number of observations.
2. The sum of deviations of the observations from their arithmetic mean is always _____.
3. _____ of distribution is that value of the variate which divides it into two equal parts.
4. _____ is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.
5. The geometric mean of a series of n positive observations is defined as the _____ root of their product.
6. _____ denotes the population in the beginning of the period.

4.8 LET US SUM UP

- Summarization of the data is a necessary function of any statistical analysis.
- Average is a value which is typical or representative of a set of data.
- Arithmetic mean is defined as the sum of observations divided by the number of observations.

- In order that the mean wage gives a realistic picture of the distribution, the wages of managers should be given less importance in its computation. The mean calculated in this manner is called weighted arithmetic mean.
- If a constant B is added (subtracted) from every observation, the mean of these observations also gets added (subtracted) by it.
- If every observation is multiplied (divided) by a constant b, the mean of these observations also gets multiplied (divided) by it.
- If some observations of a series are replaced by some other observations, then the mean of original observations will change by the average change in magnitude of the changed observations.
- Arithmetic mean is rigidly defined by an algebraic formula.
- Median of distribution is that value of the variate which divides it into two equal parts.
- Median conveys the idea of a typical observation.
- Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.
- The geometric mean of a series of n positive observations is defined as the nth root of their product.
- The harmonic mean of *n* observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals.

4.9 UNIT END ACTIVITY

Premier Automobiles Ltd. does statistical analysis for an automobile racing team. Here are the fuel consumption figures in Kilometer per litre for the team's cars in the recent races.

4.77	6.11	6.11	5.05	5.99	4.91	5.27	6.01
5.75	4.89	6.05	5.22	6.02	5.24	6.11	5.02

(a) Calculate the median fuel consumption.

(b) Calculate the mean fuel consumption.

(c) Group the given data into equally sized classes. What is the fuel consumption value of the modal classes?

(d) Which of the three measures of central tendency is best for Allison to use when she orders fuel? Explain.

4.10 KEYWORDS

Average: An average is a single value within the range of the data that is used to represent all the values in the series.

Arithmetic Mean: Arithmetic mean is defined as the sum of observations divided by the number of observations.

Geometric Mean: The geometric mean of a series of n positive observations is defined as the nth root of their product.

Harmonic Mean: The harmonic mean of n observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals.

Measure of Central Tendency: A measure of central tendency is a typical value around which other figures congregate.

Measure of Central Value: Since an average is somewhere within the range of data it is sometimes called a measure of central value.

Median: Median of distribution is that value of the variate which divides it into two equal parts.

Mode: Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.

Weighted Arithmetic Mean: Weights are assigned to different items depending upon their importance, i.e., more important items are assigned more weight.

Weighted Geometric Mean: Weighted geometric mean of observations is equal to the antilog of weighted arithmetic mean of their logarithms.

4.11 QUESTIONS FOR DISCUSSION

1. What are the functions of an average?
2. Discuss the relative merits and demerits of various types of statistical averages.
3. Compute arithmetic mean of the following series:

Marks :	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students :	12	18	27	20	17	6

4. Calculate arithmetic mean of the following data:

Mid-values :	10	12	14	16	18	20
Frequency :	3	7	12	18	10	5

5. Find out the missing frequency in the following distribution with mean equal to 30.

Class :	0-10	10-20	20-30	30-40	40-50
Frequency :	5	6	10	?	13

6. A distribution consists of three components each with total frequency of 200, 250 and 300 and with means of 25, 10 and 15 respectively. Find out the mean of the combined distribution.
7. The mean of a certain number of items is 20. If an observation 25 is added to the data, the mean becomes 21. Find the number of items in the original data.
8. The mean age of a combined group of men and women is 30 years. If the mean age of the men's group is 32 years and that for the women's group is 27 years, find the percentage of men and women in the combined group.
9. Locate median of the following data:

   65, 85, 55, 75, 96, 76, 65, 60, 40, 85, 80, 125, 115, 40

10. Find out median from the following:

No. of Workers :	1-5	6-10	11-15	16-20	21-25
No. of Factories :	3	8	13	11	5

11. Find median from the following distribution:

X :	1	2	3	4	5-9	10-14	15-19	20-25
f :	5	10	16	20	30	15	8	6

12. Give the essential requisites of a measure of 'Central Tendency'. Under what circumstances would a geometric mean or a harmonic mean be more appropriate than arithmetic mean?
13. Determine the mode of the following data:

    58, 60, 31, 62, 48, 37, 78, 43, 65, 48
14. Find geometric mean from the following daily income (in ₹) of 10 families:

    85, 70, 15, 75, 500, 8, 45, 250, 40 and 36.
15. The price of a commodity increased by 12% in 1986, by 30% in 1987 and by 15% in 1988. Calculate the average increase of price per year.
16. The population of a city was 30 lakh in 1981 which increased to 45 lakh in 1991. Determine the rate of growth of population per annum. If the same growth continues, what will be the population of the city in 1995.
17. The value of a machine depreciated by 30% in 1st year, 13% in 2nd year and by 5% in each of the following three years. Determine the average rate of depreciation for the entire period.
18. The geometric means of three groups consisting of 15, 20 and 23 observations are 14.5, 30.2 and 28.8 respectively. Find geometric mean of the combined group.
19. A sum of money was invested for 3 years. The rates of interest in the first, second and third year were 10%, 12% and 14% respectively. Determine the average rate of interest per annum.
20. The weighted geometric mean of four numbers 8, 25, 17 and 30 is 15.3. If the weights of first three numbers are 5, 3 and 4 respectively, find the weight of the fourth number.
21. The annual rates of growth of output of a factory in five years are 5.0, 6.5, 4.5, 8.5 and 7.5 percent respectively. What is the compound rate of growth of output per annum for the period?

Check Your Progress: Model Answer

1. Arithmetic Mean
2. Zero
3. Median
4. Mode
5. n^{th}
6. P_0

4.12 REFERENCE & SUGGESTED READINGS

- Bowerman, B. L., O'Connell, R. T., Murphree, E. S., & Orris, J. B. (2018). **Essentials of Business Statistics** (6th ed.). McGraw-Hill Education. ISBN: 9781259549939
- Doane, D. P., & Seward, L. E. (2019). **Applied Statistics in Business and Economics** (6th ed.). McGraw-Hill Education. ISBN: 9781260224035
- Gupta, S. C., & Kapoor, V. K. (2018). **Fundamentals of Applied Statistics** (4th ed.). Sultan Chand & Sons. ISBN: 9788180547967

- Keller, G. (2022). **Business Analytics: A Data-Driven Decision Making Approach** (1st ed.). Cengage Learning. ISBN: 9780357717828
- Evans, J. R. (2020). **Business Analytics: Methods, Models, and Decisions** (2nd ed.). Pearson. ISBN: 9780135231679

BLOCK - 3

UNIT - V

MEASURES OF DISPERSION

CONTENTS

5.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Define the terms dispersion and range
- Discuss the objectives and characteristics of a good measure of dispersion

- State the merits and demerits of mean deviation
- Explain the concept of standard deviation
- Describe the coefficient of variation

5.1 INTRODUCTION

A measure of central tendency summarizes the distribution of a variable into a single figure which can be regarded as its representative. This measure alone, however, is not sufficient to describe a distribution because there may be a situation where two or more different distributions have the same central value. Conversely, it is possible that the pattern of distribution in two or more situations is same but the values of their central tendency are different. Hence, it is necessary to define some additional summary measures to adequately represent the characteristics of a distribution. One such measure is known as the measure of dispersion or the measure of variation. So far we have discussed the measures of central tendency and dispersion of frequency distributions for their summarization and comparison with each other.

5.2 DEFINITIONS OF DISPERSION

The concept of dispersion is related to the extent of scatter or variability in observations. The variability, in an observation, is often measured as its deviation from a central value. A suitable average of all such deviations is called the measure of dispersion.

Some important definitions of dispersion are given below:

"Dispersion is the measure of variation of the items."

– A.L. Bowley

"Dispersion is the measure of extent to which individual items vary."

– L.R. Connor

"The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation or dispersion."

– Simpson and Kafka

"The degree to which numerical data tend to spread about an average value is called variation or dispersion of the data." ***– Spiegel***

Measures of central tendency are known as the averages of first order. Measures of dispersion are known as the averages of second order.

The measures of central tendencies (i.e. means) indicate the general magnitude of the data and locate only the centre of a distribution of measures. They do not establish the degree of variability or the spread out or scatter of the individual items and their deviation from (or the difference with) the means.

According to **Nciswanger**, *"Two distributions of statistical data may be symmetrical and have common means, medians and modes and identical frequencies in the modal class. Yet with these points in common they may differ widely in the scatter or in their values about the measures of central tendencies."*

Simpson and **Kafka** said, *"An average alone does not tell the full story. It is hardly fully representative of a mass, unless we know the manner in which the individual item. Scatter around it .... a further description of a series is necessary, if we are to gauge how representative the average is."*

From this discussion, we now focus our attention on the scatter or variability which is known as dispersion. Let us take the following three sets.

Students	Group X	Group Y	Group Z
1	50	45	30
2	50	50	45
3	50	55	75
$\therefore$ mean $\bar{X} \Rightarrow$	50	50	50

Thus, the three groups have same mean i.e. 50. In fact, the median of group X and Y are also equal. Now if one would say that the students from the three groups are of equal capabilities, it is totally a wrong conclusion then. Close examination reveals that in group X students have equal marks as the mean, students from group Y are very close to the mean but in the third group Z, the marks are widely scattered. It is thus clear that the measures of the central tendency is alone not sufficient to describe the data.

5.3 OBJECTIVES OF MEASURING DISPERSION

The main objectives of measuring dispersion of a distribution are mentioned below:

1. ***To test reliability of an average:*** A measure of dispersion can be used to test the reliability of an average. A low value of dispersion implies that there is greater degree of homogeneity among various items and, consequently, their average can be taken as more reliable or representative of the distribution.
2. ***To compare the extent of variability in two or more distributions:*** The extent of variability in two or more distributions can be compared by computing their respective dispersions. A distribution having lower value of dispersion is said to be more uniform or consistent.
3. ***To facilitate the computations of other statistical measures:*** Measures of dispersions are used in computations of various important statistical measures like correlation, regression, test statistics, confidence intervals, control limits, etc.
4. ***To serve as the basis for control of variations:*** The main objective of computing a measure of dispersion is to know whether the given observations are uniform or not. This knowledge may be utilised in many ways. In the words of Spurr and Bonini, *"In matters of health, variations in body temperature, pulse beat and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variations. In industrial production, efficient operation requires control of quality variations, the causes of which are sought through inspection and quality control programs"*. The extent of inequalities of income and wealth in any society may help in the selection of an appropriate policy to control their variations.

5.4 MEASURES OF DISPERSION

Various measures of dispersion can be classified into two broad categories:

1. The measures which express the spread of observations in terms of distance between the values of selected observations. These are also termed as distance measures, e.g., range, interquartile range, inter percentile range, etc.
2. The measures which express the spread of observations in terms of the average of deviations of observations from some central value. These are also termed as the averages of second order, e.g., mean deviation, standard deviation, etc.

The following are some important measures of dispersion:

(a) Range

(b) Interquartile Range

(c) Mean Deviation

(d) Standard Deviation

5.5 RANGE

The range of a distribution is the difference between its two extreme observations, i.e., the difference between the largest and smallest observations. Symbolically,

$$R = L - S$$

where R denotes range, L and S denote largest and smallest observations, respectively. R is the absolute measure of range. A relative measure of range, also termed as the coefficient of range, is defined as:

$$\text{Coefficient of Range} = L - S / L + S$$

Example: Find range and coefficient of range for each of the following data:

1. Weekly wages of 10 workers of a factory are:

   310, 350, 420, 105, 115, 290, 245, 450, 300, 375.

2. The distribution of marks obtained by 100 students:

Marks	:	0-10	10-20	20-30	30-40	40-50
No. of Students	:	6	14	21	20	18
Marks	:	50-60	60-70	70-80	80-90	90-100
No. of Students	:	10	5	3	2	1

3. The age distribution of 60 school going children:

Age (in years)	:	5-7	8-10	11-13	14-16	17-19
Frequency	:	20	18	10	8	4

Solution:

1. Range = 450 – 105 = ₹ 345

   Coefficient of Range = 450 – 105/450 + 105 = 0.62

2. Range = 100 – 0 = 100 marks

   Coefficient of Range = 100 – 0/100 + 0 = 1

3. Range = 19 – 5 = 12 Years

   Coefficient of Range = 19 – 5/19 + 5 = 0.583

5.5.1 Merits and Demerits of Range

Merits

- It is easy to understand and easy to calculate.
- It gives a quick measure of variability.

Demerits

- It is not based on all the observations.
- It is very much affected by extreme observations.
- It only gives rough idea of spread of observations.
- It does not give any idea about the pattern of the distribution. There can be two distributions with the same range but different patterns of distribution.
- It is very much affected by fluctuations of sampling.
- It is not capable of being treated mathematically.
- It cannot be calculated for a distribution with open ends.

5.5.2 Uses of Range

In spite of many serious demerits, it is useful in the following situations:

- It is used in the preparation of control charts for controlling the quality of manufactured items.
- It is also used in the study of fluctuations of, say, price of a commodity, temperature of a patient, amount of rainfall in a given period, etc.

5.6 INTERQUARTILE RANGE

Interquartile range is an absolute measure of dispersion given by the difference between third quartile (Q_3) and first quartile (Q_1).

Symbolically, Interquartile range = $Q_3 - Q_1$.

5.6.1 Interpercentile Range

The difficulty of extreme observations can also be tackled by the use of interpercentile range or simply percentile range.

Symbolically, percentile range = $P_{(100-i)} - P_i$ $(i < 50)$.

This measure excludes i% of the observations at each end of the distribution and is a range of the middle (100 – 2i)% of the observations.

Normally, a percentile range corresponding to i = 10, i.e., $P_{90} - P_{10}$ is used. Since $Q_1 = P_{25}$ and $Q_3 = P_{75}$, therefore, interquartile range is also a percentile range.

Example: Determine the interquartile range and percentile range of the following distribution:

Class Intervals : 11-13 13-15 15-17 17-19 19-21 21-23 23-25

Frequency : 8 10 15 20 12 11 4

Solution:

Class Intervals	Frequency	Less than c.f.
11-13	8	8
13-15	10	18
15-17	15	33
17-19	20	53
19-21	12	65
21-23	11	76
23-25	4	80

1. Calculation of Interquartile Range

   Calculation of Q_1

   Since N/4 = 80/4 = 20, the first quartile class is 15-17

   $\therefore$ $l_{Q_1} = 15$, $f_{Q_1} = 15$, h = 2 and C = 18

   Hence, $Q_1 = 15 + (20 - 18)/15 \times 2 = 15.27$

   Calculation of Q_3

   Since $3N/4 = 3 \times 80/4 = 60$, the third quartile class is 19-21

   $\therefore$ $l_{Q_3} = 19$, $f_{Q_3} = 12$, h = 2 and C = 53

   Hence, $Q_3 = 19 + (60 - 53)/12 \times 2 = 20.17$

   Thus, the interquartile range = 20.17 – 15.27 = 4.90

2. Calculation of Percentile Range

   Calculation of P_{10}

   Since, $10N/100 = 10 \times 80/100 = 8$, P_{10} lies in the class interval 11-13

   $\therefore$ $l_{P_{10}} = 11$, $f_{P_{10}} = 8$, h = 2 and C = 0

   Hence,

$$P_{10} = 11 + \frac{8-0}{8} \times 2 = 13$$

   Calculation of P_{90}

   Since, $90N/100 = 90 \times 80/100 = 72$, P_{90} lies in the class interval 21-23

   $\therefore$ $l_{P_{90}} = 21$, $f_{P_{90}} = 11$, h = 2 and C = 65

   Hence,

$$P_{90} = 21 + \frac{72-65}{11} \times 2 = 22.27$$

   Thus, the percentile range = $P_{90} - P_{10} = 22.27 - 13.0 = 9.27$.

5.6.2 Quartile Deviation or Semi-interquartile Range

Half of the interquartile range is called the quartile deviation or semi-interquartile range.

Symbolically, Q.D. = $Q_3 - Q_1/2$

The value of Q.D. gives the average magnitude by which the two quartiles deviate from median. If the distribution is approximately symmetrical, then $M_d \pm$ Q.D. will include about 50% of the observations and, thus, we can write $Q_1 = M_d$ – Q.D. and $Q_3 = M_d$ + Q.D.

Further, a low value of Q.D. indicates a high concentration of central 50% observations and vice-versa. Quartile deviation is an absolute measure of dispersion.

The corresponding relative measure is known as coefficient of quartile deviation defined as

$$\text{Coefficient of Q.D.} = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Analogous to quartile deviation and the coefficient of quartile deviation we can also define a percentile deviation and coefficient of percentile deviation as

$$\frac{P_{100-i} - P_i}{2} \text{ and } \frac{P_{100-i} - P_i}{P_{100-i} + P_i}$$

Example: Find the quartile deviation, percentile deviation and their coefficients from the following data:

Age (in years)	:	15	16	17	18	19	20	21
No. of Students	:	4	6	10	15	12	9	4

Solution:

Table for the Calculation of Q.D. and P.D.

Age (X)	No. of Students (f)	Less than c.f.
15	4	4
16	6	10
17	10	20
18	15	35
19	12	47
20	9	56
21	4	60

We have,

$\frac{N}{4} = \frac{60}{4} = 15$ $\quad \therefore$ $Q_1 = 17$ (by inspection)

$\frac{3N}{4} = \frac{3 \times 60}{4} = 45$ $\quad \therefore$ $Q_3 = 19$ "

$\frac{10N}{100} = \frac{10 \times 60}{100} = 6$ $\quad \therefore$ $P_{10} = 16$ "

$\frac{90N}{100} = \frac{90 \times 60}{100} = 54$ $\quad \therefore$ $P_{90} = 20$ "

Thus, Q.D. = 19 – 17/2 = 1 year and P.D. = 20 – 16 /2 = 2 years

Also, Coefficient of Q.D. = 19 – 17/19 + 17 = 0.056

and Coefficient of P.D. = 20 – 16/20 + 16 = 0.11

5.6.3 Merits and Demerits of Quartile Deviation

Merits

- It is rigidly defined.
- It is easy to understand and easy to compute.

- It is not affected by extreme observations and hence a suitable measure of dispersion when a distribution is highly skewed.
- It can be calculated even for a distribution with open ends.

Demerits

- It is not based on all the observations, hence, not a reliable measure of dispersion.
- It is very much affected by the fluctuations of sampling.
- It is not capable of being treated mathematically.

5.7 MEAN DEVIATION OR AVERAGE DEVIATION

Mean deviation is a measure of dispersion based on all the observations. It is defined as the arithmetic mean of the absolute deviations of observations from a central value like mean, median or mode. Here the dispersion in each observation is measured by its deviation from a central value. This deviation will be positive for an observation greater than the central value and negative for less than it.

5.7.1 Calculation of Mean Deviation

The following are the formulae for the computation of mean deviation (M.D.) of an individual series of observations $X_1, X_2, \ldots\ldots X_n$:

1. M.D. from $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|$

2. M.D. from $M_d = \frac{1}{n}\sum_{i=1}^{n}|X_i - M_d|$

3. M.D. from $M_0 = \frac{1}{n}\sum_{i=1}^{n}|X_i - M_0|$

In case of an ungrouped frequency distribution, the observations $X_1, X_2, \ldots\ldots X_n$ occur with respective frequencies $f_1, f_2, \ldots\ldots f_n$ such that

$$\sum_{i=1}^{n} f_i = N$$

The corresponding formulae for M.D. can be written as:

1. M.D. from $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} f_i|X_i - \bar{X}|$

2. M.D. from $M_d = \frac{1}{n}\sum_{i=1}^{n} f_i|X_i - M_d|$

3. M.D. from $M_0 = \frac{1}{n}\sum_{i=1}^{n} f_i|X_i - M_0|$

The above formulae are also applicable to a grouped frequency distribution where the symbols $X_1, X_2, \ldots\ldots X_n$ will denote the mid-values of the first, second ..... nth classes respectively.

Remarks: We state without proof that the mean deviation is minimum when deviations are taken from median.

5.7.2 Coefficient of Mean Deviation

The above formulae for mean deviation give an absolute measure of dispersion. The formulae for relative measure, termed as the coefficient of mean deviation, are given below:

1. Coefficient of M.D. from $\bar{X} = \dfrac{M.D.\ from\ \bar{X}}{\bar{X}}$

2. Coefficient of M.D. from $M_d = \dfrac{M.D.\ from\ M_d}{M_d}$

3. Coefficient of M.D. from $M_0 = \dfrac{M.D.\ from\ M_0}{M_0}$

Example: Calculate mean deviation from mean and median for the following data of heights (in inches) of 10 persons.

60, 62, 70, 69, 63, 65, 60, 68, 63, 64

Also calculate their respective coefficients.

Solution:

Calculation of M.D. from $\bar{X}$

$\bar{X}$ = 60 + 62 + 70 + 69 + 63 + 65 + 60 + 68 + 63 + 64/10 = 64.4 inches

Sum of observations greater than $\bar{X}$ = 70 + 69 + 65 + 68 = 272

Sum of observations less than $\bar{X}$ = 60 + 62 + 63 + 60 + 63 + 64 = 372

Also, k_2= 4 and k_1 = 6

M.D. from $\bar{X}$ = 1/10 [272 – 372 – (4 – 6) 64.4] = 2.88 inches

Also, coefficient of M.D. from $\bar{X}$ = 2.88/64.4 = 0.045

Coefficient of M.D. from M_d

Arranging the observations in order of magnitude, we have

60, 60, 62, 63, 63, 64, 65, 68, 69, 70

The median of the above observations is = 63 + 64/2 = 63.5 inches

Sum of observations greater than M_d = 64 + 65 + 68 + 69 + 70 = 336

Sum of observations less than M_d = 60 + 60 + 62 + 63 + 63 = 308

Also, k_2 = 5 and k_1 = 5

M.D. from $M_d = \dfrac{[336 - 308 - (5-5)63.5]}{10}$ = 2.8 inches

Also, the coefficient of M.D. from M_d = 2.8/63.5 = 0.044

Example: Calculate mean deviation from median and its coefficient from the given data:

X	:	0	1	2	3	4	5	6	7	8	9
f	:	15	45	91	162	110	95	82	26	13	2

Solution:

Calculation of Mean Deviation

X	f	Less than c.f.	|X – 4|	f|X–4|
0	15	15	4	60
1	45	60	3	135
2	91	151	2	182
3	162	313	1	162
4	110	423	0	0
5	95	518	1	95
6	82	600	2	164
7	26	626	3	78
8	13	639	4	52
9	2	641	5	10
Total				**938**

Since N/2 = 641/2 = 320.5 $\therefore$ $M_d = 4$ (by inspection)

Thus, M.D. = 938/641 =1.46 and the coefficient of M.D. = 1.46/4 = 0.365

Example: Calculate mean deviation from median for the following data:

Class Intervals :	20-25	25-30	30-40	40-45	45-50	50-55	55-60	60-70	70-80
Frequency :	6	12	17	30	10	10	8	5	2

Also calculate the coefficient of Mean Deviation.

Solution:

Calculation of Median and Mean Deviation

Class Intervals	Frequency (f)	c.f.	Mid Values	fX
20-25	6	6	22.5	135.0
25-30	12	18	27.5	330.0
30-40	17	35	35.0	595.0
40-45	30	65	42.5	1275.0
45-50	10	75	47.5	475.0
50-55	10	85	52.5	525.0
55-60	8	93	57.5	460.0
60-70	5	98	65.0	325.0
70-80	2	100	75.0	150.0

Since N/2 = 50, the median class is 40 – 45.

$\therefore$ $L_m = 40$, $f_m = 30$, h = 5 and C = 35

Hence,

$$M_d = 40 + \frac{50 - 35}{30} \times 5 = 42.5$$

Calculation of M.D.

Sum of observations which are greater than M_d = 475 + 525 + 460 + 325 + 150 = 1,935

Sum of observations which are less than M_d = 135 + 330 + 595 = 1060

No. of observations which are greater than M_d, i.e., k_2 = 10 +10 + 8 + 5 + 2 = 35

No. of observations which are less than M_d, i.e., k_1 = 6 +12 +17 = 35

Therefore,

M.D. = 1935 – 1060/100 = 8.75 and the coefficient of M.D. = 8.75/42.5 = 0.206

5.7.3 Merits, Demerits and Uses of Mean Deviation

Merits

- It is easy to understand and easy to compute.
- It is based on all the observations.
- It is less affected by extreme observations vis-a-vis range or standard deviation (to be discussed in the next section).
- It is not much affected by fluctuations of sampling.

Demerits

- It is not capable of further mathematical treatment. Since mean deviation is the arithmetic mean of absolute values of deviations, it is not very convenient to be algebraically manipulated.
- This necessitates a search for a measure of dispersion which is capable of being subjected to further mathematical treatment.
- It is not well defined measure of dispersion since deviations can be taken from any measure of central tendency.

Uses

The mean deviation is a very useful measure of dispersion when sample size is small and no elaborate analysis of data is needed. Since standard deviation gives more importance to extreme observations the use of mean deviation is preferred in statistical analysis of certain economic, business and social phenomena.

5.8 STANDARD DEVIATION

From the mathematical point of view, the practice of ignoring minus sign of the deviations, while computing mean deviation, is very inconvenient and this makes the formula, for mean deviation, unsuitable for further mathematical treatment. Further, if the signs are taken into account, the sum of deviations taken from their arithmetic mean is zero. This would mean that there is no dispersion in the observations. However, the fact remains that various observations are different from each other. In order to escape this problem, the squares of the deviations from arithmetic mean are taken and the positive square root of the arithmetic mean of sum of squares of these deviations is taken as a measure of dispersion. This measure of dispersion is known as standard deviation or root-mean square deviation. Square of standard deviation is known as variance. The concept of standard deviation was introduced by Karl Pearson in 1893.

The standard deviation is denoted by Greek letter 'σ' which is called 'small sigma' or simply sigma.

In terms of symbols

$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, for n individual observations, $X_1, X_2, \ldots\ldots X_n$, and

$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \bar{X})^2}$, for a grouped or ungrouped frequency distribution, when an observation X_i occurs with frequency f_i for $i = 1, 2, \ldots n$ and $\sum_{i=1}^{n} f_i = N$.

It should be noted here that the units of σ are same as the units of X.

5.8.1 Calculation of Standard Deviation

There are two methods of calculating standard deviation: (i) Direct Method and (ii) Short-cut Method.

Direct Method

1. ***Individual Series:*** If there are n observations $X_1, X_2, \ldots\ldots X_n$, various steps in the calculation of standard deviation are:

   (a) Find $\bar{X} = \frac{\sum X_i}{n}$

   (b) Obtain deviations $(X_i - \bar{X})$ for each i = 1, 2, ... n.

   (c) Square these deviations and add to obtain $\sum_{i=1}^{n}(X_i - \bar{X})^2$

   (d) Compute variance, i.e., $\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$

   (e) Obtain σ as the positive square root of σ^2.

   The above method is appropriate when $\bar{X}$ is a whole number. If $\bar{X}$ is not a whole number, the standard deviation is conveniently computed by using the transformed form of the above formula, given below.

$$\sigma^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 = \frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})$$

$$\frac{1}{n}\sum(X_i^2 - \bar{X}X_i)^2 - \frac{\bar{X}}{n}\sum(X_i - \bar{X}) = \frac{1}{n}\sum X_i^2 - \bar{X}\frac{\sum X_i}{n}$$

(The 2nd term is sum of deviations from $\bar{X}$, which is equal to zero.)

$$= \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum X_i^2 - \left(\frac{\sum X_i}{n}\right)^2 \text{ or}$$

= Mean of squares – Square of mean.

Example: Calculate variance and standard deviation of the weights of ten persons.

Weights (in kgs) : 45, 49, 55, 50, 41, 44, 60, 58, 53, 55

Solution:

Calculation of Standard Deviation

Let $u = X - \bar{X}$

											Total
Weights (X)	45	49	55	50	41	44	60	58	53	55	**510**
u	–6	–2	4	–1	–10	–7	9	7	2	4	**0**
u^2	36	4	16	1	100	49	81	49	4	16	**356**

From the above table

$\bar{X} = 510/10 = 51$kgs and $\sigma^2 = 356/10 = 35.6$ kgs^2

2. ***Ungrouped or Grouped Frequency Distributions:*** Let the observations X_1, X_2 ..... X_n appear with respective frequencies f_1, f_2 ..... f_n, where $\Sigma f_i = N$. As before, if the distribution is grouped, then X_1, X_2 ..... X_n will denote the mid-values of the first, second ..... n^{th} class intervals respectively. The formulae for the calculation of variance and standard deviation can be written as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \bar{X})^2 \text{ and } \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \bar{X})^2}$$

Here also, we can show that

Variance = Mean of squares – Square of the mean

Therefore, we can write

$$\sigma^2 = \frac{\sum f_i X_i^2}{N} - \left(\frac{\sum f_i X_i}{N}\right)^2 \text{ and } \sigma = \sqrt{\frac{\sum f_i X_i^2}{N} - \left(\frac{\sum f_i X_i}{N}\right)^2}$$

Example: Calculate standard deviation of the following data:

X	:	10	11	12	13	14	15	16	17	18
f	:	2	7	10	12	15	11	10	6	3

Solution:

Calculation of Standard Deviation

Let $u = X - \bar{X}$

X	**f**	**fX**	**u**	**u^2**	**fu^2**	**fX^2**
10	2	20	–4	16	32	200
11	7	77	–3	9	63	847
12	10	120	–2	4	40	1440
13	12	156	–1	1	12	2028
14	15	210	0	0	0	2940
15	11	165	1	1	11	2475
16	10	160	2	4	40	2560
17	6	102	3	9	54	1734
18	3	54	4	16	48	972
Total	**76**	**1064**			**300**	**15196**

$\bar{X} = 1064/76 = 14$

$\sigma^2 = 300/76 = 3.95$ and $\sigma = \sqrt{3.95} = 1.99$

Alternative Method

From the last column of the above table, we have

Sum of squares = 15196

Mean of squares = 15196/76 = 199.95

Thus, σ^2 = Mean of squares – Square of the mean = $199.96 - (14)^2 = 3.96$

Short-cut Method

Before discussing this method, we shall examine an important property of the variance (or standard deviation), given below:

The variance of a distribution is independent of the change of origin but not of change of scale.

Change of Origin

If from each of the observations, $X_1, X_2 \ldots\ldots X_n$, a fixed number, say A, is subtracted, the resulting values are $X_1 - A, X_2 - A \ldots\ldots X_n - A$.

We denote $X_i - A$ by d_i, where $i = 1, 2 \ldots\ldots n$ the values $d_1, d_2 \ldots\ldots d_n$ are said to be measured from A. In order to understand this, we consider the following figure.

X_i Values	:	0	1	2	3	4	5	6	7	8
d_i (= X_i – 3) Values	:	–3	–2	–1	0	1	2	3	4	5

In the above, the origin of X_i values is the point at which $X_i = 0$. When we make the transformation $d_i = X_i - 3$, the origin of d_i values shift at the value 3 because $d_i = 0$ when $X_i = 3$.

The first part of the property says that the variance of X_i values is equal to the variance of the d_i values, i.e.,

$$\sigma_X^2 = \sigma_d^2$$

Change of Scale

To make change of scale every observation is divided (or multiplied) by a suitable constant. For example, if X_i denotes inches, then $Y_i = X_i /12$ will denote feet or if X_i denotes rupees, then $Y_i = 100$

$X_i = X_i/0.01$ will denote paise, etc.

We can also have simultaneous change of origin and scale, by making the transformation $u_i = X_i - A/h$, where A refers to change of origin and h refers to change of scale.

According to second part of the property

$$\sigma_X^2 \neq \sigma_Y^2 \text{ or } \sigma_X^2 \neq \sigma_u^2$$

The relation between σ_X^2 and σ_u^2

Consider

$$\sigma_X^2 = \frac{1}{N}\sum f_i(X_i - \bar{X})^2 \quad ..(1)$$

Let $u_i = X_i - A/h, \quad \therefore \quad X_i = A + hu_i$ ..(2)

Also,

$$\Sigma f_i X_i = \Sigma f_i(A + hu_i) = AN + h\Sigma f_i u_i$$

Dividing both sides by N, we have

$$\frac{\sum f_i X_i}{N} = A + h \cdot \frac{\sum f_i u_i}{N} \text{ or } \bar{X} = A + h\bar{u} \quad ...(3)$$

Substituting the values of X_i and $\bar{X}$ in equation (1), we have

$$\sigma_X^2 = \frac{1}{N}\sum f_i(A + hu_i - A - h\bar{u})^2 = h^2\left[\frac{1}{N}\sum f_i(u_i - \bar{u})^2\right] = h^2\sigma_u^2 \quad ...(4)$$

The result shows that variance is independent of change of origin but not of change of scale. Using this, we can write down a short-cut formula for variance of X.

$$\sigma_X^2 = h^2\left[\frac{\sum f_i u_i^2}{N} - \left(\frac{\sum f_i u_i}{N}\right)^2\right] \quad ...(5)$$

Further, when only change of origin is made

$$\sigma_X^2 = \left[\frac{\sum f_i d_i^2}{N} - \left(\frac{\sum f_i d_i}{N}\right)^2\right], \text{ where } d_i = X_i - A$$

Example: Calculate standard deviation of the following series:

Weekly wages	No. of workers	Weekly wages	No. of workers
100-105	200	130-135	410
105-110	210	135-140	320
110-115	230	140-145	280
115-120	320	145-150	210
120-125	350	150-155	160
125-130	520	155-160	90

Solution:

Calculation of S.D. by using d_i (= X_i – A)

Class Intervals	No. of Workers (f)	Mid-values (X)	d = X – 127.5	fd	fd^2
100-105	200	102.5	–25	–5000	125000
105-110	210	107.5	–20	–4200	84000
110-115	230	112.5	–15	–3450	51750
115-120	320	117.5	–10	–3200	32000
120-125	350	122.5	–5	–1750	8750
125-130	520	127.5	0	0	0

Contd...

130-135	410	132.5	5	2050	10250
135-140	320	137.5	10	3200	32000
140-145	280	142.5	15	4200	63000
145-150	210	147.5	20	4200	84000
150-155	160	152.5	25	4000	100000
155-160	90	157.5	30	2700	81000
Total	**3300**			**2750**	**671750**

$$\sigma_X^2 = \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2 = \frac{671750}{3300} - \left(\frac{2750}{3300}\right)^2 = 202.87$$

$$\therefore \quad \sigma_X^2 = \sqrt{202.87} = ₹14.24$$

5.8.2 Coefficient of Variation

The standard deviation is an absolute measure of dispersion and is expressed in the same units as the units of variable X. A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation and is given by $\sigma/\bar{X} \times 100$.

This measure introduced by Karl Pearson, is used to compare the variability or homogeneity or stability or uniformity or consistency of two or more sets of data. The data having a higher value of the coefficient of variation is said to be more dispersed or less uniform, etc.

Example: Calculate standard deviation and its coefficient of variation from the following data:

Measurements : 0-5 5-10 10-15 15-20 20-25

Frequency : 4 1 10 3 2

Solution:

Calculation of $\bar{X}$ and σ

Class Intervals	Frequency (f)	Mid-values (X)	u	fu	fu^2
0-5	4	2.5	–2	–8	16
5-10	1	7.5	–1	–1	1
10-15	10	12.5	0	0	0
15-20	3	17.5	1	3	3
20-25	2	22.5	2	4	8
Total	**20**			**–2**	**28**

Here, $u = X - 12.5 / 5$

Now, $\bar{X} = 12.5 - (5 \times 2)/20 = 12$ and

$$\sigma = 5\sqrt{\frac{28}{20} - \left(\frac{2}{20}\right)^2} = 5.89$$

Thus, the coefficient of variation (CV) = $5.89 / 12 \times 100 = 49\%$

Example: The mean and standard deviation of 200 items are found to be 60 and 20 respectively. If at the time of calculations, two items were wrongly recorded as 3 and 67 instead of 13 and 17, find the correct mean and standard deviation. What is the correct value of the coefficient of variation?

Solution:

It is given that $\bar{X} = 60$, $\sigma = 20$ and $n = 200$

The sum of observations $\Sigma X_i = n\bar{X} = 200 \times 60 = 12{,}000$

To find the sum of squares of observations, we use the relation

$$\sum X_i^2 = n(\sigma^2 + \bar{X}^2)$$

From this, we can write

$$\sum X_i^2 = 200(400 + 3600) = 8{,}00{,}000$$

Further the corrected sum of observations (ΣX_i) = Uncorrected sum of observations – Sum of wrongly recorded observations + Sum of correct observations = 12,000 – (3 + 67) + (13 +17) = 11,960.

Corrected $\bar{X} = 11960/200 = 59.8$

Similarly, the corrected sum of squares:

(ΣX_i^2) = Uncorrected sum of squares – Sum of squares of wrongly recorded observations + Sum of squares of correct observations

$= 8{,}00{,}000 - (32 + 672) + (132 + 172) = 7{,}95{,}960$

Hence, corrected $\sigma^2 = 795960/200 - (59.8)^2 = 403.76$ or corrected $\sigma = 20.09$

Also, $CV = 20.09/59.8 \times 100 = 33.60$

Example: Find the missing information from the following:

	Group I	**Group II**	**Group III**	**Combined**
Number of observations	50	?	90	200
Standard deviation	6	7	?	7.746
Mean	113	?	115	116

Solution:

Let n_1, n_2, n_3 and n denote the number of observations, X_1, X_2, X_3 and X be the means and σ_1, σ_2, σ_3 and σ be the standard deviations of the first, second, third and combined group respectively.

From the given information, we can easily determine the number of observations in group II,

i.e, $n_2 = n - n_1 - n_3 = 200 - 50 - 90 = 60$.

Further the relation between means is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1 + n_2 + n_3}$$

$$\bar{X}_2 = \frac{(n_1 + n_2 + n_3)\bar{X} - n_1\bar{X}_1 - n_3\bar{X}_3}{n_2} = \frac{200 \times 116 - 50 \times 113 - 90 \times 115}{60} = 120$$

To determine σ_3, consider the following relationship between variances:

$$n\sigma^2 = n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2)$$

or $$\sigma_3^2 = \frac{n\sigma^2 - n_1(\sigma_1^2 + d_1^2) - n_2(\sigma_2^2 + d_2^2)}{n_3} - d_3^2$$

Here $d_1 = 113 - 116 = -3$, $d_2 = 120 - 116 = 4$, $d_3 = 115 - 116 = -1$

$$\therefore \quad \sigma_3^2 = \frac{200(7.746)^2 - 50(36+9) - 60(49+16)}{90} - 1 = 64 \text{ . Thus, } \sigma_3 = 8.$$

5.8.3 Properties of Standard Deviation

1. Standard deviation of a given set of observations is not greater than any other root mean square deviation, i.e.

$$\sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2} \le \sqrt{\frac{1}{n}\sum(X_i - A)^2}$$

2. Standard deviation of a given set of observations is not less than mean deviation from mean, i.e., Standard Deviation ≥ Mean Deviation from mean.

3. In an approximately normal distribution, $X \pm \sigma$ covers about 68% of the distribution, $X \pm 2\sigma$ covers about 95% of the distribution and $X \pm 3\sigma$ covers about 99%, i.e., almost whole of the distribution. This is an Empirical Rule that is based on the observations of several bell shaped symmetrical distributions. This rule is helpful in determining whether the deviation of a particular value from its mean is unusual or not. The deviations of more than 2σ are regarded as unusual and warrant some remedial action. Furthermore, all observations with deviations of more than 3σ from their mean are regarded as not belonging to the given data set.

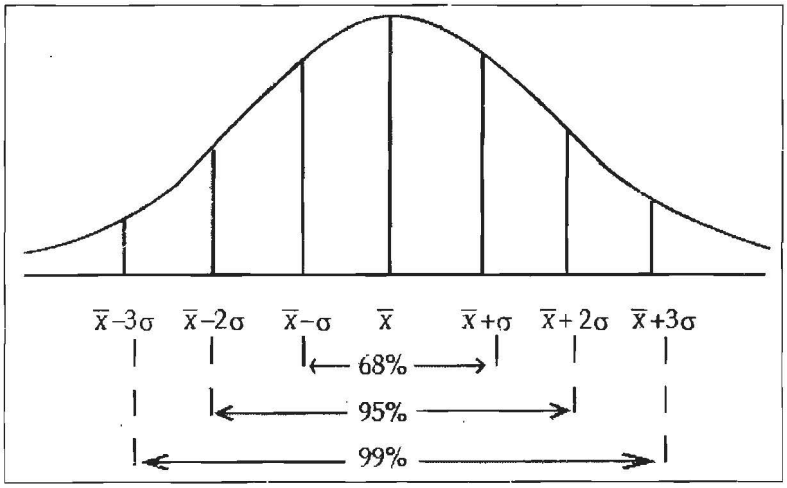


Figure 5.1: Standard Deviation

5.8.4 Merits, Demerits and Uses of Standard Deviation

Merits

- It is a rigidly defined measure of dispersion.
- It is based on all the observations.

- It is capable of being treated mathematically. For example, if standard deviations of a number of groups are known, their combined standard deviation can be computed.
- It is not very much affected by the fluctuations of sampling and, therefore, is widely used in sampling theory and test of significance.

Demerits

- As compared to the quartile deviation and range, etc., it is difficult to understand and difficult to calculate.
- It gives more importance to extreme observations.
- It depends upon the units of measurement of the observations, it cannot be used to compare the dispersions of the distributions expressed in different units.

Uses

- Standard deviation can be used to compare the dispersions of two or more distributions when their units of measurements and arithmetic means are same.
- It is used to test the reliability of mean. It may be pointed out here that the mean of a distribution with lower standard deviation is said to be more reliable.

5.8.5 Empirical Relation among Various Measures of Dispersions

Although much depends upon the nature of a frequency distribution, it has been observed that for a symmetrical or moderately skewed distribution, the following approximate results hold true.

QD ≈ 0.8453 (approximately 5/6) × MD

QD ≈ 0.6745 (approximately 2/3) × MD

QD ≈ 0.7979 (approximately 4/5) × MD

or we can say that 6 SD ≈ QD ≈ 7.5 MD

Check Your Progress

Fill in the blanks:

1. _____ is related to the extent of scatter or variability in observations.
2. A percentile range corresponding to _____.
3. Half of the interquartile range is called the _____ range.
4. _____ is a measure of dispersion based on all the observations.
5. The standard deviation is denoted by _____.
6. The deviations of more than _____ are regarded as unusual and warrant some remedial action.

5.9 LET US SUM UP

- Dispersion in statistics is a way of describing how spread out a set of data is.
- When a data set has a large dispersion, the values in the set are widely scattered; when dispersion is small the items in the set are tightly clustered.
- Range = L – S, L = largest observation and S = smallest observation.

- Coefficient of Range = L – S / L + S
- Quartile Deviation or Semi-Interquartile Range QD = $Q_3 - Q_1/2$
- Coefficient of QD = $Q_3 - Q_1 / Q_3 + Q_1$
- Standard deviation: probably the most common measure of dispersion. It tells you how spread out numbers are from the mean.
- The spread of a data set can be described by a range of descriptive statistics including variance, standard deviation and interquartile range.
- Standard deviation is the square root of the variance.
- The important advantage of interquartile range is that it can be used as a measure of variability if the extreme values are not being recorded exactly.

5.10 UNIT END ACTIVITY

"Frequency distribution may either differ in numerical size of their averages though not necessarily in their formation or they may have the same values of their averages yet differ in their respective formation". Explain and illustrate how the measures of dispersion afford a supplement to the information about frequency distribution furnished by averages.

5.11 KEYWORDS

Averages of Second Order: The measures which express the spread of observations in terms of the average of deviations of observations from some central value are termed as the averages of second order, e.g., mean deviation, standard deviation, etc.

Coefficient of Standard Deviation: A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation.

Dispersion: Dispersion is the measure of extent to which individual items vary.

Distance Measures: The measures which express the spread of observations in terms of distance between the values of selected observations. These are also termed as distance measures, e.g., range, interquartile range, interpercentile range, etc.

Interquartile Range: Interquartile Range is an absolute measure of dispersion given by the difference between third quartile (Q_3) and first quartile (Q_1) Symbolically, Interquartile range = $Q_3 - Q_1$.

Measure of Central Tendency: A measure of central tendency summarizes the distribution of a variable into a single figure which can be regarded as its representative.

Measure of Variation: The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation.

Quartile Deviation or Semi-interquartile Range: Half of the interquartile range is called the quartile deviation or semi-interquartile range.

Range: The range of a distribution is the difference between its two extreme observations, i.e., the difference between the largest and smallest observations. Symbolically, R = L – S where R denotes range, L and S denote largest and smallest observations.

Standard Deviation or Root-mean Square Deviation: The squares of the deviations from arithmetic mean are taken and the positive square root of the arithmetic mean of sum of squares of these deviations is taken as a measure of dispersion. This measure of dispersion is known as standard deviation or root-mean square deviation.

5.12 QUESTIONS FOR DISCUSSION

1. Explain briefly the meaning of (i) Range and (ii) Quartile Deviation.
2. Distinguish between an absolute measure and relative measure of dispersion. What are the advantages of using the latter?
3. What do you understand by mean deviation? Explain its merits and demerits.
4. Explain mean deviation, quartile deviation and standard deviation. Discuss the circumstances in which they may be used.
5. What do you understand by coefficient of variation?
6. Find out quartile deviation and its coefficient from the following data:

Class	0-4	5-9	10-14	15-19	20-24	25-29
Frequency	15	26	12	5	4	3

7. Find out the range of income of (a) middle 50% of workers, (b) middle 80% of the workers and hence the coefficients of quartile deviation and percentile deviation from the following data :

Wages less than	40	50	60	70	80	90	100
No. of workers	5	8	15	20	30	33	35

8. The following data denote the weights of 9 students of certain class. Calculate mean deviation from median and its coefficient.

S. No.	1	2	3	4	5	6	7	8	9
Weight	40	42	45	47	50	51	54	55	57

9. Calculate mean deviation from median for the following data:

Wages per week	50-59	60-69	70-79	80-89	90-99	100-109	110-119
No. of workers	15	40	50	60	45	90	15

10. Calculate the coefficient of mean deviation from mean and median from the following data:

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Students	2	6	12	18	25	20	10	7

11. Calculate the standard deviation of the following series:

Marks	0-10	10-20	20-30	30-40	40-50
Frequency	10	8	15	8	4

12. Calculate the standard deviation from the following data:

Age less than (in years)	10	20	30	40	50	60	70	80
No. of Persons	15	30	53	75	100	110	115	125

13. "Measures of dispersion and central tendency are complementary to each other in highlighting the characteristics of a frequency distribution". Explain this statement with suitable examples.

14. Discuss the relative advantages of coefficient of variation and standard deviation as a measure of variability.

15. Calculate range and its coefficient from the following data:

    (a) 159, 167, 139, 119, 117, 168, 133, 135, 147, 160

    (b)

Weights (lbs)	115-125	125-135	135-145	145-155	155-156	165-175
Frequency	4	5	6	3	1	1

16. The mean of 150 observations is 35 and their standard deviation is 4. Find sum and sum of squares of all the observations.

17. The mean and standard deviation of two distributions having 100 and 150 observations are 50, 5 and 40, 6 respectively. Find the mean and standard deviation of all the 250 observations taken together.

18. The mean and standard deviation of 100 items are found to be 40 and 10. If, at the time of calculations, two items were wrongly taken as 30 and 70 instead of 3 and 27, find the correct mean and standard deviation.

19. The sum and the sum of squares of a set of observations are 75 and 435 respectively. Find the number of observations if their standard deviation is 2.

20. The sum and the sum of squares of 50 observations from the value 20 are – 10 and 452 respectively. Find standard deviation and coefficient of variation.

21. The mean and standard deviation of marks obtained by 40 students of a class in statistics are 55 and 8 respectively. If there are only 5 girls in the class and their respective marks are 40, 55, 63, 75 and 87, find mean and standard deviation of the marks obtained by boys.

22. There are 60 male and 40 female workers in a factory. The standard deviations of their wages (per hour) were calculated as ₹ 8 and ₹ 11 respectively. The mean wages of the two groups were found to be equal. Compute the combined standard deviation of the wages of all the workers.

Check Your Progress: Model Answer

1. Dispersion
2. i = 10
3. Quartile deviation or Semi-interquartile
4. Mean deviation
5. σ
6. 2σ

5.13 REFERENCE & SUGGESTED READINGS

- Camm, J. D., Cochran, J. J., Fry, M. J., Ohlmann, J. W., & Anderson, D. R. (2019). **Essentials of Business Analytics** (2nd ed.). Cengage Learning. ISBN: 9781337406420
- Sharpe, N. R., De Veaux, R. D., & Velleman, P. F. (2019). **Business Statistics** (3rd ed.). Pearson. ISBN: 9780134684773
- Groebner, D. F., Shannon, P. W., & Fry, P. C. (2020). **Business Statistics: A Decision-Making Approach** (10th ed.). Pearson. ISBN: 9780134496499

UNIT - VI

MOMENTS, SKEWNESS AND KURTOSIS

CONTENTS

6.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Understand the concept of moments
- Discuss the meaning of skewness
- Describe the measure of kurtosis

6.1 INTRODUCTION

So far we have discussed the measures of central tendency and dispersion of frequency distributions for their summarisation and comparison with each other. These measures, however, do not adequately describe a frequency distribution in the sense that there could be two or more distributions with same mean and standard deviation but still different from each other with regard to shape or pattern of

distribution of observations. This implies that there is need to develop some more measures to further describe the characteristics of a distribution. These measures are known as moments, skewness and kurtosis.

6.2 MOMENTS

For a frequency distribution having observations X_1, X_2 ..... X_n with respective frequencies as f_1, f_2 ..... f_n, the r^{th} moment about mean X , is defined as:

$$\mu_r = \frac{1}{N}\sum f_i(X_i - \bar{X})^r, \text{ where } N = \Sigma f_i$$

It should be noted here that μ_r is the mean of r^{th} power of deviations of observations from their mean.

In particular, if r = 0, we have

$$\mu_0 = \frac{I}{N}\sum f_i(X_i - \bar{X})^0 = 1$$

If r = 1, we have

$$\mu_1 = \frac{1}{N}\sum f_i(X_i - \bar{X}) = 0$$

If r = 2, we have

$$\mu_2 = \frac{1}{N}\sum f_i(X_i - \bar{X})^2 = \sigma^2$$

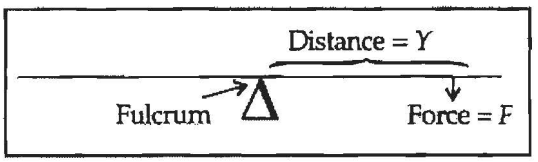
This implies that first moment of a distribution about mean is zero and second moment about mean is equal to variance.

Similarly, if r = 3, we have

$$\mu_3 = \frac{1}{N}\sum f_i(X_i - \bar{X})^3$$

The moments μ_0, μ_1, μ_2.....μ_r, etc. are also known as 'central moments'.

The term 'moment' has been adopted from physics. In physics, this term is used as a measure of force with reference to a point of support commonly known as fulcrum. The moment M of a force, applied at a distance Y from the fulcrum is given by M = F.Y, which is shown in the figure.



In statistics, the deviations are analogous to distance and frequencies are analogous to force. The value from which deviations are taken can be looked upon as fulcrum.

6.2.1 Moments about any Arbitrary Value A

Since deviations of the values can be taken from any arbitrary value A, corresponding moments about A can also be defined.

The r^{th} moment about A, denoted as μ'_r, is defined as:

$$\mu_r' = \frac{1}{N}\sum f_i(X_i - A)^r \text{ , i.e., } \mu_r' \text{ is the mean of } (X_i - A)^r \text{ values.}$$

If r = 0, we have

$$\mu_0' = \frac{1}{N}\sum f_i(X_i - A)^0 = 1$$

If r = 1, we have

$$\mu_1' = \frac{1}{N}\sum f_i(X_i - A)^1 = \frac{1}{N}\sum f_i X_i - A\frac{\sum f_i}{N} \text{ or } \mu_1' = \bar{X} - A.$$

The first moment about A is equal to the difference of $\bar{X}$ and A. Similarly, we have

$$\mu_2' = \frac{1}{N}\sum f_i(X_i - A)^2 \text{ when } r = 2,$$

$$\mu_3' = \frac{1}{N}\sum f_i(X_i - A)^3 \text{ when } r = 3, \text{etc.}$$

The moments about any arbitrary value are also known as 'raw moments'.

6.2.2 Moments about Origin

The r^{th} moment about origin, denoted as m_r, is defined as:

$$m_r = \frac{1}{N}\sum f_i(X_i - 0)^r = \frac{1}{N}\sum f_i X_i^r \text{ , i.e, } m_r \text{ is mean of } X_i^r \text{ values.}$$

Note: $m_r = \mu'_r$ if A is taken equal to zero.

If r = 1, we have

$$m_1 = \frac{1}{N}\sum f_i X_i = \bar{X}$$

Thus, mean of a distribution is also known as the first moment about origin. Further, when r = 2, 3, ... etc., we have

$$m_2 = \frac{1}{N}\sum f_i X_i^2, \; m_3 = \frac{1}{N}\sum f_i X_i^3, \ldots \text{ etc.}$$

6.2.3 Relation between Central Moments and Raw Moments

Given central moments, we can always find moments about any arbitrary value A and vice versa. First we shall obtain central moments from moments about A, i.e., given the values of μ'_1, μ'_2, μ'_3, μ'_4....., we have to find the values of m_2, m_3, m_4 ..... (Remember that m_1 is always zero).

Consider

$$\mu_r = \frac{1}{N}\sum f_i(X_i - \bar{X})^r$$

On adding and subtracting A to $(X_i - \bar{X})$, we can write

$$\mu_r = \frac{1}{N}\sum f_i[(X_i - A) - (\bar{X} - A)]^r$$

Expanding the right hand side by binomial theorem, we have

$$\mu_r = \frac{1}{N}\sum f_i[(X_i - A)^r - {}^rC_1(X_i - A)^{r-1}(\bar{X} - A) + {}^rC_2(X_i - A)^{r-2}(\bar{X} - A)^2 - {}^rC_3(X_i - A)^{r-3}(\bar{X} - A)^3 + \ldots]$$

$$= \frac{1}{N}\sum f_i(X_i - A)^r - {}^rC_1\left[\frac{1}{N}\sum f_i(X_i - A)^{r-1}\right]\mu_1' + {}^rC_2\left[\frac{1}{N}\sum f_i(X_i - A)^{r-2}\right]\mu_1'^2$$

$$- {}^rC_3\left[\frac{1}{N}\sum f_i(X_i - A)^{r-3}\right]\mu_1'^3 + \ldots$$

$(\mu'_1 = \bar{X} - A)$

$$\mu_r' - {}^rC_1\,\mu'_{r-1}\,\mu_1' + {}^rC_2\,\mu'_{r-2}\,\mu_1'^2 - {}^rC_3\,\mu'_{r-3}\,\mu_1'^3 + \ldots$$

In particular, when r = 1, we have $\mu_1 = \mu'_1 - \mu'_1 = 0$ ...(1)

When r = 2, we have,

$$\mu_2 = \mu_2' - {}^2C_1\,\mu_1'\mu_1' + {}^2C_2\,\mu_0'\,\mu_1'^2 = \mu_2' - 2\mu_1'^2 + \mu_1'^2 = \mu'_2 - \mu_1'^2 \quad \ldots(2)$$

When r = 3, we have

$$\mu_3 = \mu_3' - {}^3C_1\,\mu_2'\mu_1' + {}^3C_2\,\mu_1'\mu_1'^2 - {}^3C_3\mu_0'\,\mu_1'^3 = \mu'_3 - 3\mu_2'\mu_1' + 3\mu_1'^3 - \mu_1'^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \quad \ldots(3)$$

When r = 4, we have

$$\mu_4 = \mu_4' - {}^4C_1\mu_3'\,\mu'_1 + {}^4C_2\mu_2'\mu_1'^2 - {}^4C_3\,\mu_1'^4 + {}^4C_4\,\mu_1'^4$$

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 \quad \ldots(4)$$

In a similar way, we can express other higher ordered central moments in terms of raw moments. Using equations (2), (3) and (4), we can also express raw moments in terms of central moments as shown below:

From equation (2), we can express μ'_2 in terms of μ'_2 and μ'_1 as

$$\mu_2' = \mu_2 + \mu_1'^2 \quad \ldots(5)$$

On substituting, this value of μ'_2 in equation (3), we can get

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 \quad \ldots(6)$$

Similarly, we can obtain μ'_4 from equation (4) as

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4 \quad \ldots(7)$$

Proceeding in this way, we can also obtain the moments of higher orders.

Remarks:

(i) $\mu'_1 (= \bar{X} - A)$ is used in both types of expressions.

(ii) Since $\mu_1'^2$ is a non-negative number, $\mu_2 = \mu'_2 - \mu_1'^2$ implies that $\mu_2 \leq \mu'_2$ i.e., Variance ≤ Mean square deviation or S.D. ≤ root mean square deviation.

6.2.4 Relation between Central Moments and Moments about Origin

Here we have

$$m_1 = \frac{\bar{X}}{\bar{Y}} - 0 = \bar{X}$$

The expressions for moments of various orders can be written from the expressions given above just by replacing μ'_i by m_i. Thus, the expression for second, third and fourth central moments in terms of moments about origin are

$$\mu_2 = m_2 - m_1^2$$

$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3$$

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4$$

Similarly, the expressions for second, third and fourth moments about origin in terms of central moments are

$$m_2 = \mu_2 + m_1^2$$

$$m_3 = \mu_3 + 3\mu_2m_1 + m_1^3$$

$$m_4 = \mu_4 + 4\mu_3m_1 + 6\mu_2m_1^2 + m_1^4$$

Example: Calculate the first four moments about 30 for the following distribution and convert them into central moments.

Class Intervals	:	5-15	15-25	25-35	35-45	45-55
Frequency	:	8	12	15	9	6

Solution:

Calculation of Moments

Class Intervals	Freq. (f)	M. V. (X)	X – 30	f(X – 30)	f(X – 30)^2	f(X – 30)^3	f(X – 30)^4
5-15	8	10	–20	–160	3200	–64000	1280000
15-25	12	20	–10	–120	1200	–12000	120000
25-35	15	30	0	0	0	0	0
35-45	9	40	10	90	900	9000	90000
45-55	6	50	20	120	2400	48000	960000
Total	**50**			**–70**	**7700**	**–19000**	**2450000**

$$\mu_1' = \frac{-70}{50} = -1.40,\ \mu_2' = \frac{7700}{50} = 154,\ \mu_3' = \frac{-19000}{50} = -380,\ \mu_4' = \frac{2450000}{50} = 49000$$

Conversion into central moments

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 154 - (-1.4)^2 = 152.04$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -380 - 3 \times 154\,(-1.4) + 2 \times (-1.4)^3 = 261.31$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 49000 - 4 \times (-380)(-1.4) + 6 \times 154 \times (1.4)^2 - 3 \times (-1.4)^4 = 48671.52$$

Example: The first two moments of a distribution about the value 5 are 2 and 20. Find mean and variance of the distribution.

Solution:

We know that

$$\mu'_1 = \bar{X} - A$$

or

$$\bar{X} = \mu'_1 + A = 2 + 5 = 7$$

Also, $\mu_2 = \mu'_2 - \mu'^2_2 = 20 - 4 = 16$

Mean = 7 and Variance = 16

Example: The first four moments of a distribution about 4 are as given below:

$$\mu_1' = 1,\ \mu_2' = 4,\ \mu_3' = 10 \text{ and } \mu_4' = 45$$

Find mean of the distribution and calculate the first four moments about mean and also the first four moments about origin.

Solution:

We know that

$$m_1 = 0 \text{ and } \mu_2 = \mu_2' - \mu_1'^2 = 4 - 1 = 3$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 10 - 12 + 2 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 45 - 4 \times 10 + 6 \times 4 - 3 = 26$$

Moments about origin.

$$m_1 = \bar{X} = \mu_1' + A = 1 + 4 = 5$$

$$m_2 = m_2 + m_1^2 = 3 + 25 = 28$$

$$m_3 = m_3 + 3m_2m_1 + m_1^3 = 0 + 45 + 125 = 170$$

$$m_4 = m_4 + 4m_3m_1 + 6m_2m_1^2 + m_1^4 = 26 + 0 + 18 \times 25 + 625 = 1101$$

Example: The first four moments from mean of a distribution are 0, 3.2, 3.6 and 20. The mean value is 11. Calculate the first four moments about zero and about 10.

Solution:

We are given

$$\mu_1 = 0,\ \mu_2 = 3.2,\ \mu_3 = 3.6,\ \mu_4 = 20 \text{ and } \bar{X} = 11$$

The first four moments about origin are:

$$m_1 = 11$$

$$m_2 = \mu_2 + m_1^2 = 3.2 + 121 = 124.2$$

$$m_3 = m_3 + 3\mu_2 m_1 + m_1^3 = 3.6 + 3 \times 3.2 \times 11 + 11^3 = 1440.2$$

$$m_4 = m_4 + 4\mu_3 m_1 + 6\mu_2 m_1^2 + m_1^4$$

$$= 20 + 4 \times 3.6 \times 11 + 6 \times 3.2 \times 121 = 11^4 = 17142.6$$

The first four moments about 10 are

$$\mu_1' = \bar{X} - A = 11 - 10 = 1,$$

$$\mu_2' = \mu_2 + \mu_1'^2 = 3.2 + 1 = 4.2$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 = 3.6 + 3 \times 3.2 + 1 = 14.2$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4 = 20 + 4 \times 3.6 + 6 \times 3.2 + 1 = 54.6$$

6.2.5 Effect of Change of Scale and Origin on Moments

Let

$$u_i = \frac{X_i - A}{h}$$

where A refers to change of origin and h refers to change of scale. From the above, we can write $X_i = A + hu_i$ and

$$\bar{X} = A + h\bar{u}$$

The r^{th} moment about mean can be written as:

$$\mu_r(X) = \frac{1}{N}\sum f_i(X_i - \bar{X})^r = \frac{1}{N}\sum f_i[A + hu_i - A - h\bar{u}]^r$$

$$= h^r \cdot \frac{1}{N}\sum f_i(u_i - \bar{u})^r = h^r \mu_r(u)$$

This result shows that r^{th} moment of X-values about mean is h^r times the r^{th} moment of u-values about its mean. This result also shows that central moments are independent of change of origin but not of change of scale.

Further, we can write

$$X_i - A = hu_i$$

$$\mu_r'(X) = \frac{1}{N}\sum f_i(X_i - A)^r = h^r m_r(u)$$

6.2.6 Charlier's Check of Accuracy

Charlier has given the following identities which can be used to check the accuracy of calculations of moments from a frequency distribution:

Moment	Chariler's Check Formula
First	$\sum f(X+1) = \sum fX + Nd$
Second	$\sum f(X+1)^2 = \sum fX^2 + 2\sum fX + N$
Third	$\sum f(X+1)^3 = \sum fX^3 + 3\sum fX^2 + 3\sum fX^2 + N$
Fourth	$\sum f(X+1)^4 = \sum fX^4 + 4\sum fX^3 + 6\sum fX^2 + 4\sum fX + N$

6.2.7 Sheppard's Correction for Grouping

In case of a grouped or continuous frequency distribution, the calculation of moments is based upon the assumption that all observations in a class are equal to its middle value. This assumption leads to a systematic error in the even ordered moments. The following corrections are suggested by W.F. Sheppard.

Since $\mu_1 = 0$, therefore, there is no need of correction.

Corrected $\mu_2 = \mu_2 - \frac{h^2}{12}$ where h is class interval.

Third moment, μ_3 needs no correction.

Corrected $\mu_4 = \mu_4 - \frac{\mu_2 h^2}{2} + \frac{7h^4}{240}$, etc.

Where μ_2, μ_4 appearing on the right hand side of the above equations are uncorrected values.

Note: These corrections are valid for bell shaped distributions with flat tails and are not applicable to J-shaped or U-shaped distributions.

6.2.8 Coefficients Based on Moments

Alpha Coefficients

The moments, discussed so far, have their units depending upon the units of variable X. For example, if X is measured in inches, then the respective units of μ_1, μ_2, μ_3 ... are inches, inches^2, inches^3 ... etc. Thus, in order that the moments of two or more distributions are comparable, it is necessary to convert them into coefficients.

Transform the variable X_i into another variable

$$z_i = \frac{X_i - \bar{X}}{\sigma}$$

where z_i is a variable with mean zero and standard deviation unity. This variable is independent of the units of measurements and hence, its moments will also be pure numbers independent of units. Various moments of z about zero are called α-coefficients. The r^{th} order moment of z about zero, denoted by α_r is given by

$$\alpha_r = \frac{1}{N}\sum f_i z_i^r = \frac{1}{N}\sum f_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^r$$

On taking r = 1, 2, ... etc., various α-coefficients can be written as

$$\alpha_1 = \frac{1}{N}\sum f_i z_i = \frac{1}{N}\sum f_i \left(\frac{X_i - \bar{X}}{\sigma}\right) = \frac{1}{\sigma}\cdot\frac{\sum f_i (X_i - \bar{X})}{N} = \frac{\mu_1}{\sigma} = 0$$

$$\alpha_2 = \frac{1}{N}\sum f_i z_i^2 = \frac{1}{N}\sum f_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{1}{\sigma^2}\cdot\frac{\sum f_i (X_i - \bar{X})^2}{N} = \frac{\mu_2}{\sigma^2} = 1$$

$$\alpha_3 = \frac{1}{N}\sum f_i z_i^3 = \frac{1}{N}\sum f_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^3 = \frac{1}{\sigma^3}\cdot\frac{\sum f_i (X_i - \bar{X})^3}{N} = \frac{\mu_3}{\sigma^3}$$

$$\alpha_4 = \frac{1}{N}\sum f_i z_i^4 = \frac{1}{N}\sum f_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^4 = \frac{1}{\sigma^4}\cdot\frac{\sum f_i (X_i - \bar{X})^4}{N} = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$

Beta Coefficients

Karl Pearson suggested two Beta coefficients, β_1 and β_2, that are related to β-coefficients:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \alpha_3^2, \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \alpha_4$$

Gamma Coefficients

The gamma coefficients, denoted as γ_1 and γ_2, were suggested by R.A. Fisher and are related to α and β coefficients as given below:

$\gamma_1 = \pm\sqrt{\beta_1} = \alpha_3$, where the sign of $\sqrt{\beta_1}$ is taken as the sign of α_3.

$\gamma_2 = \beta_2 - 3 = \alpha_4 - 3$

6.3 SKEWNESS

Skewness of a distribution refers to its asymmetry. The symmetry of a distribution implies that for a given deviation from a central value, there is equal number of observations on either side of it. If the distribution is asymmetrical or skewed, its frequency curve would have a prolonged tail either towards its left or towards its right hand side. Thus, the skewness of a distribution is defined as the departure from symmetry. We may note here that there may be a situation where two or more frequency distributions are same with regard to mean and variance but not so with regard to skewness.

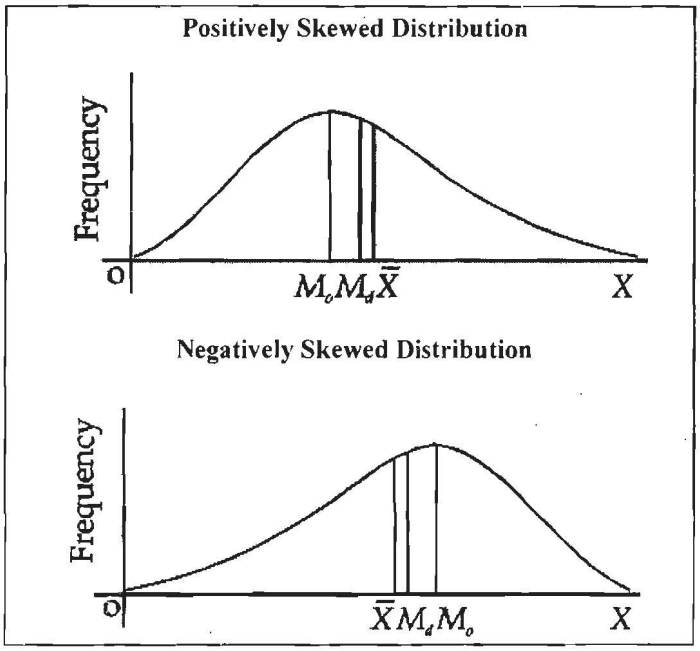


Figure 6.1: Positively vs. Negatively Skewed Distribution

In a symmetrical distribution, mean, median and mode are equal and the ordinate at mean divides the frequency curve into two parts such that one part is the mirror image of the other, positive skewness results if some observations of high magnitude are added to a symmetrical distribution so that the right hand tail of the frequency curve gets elongated. In such a situation, we have Mode < Median < Mean. Similarly, negative skewness results when some observations of low magnitude are added to the

distribution so that left hand tail of the frequency curve gets elongated and we have Mode > Median > Mean.

6.3.1 Measures of Skewness

A measure of skewness gives the extent and direction of skewness of a distribution. As in case of dispersion, we can define the absolute and the relative measures of skewness. Various measures of skewness can be divided into three broad categories: Measures of Skewness based on

- X, M_d and M_o
- Quartiles or percentiles
- Moments

Measure of Skewness based on X, M_d and M_o

This measure was suggested by Karl Pearson. According to this method, the difference between $\bar{X}$ and M_o can be taken as an absolute measure of skewness in a distribution, i.e., absolute measure of skewness = $\bar{X} - M_o$.

Alternatively, when mode is ill defined and the distribution is moderately skewed, the above measure can also be approximately expressed as $3(\bar{X} - M_d)$.

A relative measure, known as Karl Pearson's Coefficient of Skewness, is given by

$$S_K = \frac{\bar{X} - M_o}{\sigma} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} \text{ or } S_K = \frac{3(\bar{X} - M_d)}{\sigma}$$

We note that:

if $S_k > 0$, the distribution is positively skewed,

if $S_k < 0$, the distribution is negatively skewed and

if $S_k = 0$, the distribution is symmetrical.

Measure of Skewness based on Quartiles or Percentiles

(a) ***Using Quartiles:*** This measure, suggested by Bowley, is based upon the fact that Q_1 and Q_3 are equidistant from median of a symmetrical distribution, i.e., $Q_3 - M_d = M_d - Q_1$.

Therefore, $(Q_3 - M_d) - (M_d - Q_1)$ can be taken as an absolute measure of skewness. A relative measure, known as Bowley's Coefficient of Skewness, is defined as

$$S_Q = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 - 2M_d + Q_1}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

The value of S_Q will lie between – 1 and + 1.

It may be noted here that S_k and S_Q are not comparable, though, in the absence of skewness, both of them are equal to zero.

(b) ***Using Percentiles:*** Bowley's measure of skewness leaves 25% observations on each extreme of the distribution and hence is based only on the middle 50% of the observations. As an improvement to this, Kelly suggested a measure based on the middle 80% of the observations.

Kelly's absolute measure of Skewness = $(P_{90} - P_{50}) - (P_{50} - P_{10})$ and

Kelly's Coefficient of Skewness $S_p = \dfrac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})} = \dfrac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$

(We note that $P_{50} = M_d$)

Measure of Skewness based on Moments

This measure is based on the property that all odd ordered moments of a symmetrical distribution are zero. Therefore, a suitable α-coefficient can be taken as a relative measure of skewness.

Since $\alpha_1 = 0$ and $\alpha_2 = 1$ for every distribution, these do not provide any information about the nature of a distribution. The third α-coefficient, i.e., α_3 can be taken as a measure of the coefficient of skewness. The skewness will be positive, negative or zero (i.e. symmetrical distribution) depending upon whether $\alpha_3 > 0$, < 0 or $= 0$. Thus, the coefficient of skewness based on moments is given as

$$S_M = \alpha_3 = \frac{\mu_3}{\sigma_3} = \pm\sqrt{\beta_1} = \gamma_1$$

Alternatively, the skewness is expressed in terms of β_1. Since β_1 is always a non-negative number, the sign of skewness is given by the sign of μ_3.

Example: Calculate the Karl Pearson's coefficient of skewness from the following data:

Size	:	1	2	3	4	5	6	7
Frequency	:	10	18	30	25	12	3	2

Solution:

To calculate Karl Pearson's coefficient of skewness, we first find $\bar{X}$, M_o and σ from the given distribution.

Size (X)	Frequency (f)	d = X – 4	fd	fd^2
1	10	–3	–30	90
2	18	–2	–36	72
3	30	–1	–30	30
4	25	0	0	0
5	12	1	12	12
6	3	2	6	12
7	2	3	6	18
Total	**100**		**–72**	**234**

$$\bar{X} = A + \frac{\sum fd}{N} = 4 + \frac{-72}{100} = 3.28$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{234}{100} - \left(\frac{-72}{100}\right)^2} = 1.35$$

Also, M_o (by inspection) = 3.00

$$S_K = \frac{\bar{X} - M_0}{\sigma} = \frac{3.28 - 3.00}{1.35} = 0.207$$

Since S_k is positive and small, the distribution is moderately positively skewed.

Example: Calculate Karl Pearson's coefficient of skewness from the following data:

Weights (lbs)	No. of Students
90-100	4
100-110	10
110-120	17
120-130	22
130-140	30
140-150	23
150-160	16
160-170	5
170-180	3

Solution:

Calculation of $\bar{X}$, σ and M_0

Class Intervals	Frequency (f)	Mid-points (X)	$u = \frac{X-135}{10}$	fu	fu^2
90-100	4	95	–4	–16	64
100-110	10	105	–3	–30	90
110-120	17	115	–2	–34	68
120-130	22	125	–1	–22	22
130-140	30	135	0	0	0
140-150	23	145	1	23	23
150-160	16	155	2	32	64
160-170	5	165	3	15	45
170-180	3	175	4	12	48
Total	**130**			**–20**	**424**

1. $\bar{X} = A + h\frac{\sum fu}{N} = 135 + 10 \times \frac{-20}{130} = 133.46$

2. $\sigma = h \times \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 10 \times \sqrt{\frac{424}{130} - \left(\frac{-20}{130}\right)^2} = 18.0$

3. $M_0 = L_m + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$

By inspection, the modal class is 130-140.

$\therefore$ $L_m = 130$, $\Delta_1 = 30 - 22 = 8$, $\Delta_2 = 30 - 23 = 7$ and $h = 10$

Thus, $M_o = 130 + \frac{8}{15} \times 10 = 135.33$

Hence, $S_k = \frac{\bar{X} - M_o}{\sigma} = \frac{133.46 - 135.33}{18.0} = -0.10$, *i.e.*, the distribution is moderately negatively skewed.

Example: Calculate Karl Pearson's coefficient of skewness from the following data:

Class Intervals :	40-60	30-40	20-30	15-20	10-15	5-10	3-5	0-3
Frequency :	25	15	12	8	6	4	3	2

Solution:

Since mode is ill defined, skewness will be computed by the use of the median.

Calculation of $\bar{X}$, σ and M_d

Class Intervals	Freq. (f)	M.V. (X)	d = X – 25	fd	fd^2	Less than (c.f.)
0-3	2	1.5	–23.5	–47.0	1104.5	2
3-5	3	4.0	–21.0	–63.0	1323.0	5
5-10	4	7.5	–17.5	–70.0	1225.0	9
10-15	6	12.5	–12.5	–75.0	937.5	15
15-20	8	17.5	–7.5	–60.0	450.0	23
20-30	12	25.0	0.0	0.0	0.0	35
30-40	15	35.0	10.0	150.0	1500.0	50
40-60	25	50.0	25.0	625.0	15625.0	75
Total	**75**			**460.0**	**22165.0**	

1. $\bar{X} = 25 + \frac{460}{75} = 31.13$

2. $\sigma = \sqrt{\frac{22165}{75} - \left(\frac{460}{75}\right)^2} = 16.06$

3. Since $\frac{N}{2} = \frac{75}{2} = 37.5$, median class is 30-40.

Thus,

$L_m = 30, C = 35, f_m = 15, h = 10$

$M_d = 30 + \frac{37.5 - 35}{15} \times 10 = 31.67$

$S_k = \frac{3(\bar{X} - M_d)}{\sigma} = \frac{3(31.13 - 31.67)}{16.06} = -0.10$

Example: Calculate Bowley's coefficient of skewness from the following data:

Class Intervals :	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency :	7	10	20	13	17	10	14	9

Solution:

Calculation of M_d, Q_1 and Q_3

Class Intervals	Frequency (f)	Less than (c.f.)
0-5	7	7
5-10	10	17
10-15	20	37
15-20	13	50
20-25	17	67
25-30	10	77
30-35	14	91
35-40	9	100
Total	**100**	

Since N/2 = 50, the median class is 15-20.

Thus, $L_m = 15$, $f_m = 13$, $C = 37$, $h = 5$, hence

$$M_d = 15 + \frac{50 - 37}{13} \times 5 = 20$$

Since N/4 = 25, the first quartile class is 10-15.

Thus, $L_{Q_1} = 10$, $f_{Q_1} = 20$, $C = 17$, $h = 5$, hence

$$Q_1 = 10 + \frac{25 - 17}{20} \times 5 = 12$$

Since 3N/4 = 75, the third quartile class is 25-30.

Thus, $L_{Q_3} = 25$, $f_{Q_3} = 10$, $C = 67$, $h = 5$, hence

$$Q_3 = 25 + \frac{75 - 67}{10} \times 5 = 29$$

Bowley's Coefficient of Skewness $S_Q = \frac{29 - 2 \times 20 + 12}{29 - 12} = \frac{1}{17} = 0.06$

Thus, the distribution is approximately symmetrical.

Example: In a frequency distribution, the coefficient of skewness based upon quartiles is 0.6. If the sum of upper and lower quartiles is 100 and median is 38, find the values of upper and lower quartiles.

Solution:

It is given that $Q_3 + Q_1 = 100$, $M_d = 38$ and $S_Q = 0.6$

Substituting these values in Bowley's formula, we get

$$0.6 = \frac{100 - 2 \times 38}{Q_3 - Q_1} \Rightarrow Q_3 - Q_1 = 40$$

Adding the equations $Q_3 + Q_1 = 100$ and $Q_3 - Q_1 = 40$, we get

$2Q_3 = 140$ or $Q_3 = 70$

Also $Q_1 = 30$ $(Q_1 + Q_3 = 100)$.

Example: Calculate the Kelly's coefficient of skewness from the following data :

Wages (₹)	No. of Workers
800-900	10
900-1000	33
1000-1100	47
1100-1200	110
1200-1300	160
1300-1400	80
1400-1500	60

Solution:

Calculation of P_{10}, P_{50} and P_{90}

Class Intervals	No. of Workers (f)	Less than (c.f.)
800-900	10	10
900-1000	33	43
1000-1100	47	90
1100-1200	110	200
1200-1300	160	360
1300-1400	80	440
1400-1500	60	500
Total	**500**	

1. Since $\frac{10}{100}N = \frac{10 \times 500}{100} = 50$. P_{10} lies is the interval 1000-1100.

   Thus, $L_{P_{10}} = 1000$, $C = 43$, $f_{P_{10}} = 47$, $h = 100$

   Hence, $P_{10} = 1000 + \frac{50-43}{47} \times 100 = ₹1014.89$

2. Since $\frac{50}{100}N = 250$, P_{50} lies in the interval 1200-1300.

   Thus, $L_{P_{50}} = 1200$, $C = 200$, $f_{P_{50}} = 160$, $h = 100$

   Hence, $P_{50} = 1200 + \frac{250-200}{160} \times 100 = ₹12.31.25$

3. Since $\frac{90}{100}N = 450$, P_{90} lies in the class 1400-1500.

   Thus, $L_{P_{90}} = 1400$, $C = 440$, $f_{P_{90}} = 60$, $h = 100$.

   Hence, $P_{90} = 1400 + \frac{450-440}{60} \times 100 = ₹1416.67$

$$\therefore S_P = \frac{1416.67 + 1014.89 - 2 \times 1231.25}{1416.67 - 1014.89} = \frac{-30.94}{401.78} = -0.08$$

Example: Compute the moment measure of skewness from the following distribution:

Marks obtained	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	8	14	22	26	15	10	5

Solution:

Calculation of Skewness

Class Intervals	Freq. (f)	M.V. (X)	$u = \frac{X-35}{10}$	fu	fu^2	fu^3
0-10	8	5	–3	–24	72	–216
10-20	14	15	–2	–28	56	–112
20-30	22	25	–1	–22	22	–22
30-40	26	35	0	0	0	0
40-50	15	45	1	15	15	15
50-60	10	55	2	20	40	80
60-70	5	65	3	15	45	135
Total	**100**			**–24**	**250**	**–120**

$$\mu_1' = h\frac{\sum fu}{N} = \frac{10 \times (-24)}{100} = -2.4,\ \mu_2' = h^2\frac{\sum fu^2}{N} = \frac{100 \times 250}{100} = 250 \text{ and}$$

$$\mu_3' = h^3\frac{\sum fu^3}{N} = \frac{1000 \times (-120)}{100} = -1200$$

Thus,

$\mu_2 = 250 - (-2.4)^2 = 244.24$ and $\mu_3 = -1200 + 3 \times 250 \times 2.4 + 2(-2.4)^3 = 572.35$

Hence,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(572.35)^2}{(244.24)^3} = 0.02248$$

Since the value of β_1 is small and μ_3 is positive, therefore, the distribution is moderately positively skewed.

Since β_1 is a coefficient, its value can directly be obtained from moments of u, i.e., the moments without adjustment by the scale factor h. Let us denote various moments of u as follows:

$$\delta_1' = \frac{\sum fu}{N} = \frac{-24}{100} = -0.24,\ \delta_2' = \frac{\sum fu^2}{N} = \frac{250}{100} = 2.50,\ \delta_3' = \frac{\sum fu^3}{N} = \frac{-120}{100} = -1.2$$

$$\therefore\quad \delta_2 = \delta_2' - \delta_1'^2 = 2.50 - (-0.24)^2 = 2.4424$$

Note: At least 4 places after decimal should be taken to get the correct results.

$$\delta_3 = \delta_3' - 3\delta_2'\delta_1' + 2\delta_1'^3 = -1.2 - 3 \times 2.5 \times (-0.24) + 2 \times (-0.24)^3 = 0.5724$$

$$\therefore\quad \beta_1 = \frac{\delta_3^2}{\delta_2^3} = \frac{(0.5724)^2}{(2.4424)^3} = 0.02249$$

It may also be pointed out that the central moments can also be obtained from δ_2, δ_3, etc., by suitable multiplication of the scale factor.

Example: If the first three moments of an empirical frequency distribution about the value 2 are 1, 16 and 40. Examine the skewness of the distribution.

Solution:

We are given raw moments; $\mu'_1 = 1$, $\mu'_2 = 16$ and $\mu'_3 = 40$, which should be converted into central moments.

Now

$$\mu_2 = \mu'_2 - \mu'_1 = 16 - 1 = 15$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu' = -40 - 3 \times 16 + 2 = -86$$

$$\gamma_1 = -\sqrt{\beta_1} = -\frac{86}{(15)^{3/2}} = -1.48$$

6.4 KURTOSIS

This is another measure of the shape of a frequency curve. While skewness refers to the extent of lack of symmetry, kurtosis refers to the extent to which a frequency curve is peaked. Kurtosis is a Greek word which means bulginess. In statistics, the word is used for a measure of the degree of peakedness of a frequency curve.

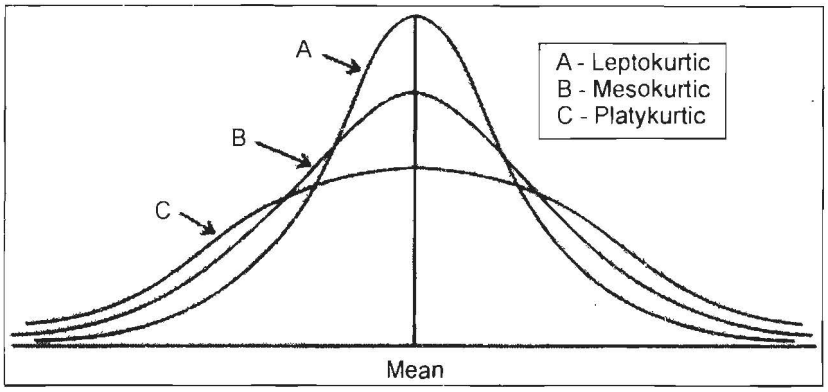


Figure 6.2: Degree of Peakedness of a Frequency Curve

Karl Pearson, in 1905, introduced three types of curves depending upon the shape of their peaks. These three shapes are known as Mesokurtic, Leptokurtic and Platykurtic. A mesokurtic shaped curve is neither too peaked nor too flattened. This in fact is the frequency curve of a normal distribution. A curve that is more peaked than a normal curve is known as leptokurtic while a relatively flat topped curve is known as platykurtic. The three types of curves are shown in the Figure 6.2.

6.4.1 Measures of Kurtosis

A measure of the coefficient of kurtosis, given by Karl Pearson, is

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

This is also equal α_4.

For a normal distribution (i.e., mesokurtic curve) the value of $\beta_2 = 3$. When $\beta_2 > 3$, the curve is more peaked than the normal and is called a leptokurtic curve. Further, when $\beta_2 < 3$, the curve is less peaked than the normal and is called a platykurtic curve.

The measure of kurtosis can also be expressed in terms of γ_2. Since $\gamma_2 = b_2 - 3$, we can write $\gamma_2 = 0$ for mesokurtic, $\gamma_2 > 0$ for leptokurtic and $\gamma_2 < 0$ for platykurtic curve.

Example: Find standard deviation and kurtosis of the following series by the method of moments:

Class Intervals	:	0-10	10-20	20-30	30-40	40-50
Frequency	:	10	20	40	20	10

Solution:

Calculation of Moments

Class Intervals	Frequency (f)	Mid-values (X)	$u = \frac{X-25}{10}$	fu	fu^2	fu^4
0-10	10	5	–2	–20	40	160
10-20	20	15	–1	–20	20	20
20-30	40	25	0	0	0	0
30-40	20	35	1	20	20	20
40-50	10	45	2	20	40	160
Total	**100**			**0**	**120**	**360**

Since $\Sigma fu = 0$, $\bar{X} = 25$ and the calculated moments will be central.

$$\mu_2 = h^2 \frac{\sum fu^2}{N} = 100 \times \frac{120}{100} = 120$$

$$\mu_4 = h^4 \frac{\sum fu^4}{N} = 10000 \times \frac{360}{100} = 36000$$

Thus, measure of kurtosis:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{36000}{14400} = 2.5$$

Since this value is less than 3, the distribution is platykurtic.

The standard deviation:

$\sigma = \sqrt{120} = 10.95$

Example: The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Calculate the moment measures of skewness and kurtosis of the distribution and comment upon the results.

Solution:

The moment measures of skewness and kurtosis are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3$$

Since β_1 is very small, the distribution is approximately symmetrical. Further, $\beta_2 = 3$, therefore, the curve is mesokurtic. The above calculations show that the given distribution is approximately normal.

Example: The following data are given to an economist for the purpose of economic analysis. The data refers to the length of life of a certain type of batteries.

n = 100, $\Sigma fd = 50$, $\Sigma fd^2 = 1970$, $\Sigma fd^3 = 2948$ and $\Sigma fd^4 = 86{,}752$. Here $d = X - 48$.

Do you think that the distribution is platykurtic?

Solution:

We can calculate raw moments, from the given values, as given below

$$\mu_1' = \frac{\sum fd}{N} = \frac{50}{100} = 0.5,\ \mu_2' = \frac{\sum fd^2}{N} = \frac{1970}{100} = 19.7,$$

$$\mu_3' = \frac{\sum fd^3}{N} = \frac{2948}{100} = 29.48,\ \mu_4' = \frac{\sum fd^4}{N} = \frac{86752}{100} = 867.52$$

To calculate β_2, we compute μ_2 and μ_4, as given below

$$\mu_2 = \mu_2' - \mu_1'^2 = 19.7 - 0.5^2 = 19.45$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 867.52 - 4 \times 29.48 \times 0.5 + 6 \times 19.7 \times (0.5)^2 - 3 \times (0.5)^4 = 837.9$$

Now,

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{837.9}{(19.45)^2} = 2.2$$

which is less than 3, therefore, the distribution is platykurtic.

Check Your Progress

Fill in the blanks:

1. μ_r is the mean of _____ power of deviations of observations from their mean.
2. The term 'moment' has been adopted from _____.
3. The gamma coefficients, denoted as γ_1 and γ_2, were suggested by _____.
4. Skewness of a distribution refers to its _____.
5. The skewness is expressed in terms of _____.
6. Kurtosis is a Greek word which means _____.

6.5 LET US SUM UP

- Moments about mean are generally used in statistics. We use a Greek alphabet read as mu for these moments. Consider a mass attached at each point proportional to its frequency and take moments about the mean.
- First, second, third and fourth moments can be used as a measure of Central Tendency, Variation (dispersion), asymmetry and peaked-ness of the curve. We have understood the first four moments about mean in this lesson, i.e., μ_1, μ_2, μ_3, and μ_4.

- Measures of Skewness and Kurtosis, like measures of central tendency and dispersion, study the characteristics of a frequency distribution.
- Averages tell us about the central value of the distribution and measures of dispersion tell us about the concentration of the items around a central value.
- When two or more symmetrical distributions are compared, the difference in them is studied with 'Kurtosis'.
- When two or more symmetrical distributions are compared, they will give different degrees of Skewness. These measures are mutually exclusive i.e. the presence of skewness implies absence of kurtosis and vice-versa.
- Bowley's method of skewness is based on the values of median, lower and upper quartiles. This method suffers from the same limitations which are in the case of median and quartiles. Wherever positional measures are given, skewness should be measured by Bowley's method.
- Bowley's method is also used in case of 'open-end series', where the importance of extreme values is ignored.

6.6 UNIT END ACTIVITY

The length of stay on the cancer floor of XYZ Hospital was organized into a frequency distribution. The mean length of stay was 28 days, the medial 25 days and modal length is 23 days. The standard deviation was computed to be 4.2 days. Is the distribution symmetrical, or skewed? What is the coefficient of skewness? Interpret.

6.7 KEYWORDS

Moments: In statistics, moments are certain constant values in a given distribution which help us to ascertain the nature and form of distribution.

Skewness: Skewness is refers to the symmetry of the distribution.

Kurtosis: Kurtosis is the degree of flatness or 'peakedness' in the region of mode of a frequency curve.

Coefficient of Kurtosis: It a measure of the relative peakedness of the top of a frequency curves.

Measure of Skewness: Measure of skewness is the technique to indicate the direction and extent of skewness in the distribution values in the data set.

Moment of Order: It is defined as the arithmetic mean of the r^{th} power of deviations of observations.

Platykurtic: Negative kurtosis indicates a flatter distribution than the normal distribution, and called as platykurtic.

Leptokurtic: A positive kurtosis means more peaked curve, called Leptokurtic.

Mesokurtic: Peakedness of normal distribution is called Mesokurtic.

6.8 QUESTIONS FOR DISCUSSION

1. Define coefficients based on moments.
2. Explain moments about origin.
3. Discuss Charlier's check of accuracy.
4. What is skewness?

5. What are the characteristics of a good measure of skewness?
6. How do you calculate Bowley's coefficient of skewness?
7. What do you understand by measures of Kurtosis?
8. Explain the terms Platykurtic, Leptokurtic and Mesokurtic.
9. Define moments of the distribution about mean.
10. It is given that $\Sigma fx' = -10$, $\Sigma fx'^2 = 400$, $\Sigma fx'^3 = -1000$, $\Sigma fx'^4 = 5000$ and N = Value of W 10. Find the first four central moments.
11. Given the following information: Mean = 10, Variance = 16 and $\sqrt{\beta_1} = +1$.

    Obtain first three moments about origin.
12. For a mesokurtic distribution the first moment about 7 is 23 and the second moment about origin is 1000. Find the coefficient of variation and the fourth moment about mean.
13. For a normal distribution, the first moment about origin is 35 and the second moment about 35 is 10. Find the first four central moments.
14. Describe the relation between central moments and moments about origin.
15. How do you calculate Karl Pearsons's coefficient of skewness?
16. What are the properties of moments?
17. The first three moments of a distribution about the value 2 are 1, 16 and – 40. Show that the mean of the distribution is 3, the variance is 15 and $\mu_3 = -86$. Also, show that the first three moments about 0 are 3, 24 and 76.
18. For a distribution of 100 items, $\bar{X} = 54$, $\sigma = 3$, $\beta_1 = 0$ and $\beta_2 = 3$. If two observations 64 and 50 were wrongly recorded as 62 and 52, find the corrected value of the above measures.
19. The following data are given to an economist for the purpose of analysis:

    $n = 100$, $\Sigma fd = 100$, $\Sigma fd^2 = 2000$, $\Sigma fd^3 = 300$, $\Sigma fd^4 = 60{,}000$,

    where $d = X - 48$. Calculate $\bar{X}$, μ_2, μ_3 and μ_4.
20. For a distribution, the first four moments are 1, 7, 38 and 155 respectively.

    (i) Compute the moment coefficients of skewness and kurtosis

    (ii) Is the distribution mesokurtic? Give reasons.
21. The standard deviation of a distribution is 4. What should be the value of fourth central moment in order that the distribution be mesokurtic.
22. For a distribution the mean is 10, the standard deviation is 4, $\sqrt{\beta_1} = 1$ and $\beta_2 = 4$. Obtain the first four moments about origin.
23. In a certain distribution, the first four moments about the point 4 are – 1.5, 17, – 30 and 108. Calculate β_1 and β_2 and comment on the nature of the frequency curve as regards to skewness and kurtosis.
24. For a certain normal distribution, the first moment about 8 is 22 and the fourth moment about 30 is 243. Find mean and standard deviation of the distribution.
25. The sum of 20 observations is 300, sum of squares is 5000 and median is 15. Find coefficient of variation and skewness.

26. The first three moments of a distribution about a value 3 of the variable are 2, 10 and 30 respectively. Obtain the first three moments about origin. What are the mean, variance and $\sqrt{\beta_1}$ of such a distribution?

27. The first four moments from mean of a distribution are 0, 3.2, 3.6 and 20. The mean value is 11. Calculate the first four moments about zero and about 10.

28. Compute the moment measure of skewness from the following distribution:

Marks obtained	:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	:	8	14	22	26	15	10	5

Check Your Progress: Model Answer

1. r^{th}
2. Physics
3. R.A. Fishe
4. Asymmetry
5. β_1
6. Bulginess

6.9 REFERENCE & SUGGESTED READINGS

- Groebner, D. F., Shannon, P. W., Fry, P. C., & Smith, K. D. (2022). **Business Statistics: A Decision-Making Approach** (11th ed.). Pearson. ISBN: 9780136681503
- Albright, S. C., Winston, W. L., & Zappe, C. (2021). **Data Analysis and Decision Making** (6th ed.). Cengage Learning. ISBN: 9780357131785
- Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2020). **Statistics for Business and Economics** (14th ed.). Cengage Learning. ISBN: 9780357114474
- Jaggia, S., Kelly, A., &Lertwachara, K. (2021). **Essentials of Business Statistics** (2nd ed.). McGraw-Hill Education. ISBN: 9781260205799
- Bowerman, B. L., O'Connell, R. T., Murphree, E. S., & Orris, J. B. (2018). **Essentials of Business Statistics** (6th ed.). McGraw-Hill Education. ISBN: 9781259549939

UNIT - VII

ANALYSIS OF TIME SERIES

CONTENTS

7.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Discuss the meaning of time series
- Know the objectives of time series analysis
- Describe variations in time series
- Define trend analysis
- Explain seasonal variations and irregular variations
- Discuss decomposition of time series

7.1 INTRODUCTION

Time has strange, fascinating and little understood properties. Virtually every process on earth is determined by a time variable. One of the most frequently encountered managerial decision situations involving forecasting is to measure the effect that time has on the sales of a product, the market price of a security, the output of individuals, work shifts, companies, industries, societies and so on. A fundamental conceptual model in all of these situations is the product life cycle concept which goes through four stages – introduction, growth, maturity and decline. Let us look at this concept in greater detail before we apply it. A series of observations, on a variable, recorded after successive intervals of time is called a time series. The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, an hour, etc. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

7.2 MEANING AND OBJECTIVES OF TIME SERIES

A series of observations, on a variable, recorded after successive intervals of time is called a time series. It should be noted here that the time series data are bivariate data in which one of the variables is time. This variable will be denoted by t. The symbol Y_t will be used to denote the observed value, at point of time t, of the other variable. If the data pertains to n periods, it can be written as (t, Y_t), $t = 1, 2, ... n$.

7.2.1 Objectives of Time Series Analysis

The analysis of time series implies its decomposition into various factors that affect the value of its variable in a given period. It is a quantitative and objective evaluation of the effects of various factors on the activity under consideration.

There are two main objectives of the analysis of any time series data:

1. To study the past behaviour of data.
2. To make forecasts for future.

The study of past behaviour is essential because it provides us the knowledge of the effects of various forces. This can facilitate the process of anticipation of future course of events and, thus, forecasting the value of the variable as well as planning for future.

7.2.2 Analysis of Time Series

As mentioned earlier, the purpose of analysis of a time series is to decompose Y_t into various components. However, before doing this, we have to make certain assumptions regarding the manner in which these components have combined themselves to give the value Y_t. Very often it is assumed that Y_t is given by either the summation or the multiplication of various components, and accordingly you may assume two type of models, i.e., additive model or multiplicative model.

1. ***Additive Model:*** This model is based on the assumption that the value of the variable of a time series, at a point of time t, is the sum of the four components. Using symbols, we can write

   $Y_t = T_t + S_t + C_t + R_t$, where T_t, S_t, C_t and R_t are the values of trend, seasonal, cyclical and random components respectively, at a point of time t.

   This model assumes that all the four components of time series act independently of one another. This assumption implies that one component has no effect on the other(s) irrespective of their magnitudes.

2. ***Multiplicative Model:*** This model assumes that Y_t is given by the multiplication of various components. Symbolically, we can write

   $Y_t = T_t \times S_t \times C_t \times R_t$

   This model implies that although the four components may be due to different causes, these are, strictly speaking, not independent of each other. For example, the seasonal component may be some percentage of trend. Similarly, we can have other components expressed in terms of certain percentage.

   There is, in fact, very little agreement amongst the experts about the validity of the models assumed above. It is not very certain that the components combine themselves in the manner mentioned in the two models. Consequently, various mixed type of models have also been suggested, such as

   $Y_t = T_t.S_t.C_t + R_t$

   or $Y_t = T_t.C_t + S_t.R_t$ or $Y_t = T_t + C_t.S_t.R_t$, etc.

Out of all the models, given above, the additive and the multiplicative models are often used. The two models, when applied to the same data, would give different answers.

7.3 COMPONENTS OF A TIME SERIES

An observed value of a time series, Y_t, is the net effect of many types of influences such as changes in population, techniques of production, seasons, level of business activity, tastes and habits, incidence of fire floods, etc. It may be noted here that different types of variables may be affected by different types of factors, e.g., factors affecting the agricultural output may be entirely different from the factors affecting industrial output. However, for the purpose of time series analysis, various factors are classified into the following three general categories applicable to any type of variable.

1. Trend analysis
2. Periodic or Oscillatory Variations
   (a) Cyclical Variations
   (b) Seasonal Variations
3. Random or Irregular Variations

7.4 SECULAR TREND

Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time. Most of the business and economic time series would reveal a tendency to increase or to decrease over a number of years. For example, data regarding industrial production, agricultural production, population, bank deposits, deficit financing, etc., show that, in general, these magnitudes have been rising over a fairly long period. As opposed to this, a time series may also reveal a declining trend, e.g., in the case of substitution of one commodity by another, the demand of the substituted commodity would reveal a declining trend such as the demand for cotton clothes, demand for coarse grains like bajra, jowar, etc. With the improved medical facilities, the death rate is likely to show a declining trend, etc. The change in trend, in either case, is attributable to the fundamental forces such as changes in population, technology, composition of production, etc.

According to A.E. Waugh, secular trend is, *"that irreversible movement which continues, in general, in the same direction for a considerable period of time"*. There are two parts of this definition; (i) movement in same direction, which implies that if the values are increasing (or decreasing) in successive periods, the tendency continues; and (ii) a considerable period of time. There is no specific period which can be called as a long period. Long periods are different for different situations. For example, in cases of population or output trends, the long period could be 10 years while it could be a month for the daily demand trend of vegetables.

7.4.1 Objectives of Measuring Trend

There are four main objectives of measuring trend of a time series data:

1. To study past growth or decline of the series. On ignoring the short-term fluctuations, trend describes the basic growth or decline tendency of the data.
2. Assuming that the same behaviour would continue in future also, the trend curve can be projected into future for forecasting.
3. In order to analyse the influence of other factors, the trend may first be measured and then eliminated from the observed values.
4. Trend values of two or more time series can be used for their comparison.

7.4.2 Measurement of Secular Trend

The following are the principal methods of measuring trend from a given time series:

1. Graphic or Free Hand Curve Method
2. Method of Averages
   (a) Method of Selected Points
   (b) Method of Semi-averages
   (c) Method of Moving Averages
3. Mathematical Trends
   (a) Method of Least Squares
      (i) Fitting of Linear Trend
      (ii) Fitting of Parabolic Trend
      (iii) Fitting of Exponential Trend
   (b) Method of Selected Points and Method of Semi-averages
      (i) Modified Exponential Curve
      (ii) Gompertz Curve
      (iii) Logistic Curve

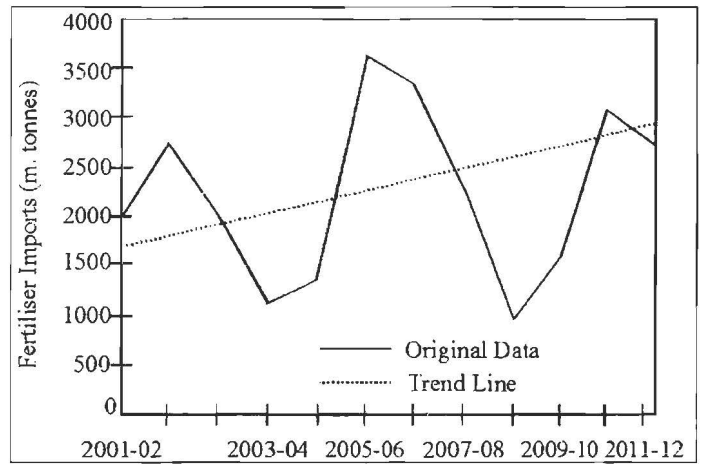
Graphic or Free Hand Curve Method

This is the simplest method of studying the trend. The given time series data are plotted on a graph paper by taking time on X-axis and the other variable on Y-axis. A smooth line or curve, drawn through the plotted points, would represent the trend of the given data.

Example: Determine the trend of the following time series data by graphical method.

Years	2001–02	2002–03	2003–04	2004–05	2005–06	2006–07
Fert. Imports	2005	2759	2041	1132	1355	3624
Years	2007–08	2008–09	2009–10	2010–11	2011–12	2012–13
Fert. Imports	3399	2310	984	1608	3114	2758

Solution:



The dotted line, shown in the figure, is the required trend line. This line can be extended to get the predicted values for future.

Merits

(i) It is a simple method of estimating trend which requires no mathematical calculations.

(ii) It is a flexible method as compared to rigid mathematical trends and, therefore, a better representative of the trend of the data.

(iii) If the observations are relatively stable, the trend can easily be approximated by this method.

Demerits

(i) It is a subjective method. The values of trend, obtained by different statisticians would be different and hence, not reliable.

(ii) Prediction made on the basis of this method is of little value.

Method of Averages

The following are the method of averages:

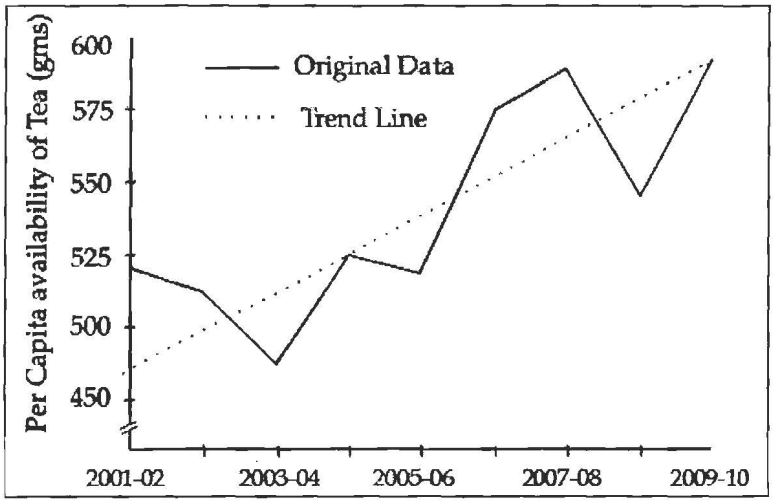
Method of Selected Points

In this method, two points, considered to be the most representative or normal, are joined by a straight line to get secular trend. This, again, is a subjective method since different persons may have different opinions regarding the representative points. Further, only linear trend can be determined by this method.

Example: Determine the trend of the following time series data by the method of selected points:

Years	2001-02	2002-03	2003-04	2004-05	2005-06
Per Capita availability of Tea (gms)	521	511	462	525	518
Years	2006-07	2007-08	2008-09	2009-10	
Per Capita availability of Tea (gms)	575	589	546	593	

Solution:



In the above figure, we have taken the years 2004-2005 and 2009-2010 as the normal years. The corresponding points are joined by a straight line to get the trend of the observed values.

Method of Semi-averages

The given time series data are divided into two equal parts and the arithmetic mean of the values of each part is computed. The computed means are termed as semi-averages. Each semi-average is paired with the centre of time period of its part. The two pairs are then plotted on a graph paper and the points are joined by a straight line to get the trend. It should be pointed out here that in case of odd number of observations, the two equal parts are obtained by dropping the middle most observation.

Merits

- It is simple method of measuring trend.
- It is an objective method because anyone applying this to a given data would get identical trend values.

Demerits

- This method can give only a linear trend of the data irrespective of whether it exists or not.
- This is only a crude method of measuring trend, since we do not know whether the effect of other components is completely eliminated or not.

Method of Moving Average

This method is based on the principle that the total effect of periodic variations at different points of time in its cycle gets completely neutralized, i.e., $\Sigma S_t = 0$ in one year and $\Sigma C_t = 0$ in the period of cyclical variations.

In the method of moving average, successive arithmetic averages are computed from overlapping groups of successive values of a time series. Each group includes all the observations in a given time interval, termed as the period of moving average. The next group is obtained by replacing the oldest value by the next value in the series. The averages of such groups are known as the moving averages.

The moving average of a group is always shown at the centre of its period. The process of computing moving averages smoothens out the fluctuations in the time series data. It can be shown that if the trend is linear and the oscillatory variations are regular, the moving average with period equal to the period of oscillatory variations would completely eliminate them. Further, the effect of random variations would get minimised because the average of a number of observations must lie between the smallest and the largest observation. It should be noted here that the larger is the period of moving average the more would be the reduction in the effect of random component but more information is lost at the two ends of data.

When the trend is non-linear, the moving averages would give biased rather than the actual trend values.

Let $Y_1, Y_2, \ldots\ldots Y_n$ be the n values of a time series for successive time periods 1, 2, ... n respectively. The calculation of 3-period and 4-period moving averages are shown in the following tables:

Time Period	*Values of Y*	*3 - period M.A.*
1	Y_1	...
2	Y_2	$\frac{Y_1+Y_2+Y_3}{3}$
3	Y_3	$\frac{Y_2+Y_3+Y_4}{3}$
4	Y_4	$\frac{Y_3+Y_4+Y_5}{3}$
5	Y_5	...
⋮	⋮	⋮
n	Y_n	...

Time Period	*Values of Y*	*4 - period M.A.*	*Centered Values*
1	Y_1	...	...
2	Y_2	...	...
3	Y_3	$\frac{Y_1+Y_2+Y_3+Y_4}{4}=A_1$	$\frac{A_1+A_2}{2}$
4	Y_4	$\frac{Y_2+Y_3+Y_4+Y_5}{4}=A_2$	$\frac{A_2+A_3}{2}$
5	Y_5	$\frac{Y_3+Y_4+Y_5+Y_6}{4}=A_3$	
⋮	⋮	⋮	⋮
n	Y_n	...	...

It should be noted that, in case of 3-period moving average, it is not possible to get the moving averages for the first and the last periods. Similarly, the larger is the period of moving average the more information will be lost at the ends of a time series.

When the period of moving average is even, the computed average will correspond to the middle of the two middle most periods. These values should be centered by taking arithmetic mean of the two successive averages. The computation of moving average in such a case is also illustrated in the above table.

Example: Determine the trend values of the following data by using 3-year moving average. Also find short-term fluctuations for various years, assuming additive model. Plot the original and the trend values on the same graph.

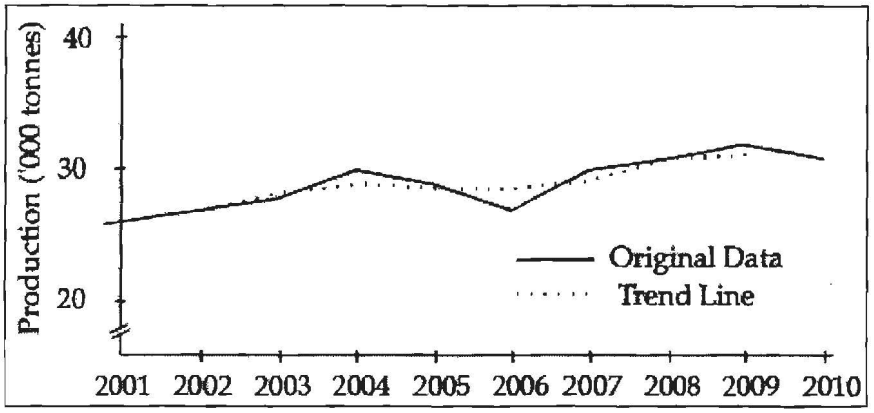
Years	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Production (000' tonnes)	26	27	28	30	29	27	30	31	32	31

Solution:

Calculation of Trend and Short-term Fluctuations

Years	Production (Y)	3-Years Moving Total	3-Years M.A. or Trend values (T)	Short-term fluctuations (Y – T)
2001	26	...	...	...
2002	27	81	27.00	0.00
2003	28	85	28.33	–0.33
2004	30	87	29.00	1.00
2005	29	86	28.67	0.33
2006	27	86	28.67	–1.67
2007	30	88	29.33	0.67
2008	31	93	31.00	0.00
2009	32	94	31.33	0.67
2010	31	...	...	...

Graphical Presentation of Y and T Values



Merits

- This method is easy to understand and easy to use because there are no mathematical complexities involved.
- It is an objective method.
- It is a flexible method in the sense that if a few more observations are added, the entire calculations are not changed.
- When the period of oscillatory movements is equal to the period of moving average, these movements are completely eliminated.
- By the indirect use of this method, it also possible to isolate seasonal, cyclical and random components.

Demerits

- It is not possible to calculate trend values for all the items of the series. Some information is always lost at its ends.
- This method can determine accurate values of trend only if the oscillatory and random fluctuations are uniform in terms of period and amplitude and the trend is,

at least, approximately linear. However, these conditions are rarely met in practice. When the trend is not linear, the moving averages will not give correct values of trend.

- The selection of period of moving average is a difficult task and a great deal of care is needed to determine it.
- Like arithmetic mean, the moving averages are too much affected by extreme values.
- The trend values obtained by moving averages may not follow any mathematical pattern and thus, cannot be used for forecasting, which perhaps is the main task of any time series analysis.

Mathematical Trends

The method of fitting a mathematical trend to given time series data is perhaps the most popular and satisfactory. The form of mathematical equation used for the determination of trend depends upon the nature of the broad idea of trend, obtained by graphic representation of data or otherwise. Some popularly known forms of trend are linear, parabolic, exponential and growth curves. A brief description of these is given below:

- ***Linear Trend:*** The general form of a linear trend is given by the equation $Y_t = a + bt$, where t denotes time, Y_t is the trend value (note that trend values, in mathematical models, will be denoted by Y_t rather than by T_t for the sake of convenience) of variable at time t and a (> 0) and b (a real number) are constants. The constant a can be interpreted as the value of trend (Y_t) when t = 0 and b gives the change in Y_t per unit change in time. It should be noted that the rate of change of Y_t is always constant in case of a linear trend. This implies that for equal absolute changes in t, there are correspondingly equal absolute changes in Y_t. Further, a linear trend can be rising or falling according as b > 0 or < 0, as shown in Figure 7.1.

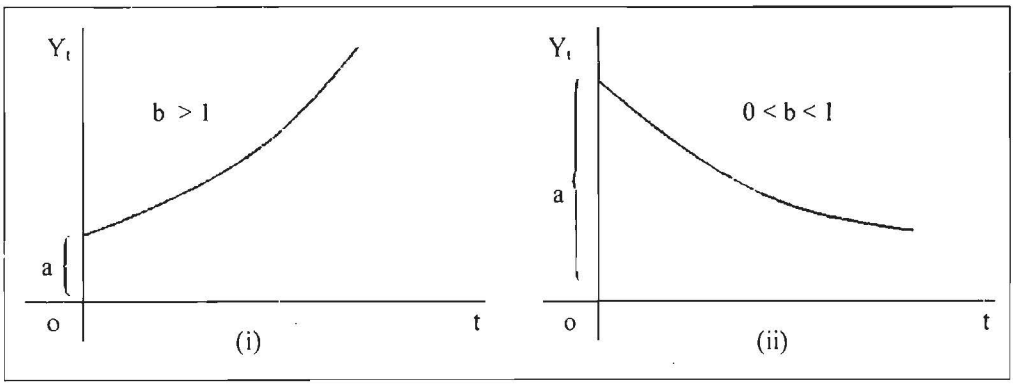


Figure 7.1: Rising and Falling Linear Trend

- ***Parabolic Trend:*** The general form of a parabolic trend is $Y_t = a + bt + ct^2$, where a, b and c are constants. Here the rate of change of Y_t is different at different time periods. The possible shapes of parabolic trends are shown in Figure 7.2.

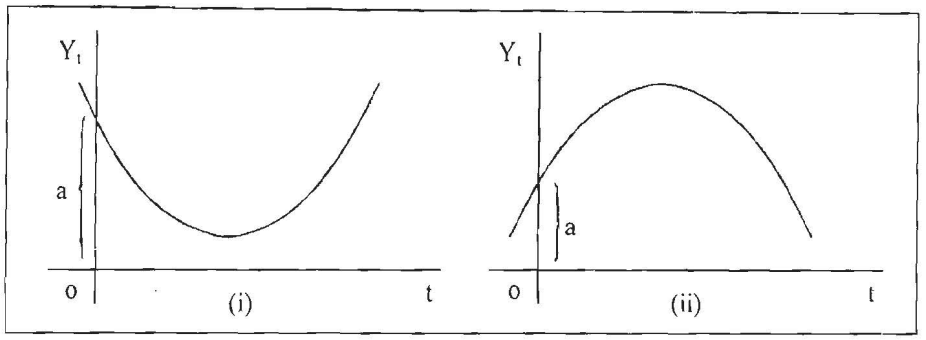


Figure 7.2: Increasing and Decreasing Parabolic Trend

We note that the rate of change of Y_t is increasing in the first case while it is decreasing in the second.

- ***Exponential Trend:*** The general form of an exponential trend is given by the equation $Y_t = a.b^t$, where a and b are positive constants. This implies that values Y_t changes by a constant percentage per unit of time. For example, if a = 50 and b = 1.05, then

  $Y_1 = 50 \times 1.05 \Rightarrow$ 5% increase in the value of a.

  Similarly, $Y_2 = 50 \times (1.05)^2 = Y_1 \times 1.05 \Rightarrow$ 5% increase in the value of Y_1 and so on. We note that when $b > 1$, the exponential trend is increasing. In a similar way, it would be decreasing when $0 < b < 1$, as shown in Figure 7.3.

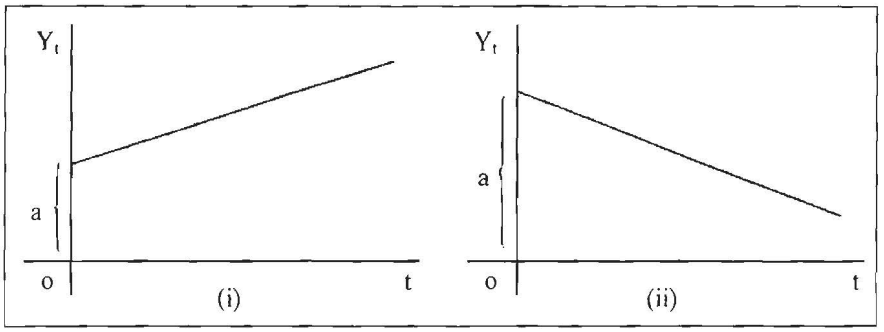


Figure 7.3: Increasing and Decreasing Exponential Trend

- ***Growth Curves:*** The mathematical trends like straight line, parabola or exponential curve are such that they would show either increasing or decreasing or first increasing (decreasing) and then decreasing (increasing) trends throughout the entire period, without any limit. In most of the business and economic time series, we find that although the trend is increasing (or decreasing), it remains less (or greater) than a particular value. Such a behaviour is described by means of growth curves. Some important growth curves are Modified Exponential Curve, Gompertz Curve and Logistic or Pearl-Reed Curve.

  ❖ *Modified Exponential Curve:* The general form of a modified exponential curve is given by

  $Y_t = k + a.b^t$, where k, a and b are constants.

  For a fixed value of k, different trends can be obtained from various combinations of the values of a and b as shown below.

  (a) When $a < 0$ and $0 < b < 1$: This curve is shown in Figure 7.4(i), which indicates that Y_t increases at decreasing rate per unit of time and always remains less than k.

(b) When a > 0 and 0 < b < 1: This curve is shown in Figure 7.4(ii), which indicates that Y_t decreases at increasing rate per unit of time and always remains greater than k.

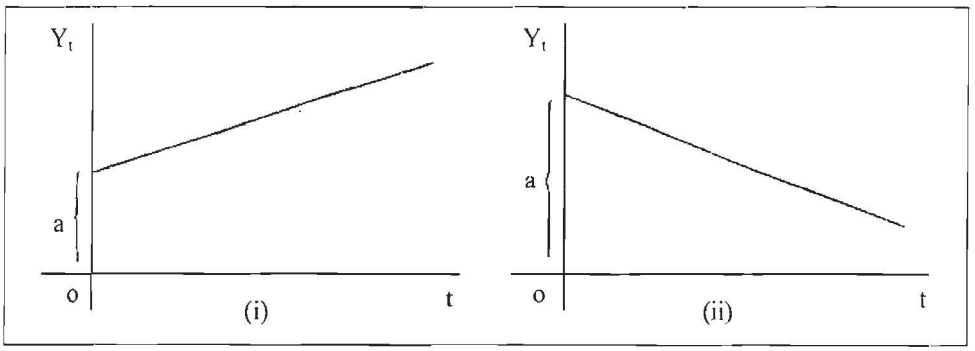


Figure 7.4: Growth Curve: Increasing and Decreasing Modified Exponential Curve

(c) When a < 0 and b > 1: This curve is shown in Figure 7.5(i), which indicates that the greatest value of Y_t is k + a and it decreases at decreasing rate per unit of time.

(d) When a > 0 and b > 1: This is shown in Figure 7.5(ii), which indicates that the lowest value of Y_t is k + a and it increases at increasing rate.

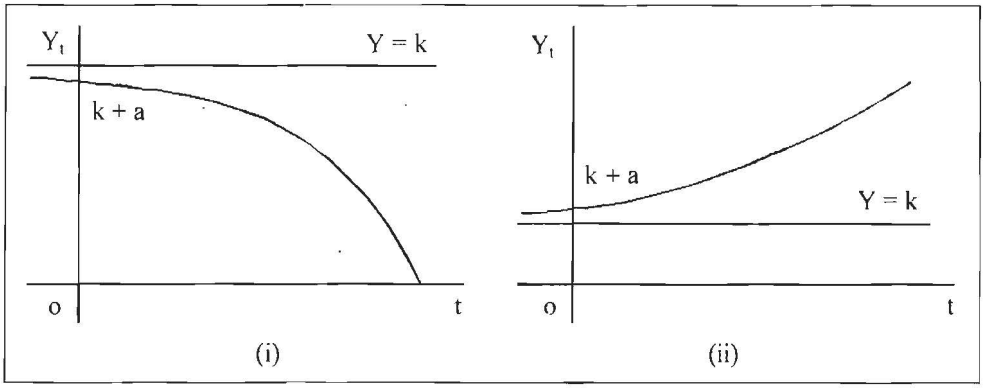


Figure 7.5: Growth Curve: Increasing and Decreasing Modified Exponential Curve

- *Gompertz Curve:* The general form of Gompertz curve is given by

  $Y_t = K.a^{b^t}$, where k, a and b are constants.

  Taking log of both sides, we have

  $\log Y_t = \log k + b^t.\log a$, which is of modified exponential form.

- *Logistic or Pearl-Reed Curve:* In the case of a modified exponential curve, the rate of increase (or decrease) of Y_t was either increasing or decreasing throughout the whole period. In most of the business and economic time series, we may have a situation where the rate of increase of Y_t is very high in the early stages and then gradually declines and reaches a saturation. This type of behaviour can be approximated by a Logistic curve, which is identical to a modified exponential curve.

The general form of a Logistic curve is given by $Y_t = \frac{K}{1+e^{f(t)}}$, where k is a constant and f(t) is a polynomial of degree m.

Note: If m = 3, the polynomial can be written as $f(t) = a + bt + ct^2 + dt^3$, where a, b, c and are constants.

A polynomial of degree one is, often, considered and thus, we can write the Logistic curve as $Y_t = \frac{K}{1 + e^{a+bt}}$ ... (1)

Further, we can also use base 10 instead of e. Thus, we get $Y_t = \frac{K}{1 + 10^{a+bt}}$

When $b < 0$, $a + bt$ tends to be negative for large values of t and hence e^{a+bt} or 10^{a+bt} tends to be zero. Thus, $Y_t \to k$ for large values of t. Similarly, $Y_t \to 0$ as $t \to -\infty$. Hence, the curve lies between two asymptotes $Y_t = k$ and $Y_t = 0$, as shown in Figure 7.6(a).

Further, when $b > 0$, the curve would again lie between these asymptotes. However, Y_t would be decreasing, rather than increasing, in this case, as shown Figure 7.6(b).

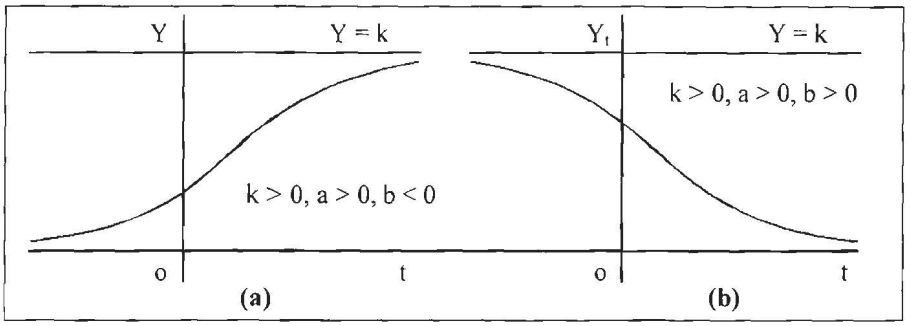


Figure 7.6: Growth Curve: Increasing and Decreasing Logistic or Pearl-Reed Curve

It can be shown that a Logistic curve is closely related to a modified exponential curve. Taking the reciprocal of the Logistic curve given by (1), we get

$$\frac{1}{Y_t} = \frac{1 + e^{a+bt}}{k} = \frac{1}{k} + \frac{e^{a+bt}}{k} = \frac{1}{k} + \frac{e^a}{k} \times e^{bt}$$

Let $\frac{1}{k} = \alpha$ $\frac{e^a}{k} = A$ and $e^b = B$

$\therefore$ $\frac{1}{Y_t} = \alpha + A.B^t$, which is identical to the modified exponential form.

Method of Least Squares

This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimised. We shall use this method in the fitting of following trends:

(a) Linear Trend

(b) Parabolic Trend

(c) Exponential Trend

Fitting of Linear Trend

Given the data (Y_t, t) for n periods, where t denotes time period such as year, month, day, etc., we have to find the values of the two constants, a and b, of the linear trend equation $Y_t = a + bt$.

Using the least square method. the normal equation for obtaining the values of a and b are:

$$\Sigma Y_t = na + b\ \Sigma t$$

$$\Sigma tY_t = a\Sigma t + b\Sigma t^2$$

Let $X = t - A$, such that $\Sigma X = 0$, where A denotes the year of origin.

The above equations can also be written as

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

(Dropping the subscript t for convenience).

Since X = 0, we can write $a = \frac{\sum Y}{n}$ and $b = \frac{\sum XY}{\sum X^2}$.

This distinction will become obvious from the following two examples.

Example: Fit a straight line trend to the following data and estimate the likely profit for the year 2016. Also calculate various trend values.

Years	:	2007	2008	2009	2010	2011	2012	2013
Profit (in lacs of ₹)	:	60	72	75	65	80	85	95

Solution:

Calculation Table

Years (t)	Y	X = t – 1980	XY	X^2	Trend Values
2007	60	–3	–180	9	61.42
2008	72	–2	–144	4	66.28
2009	75	–1	–75	1	71.14
2010	65	0	0	0	76.00
2011	80	1	80	1	80.86
2012	85	2	170	4	85.72
2013	95	3	285	9	90.58
Total	**532**	**0**	**136**	**28**	

From the table, we can write $a = \frac{532}{7} = 76$ (n = 7, the no. of observations) and

$b = \frac{136}{28} = 4.86$

Thus, the fitted line of trend is $Y = 76 + 4.86X$

Thus, the appropriate way of writing the trend equation would be:

$Y = 76 + 4.86X$, where (i) year of origin = 1st July 2010 (the year in which X = 0), (ii) unit of X = 1 year and (iii) Y's are annual figures of profits.

Calculation of Trend Values: Trend value of a particular year is obtained by substituting the associated value of X in the trend equation. For example, X = – 3 for 2007, therefore, trend for 2007 is $Y = 76 + 4.86\ (-3) = 61.42$.

Alternatively, trend values can be calculated as follows:

We know that a is the trend value in the year of origin and b gives the rate of change per unit of time. Thus, the trend for 2010 = 76, for 2009 = 76 – 4.86 = 71.14, for 2008 = 71.14 – 4.86 = 66.28 and for 2007 = 66.28 – 4.86 = 61.42, etc. Similarly, trend for 2011 = 76 + 4.86 = 80.86, for 2012 = 80.86 + 4.86 = 85.72, etc.

Prediction of Trend for a Year: Using the trend equation, we can predict a trend value for a year which doesn't belong to the observed data. To predict the value for 2016, the associated value of X = 6. Substituting this in the trend equation we get Y = 76 + 6 × 4.86 = ₹ 105.16 lacs.

Remarks: The prediction of trend is only valid for periods that are not too far from the observed data.

Example: Fit a straight line trend, by the method of least squares, to the following data. Assuming that the same rate of change continues, what would be the predicted sales for 2014?

Years	:	2008	2009	2010	2011	2012	2013
Sales (in '000 ₹)	:	15	17	20	21	23	24

Solution:

We note that n is even in the given example.

Calculation Table

Years (t)	Sales (Y)	d = t – 2010.5	X = 2d	XY	X^2	Trend Values
2008	15	–2.5	–5	–75	25	15.45
2009	17	–1.5	–3	–51	9	17.27
2010	20	–0.5	–1	–20	1	19.09
2011	21	0.5	1	21	1	20.91
2012	23	1.5	3	69	9	22.73
2013	24	2.5	5	120	25	24.55
Total	**120**		**0**	**64**	**70**	

From the above table, we can write

$$a = \frac{120}{6} = 20 \quad \text{and} \quad b = \frac{64}{70} = 0.91$$

∴ The fitted trend line is Y = 20 + 0.91 X

Year of origin : Middle of 2010 and 2011 or 1st Jan. 2011

Unit of X : ½ year (Since X changes by 2 units in one year)

Nature of Y values: Annual figures of sales.

Calculation of Trend Values

Trend for 2010 = 20 – 0.91 = 19.09

Trend for 2009 = 19.09 – 2 × 0.91 = 17.27

Trend for 2011 = 20 + 0.91 = 20.91

Trend for 2012 = 20.91 + 2 × 0.91 = 22.73, etc.

To predict the sales for 2014, we note that X = 7

Thus, the predicted sales = 20 + 7 × 0.91 = ₹ 26.37 (thousand).

Shifting of Origin of a Trend Equation: Let Y = a + bX be the equation of linear trend, with 2006 as the year of origin and unit of X equal to 1 year.

To shift origin of the above equation, say to 2011, we proceed as follows: The associated value of X for 2011 is 5. Thus, the trend for 2011 = a + 5b. We know that a linear trend equation is given by Y = trend value in the year of origin + bX. Thus, we can write the trend equation, with origin at 2011, as Y = a + 5b + bX = a + b (X + 5). This implies that the required equation can be obtained by replacing X by X + 5 in the original trend equation.

Similarly, the trend equation with 2005 as origin can be written as Y = a + b (X – 1) = (a – b) + bX.

Further, if the unit of X is given to be half year, the trend equation with 2011 as the year of origin can be written as Y = a + b (X + 10) = (a + 10b) + bX.

Fitting of Parabolic Trend

The mathematical form of a parabolic trend is given by $Y_t = a + bt + ct^2$ or $Y = a + bt + ct^2$ (dropping the subscript for convenience). Here a, b and c are constants to be determined from the given data.

Using the method of least squares, the normal equations for the simultaneous solution of a, b and c are:

$$\Sigma Y = na + b\ \Sigma t + c\Sigma t^2$$

$$\Sigma tY = a\Sigma t + b\Sigma t^2 + c\Sigma t^3$$

$$\Sigma t^2Y = a\Sigma t^2 + b\Sigma t^3 + c\Sigma t^4$$

By selecting a suitable year of origin, i.e., define X = t - origin such that X = 0, the computation work can be considerably simplified. Also note that if X = 0, then X^3 will also be equal to zero. Thus, the above equations can be rewritten as:

$$\Sigma Y = na + c\Sigma X^2 \quad \text{... (i)}$$

$$\Sigma XY = b\Sigma X^2 \quad \text{... (ii)}$$

$$\Sigma X^2Y = a\Sigma X^2 + c\Sigma X^4 \quad \text{... (iii)}$$

From equation (ii), we get

$$b = \frac{\sum XY}{\sum X^2} \quad \text{... (iv)}$$

Further, from equation (i), we get

$$a = \frac{\sum Y - c\sum X^2}{n} \quad \text{... (v)}$$

And from equation (iii), we get

$$c = \frac{n\sum X^2Y - \left(\sum X^2\right)\left(\sum Y\right)}{n\sum X^4 - \left(\sum X^2\right)^2} \quad \text{.... (vi)}$$

Thus, equations (iv), (v) and (vi) can be used to determine the values of the constants a, b and c.

Example: Fit a parabolic trend $Y = a + bt + ct^2$ to the following data, where t denotes years and Y denotes output (in thousand units).

t	2005	2006	2007	2008	2009	2010	2011	2012	2013
Y	2	6	7	8	10	11	11	10	9

Also compute the trend values. Predict the value for 2014.

Solution:

Calculation Table

t	Y	X = t – 1985	XY	X^2Y	X^2	X^3	X^4	Trend Values
2005	2	–4	–8	32	16	–64	256	2.28
2006	6	–3	–18	54	9	–27	81	5.02
2007	7	–2	–14	28	4	–8	16	7.22
2008	8	–1	–8	8	1	–1	1	8.88
2009	10	0	0	0	0	0	0	10.00
2010	11	1	11	11	1	1	1	10.58
2011	11	2	22	44	4	8	16	10.62
2012	10	3	30	90	9	27	81	10.12
2013	9	4	36	144	16	64	256	9.08
Total	**74**	**0**	**51**	**411**	**60**	**0**	**708**	

From the above table, we can write

$$b = \frac{51}{60} = 0.85$$

$$c = \frac{9 \times 411 - 60 \times 74}{9 \times 708 - (60)^2} = -0.27$$

$$a = \frac{74 - (-0.27) \times 60}{9} = 10.0$$

The fitted trend equation is $Y = 10.0 + 0.85X - 0.27X^2$,

with origin = 2009 and unit of X = 1 year.

Various trend values are calculated by substituting appropriate values of X in the above equation. These values are shown in the last column of the above table.

The predicted value for 2014 is given by

$$Y = 10.0 + 0.85 \times 5 - 0.27 \times 25 = 7.5$$

Fitting of Exponential Trend

The general form of an exponential trend is $Y = a.b^t$, where a and b are constants to be determined from the observed data.

Taking logarithms of both sides, we have $\log Y = \log a + t \log b$.

This is a linear equation in log Y and t and can be fitted in a similar way as done in case of linear trend. Let A = log a and B = log b, then the above equation can be written as log Y = A + Bt.

The normal equations, based on the principle of least squares are

$$\Sigma \log Y = nA + B\Sigma t$$

and $$\Sigma t \log Y = A\Sigma t + B\Sigma t^2$$

By selecting a suitable origin, i.e., defining X = t - origin, such that X = 0, the computation work can be simplified. The values of A and B are given by

$$A = \frac{\sum \log Y}{n} \text{ and } B = \frac{\sum X \log Y}{\sum X^2} \text{ respectively.}$$

Thus, the fitted trend equation can be written as log Y = A + BX

or Y = Antilog [A + BX] = Antilog [log a + X log b]

= Antilog [log $a.b^X$] = $a.b^X$.

Example: Fit an exponential trend Y = a.bt to the following data:

Census Year (t)	1963	1973	1983	1993	2003	2013
Population of India (in Crores)	31.9	36.1	43.9	54.8	68.3	84.4

Predict the population for 2023.

Solution:

Calculation Table

Census Year t	**Population Y**	$X = \frac{(t - 1988)}{5}$	**log Y**	**X log Y**	X^2
1963	31.9	–5	1.5038	–7.5190	25
1973	36.1	–3	1.5575	–4.6725	9
1983	43.9	–1	1.6425	– 1.6425	1
1993	54.8	1	1.7388	1.7388	1
2003	68.3	3	1.8344	5.5032	9
2013	84.4	5	1.9263	9.6315	25
Total		**0**	**10.2033**	**3.0395**	**70**

From the above table, we get $A = \frac{10.2033}{6} = 1.70$ and $B = \frac{3.0395}{70} = 0.043$

Further, a = antilog 1.70 = 50.12 and b = antilog 0.043 = 1.10

Thus, the fitted trend equation is $Y = 50.12(1.10)^X$.

Origin : 1st July, 1988 and unit of X = 5 years.

The trend values can be computed by the equation Y = antilog [1.70 + 0.043X]. Further, the prediction of population for 2023 is obtained by substituting X = 7, in the above equation.

Y = antilog[1.70 + 0.043 7] = antilog[2.001] = 100.2 crores

Merits of Least Squares Method

- Given the mathematical form of the trend to be fitted, the least squares method is an objective method.
- Unlike the moving average method, it is possible to compute trend values for all the periods and predict the value for a period lying outside the observed data.
- The results of the method of least squares are most satisfactory because the fitted trend satisfies the two important properties, i.e., (i) $(Y_o - Y_t) = 0$ and (ii) $(Y_o - Y_t)^2$ is minimum. Here Y_o denotes the observed value and Y_t denotes the calculated trend value.

The first property implies that the position of fitted trend equation is such that the sum of deviations of observations above and below this is equal to zero. The second property implies that the sum of squares of deviations of observations, about the trend equation, are minimum.

Demerits of Least Squares Method

- As compared with the moving average method, it is a cumbersome method.
- It is not flexible like the moving average method. If some observations are added, then the entire calculations are to be done once again.
- It can predict or estimate values only in the immediate future or past.
- The computation of trend values, on the basis of this method, doesn't take into account the other components of a time series and hence not reliable.
- Since the choice of a particular trend is arbitrary, the method is not, strictly, objective.
- This method cannot be used to fit growth curves, the pattern followed by the most of the economic and business time series.

7.5 PERIODIC OR OSCILLATORY VARIATIONS

These variations, also known as oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation. These oscillations are shown in Figure 7.7:

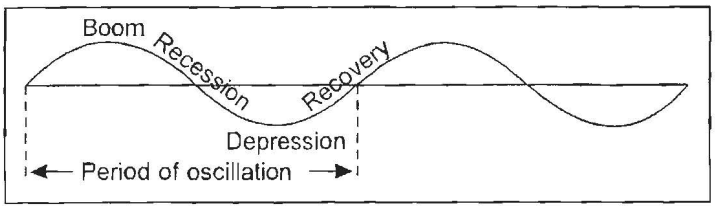


Figure 7.7: Periodic Variations

The oscillatory movements are termed as Seasonal Variations if their period of oscillation is equal to one year and as Cyclical Variations if the period is greater than one year.

A time series, where the time interval between successive observations is less than or equal to one year, may have the effects of both the seasonal and cyclical variations. However, the seasonal variations are absent if the time interval between successive observations is greater than one year.

Although the periodic variations are more or less regular, they may not necessarily be uniformly periodic, i.e., the pattern of their variations in different periods may or may

not be identical in respect of time period and size of periodic variations. For example, if a cycle is completed in five years then its following cycle may take greater or less than five years for its completion.

7.5.1 Cyclical Variations

Cyclical variations are revealed by most of the economic and business time series and, therefore, are also termed as trade (or business) cycles. Any trade cycle has four phases which are respectively known as boom, recession, depression and recovery phases. Various phases repeat themselves regularly one after another in the given sequence. The time interval between two identical phases is known as the period of cyclical variations.

Objectives of Measuring Cyclical Variations

The main objectives of measuring cyclical variations are:

(i) To analyse the behaviour of cyclical variations in the past.

(ii) To predict the effect of cyclical variations so as to provide guidelines for future business policies.

Measurement of Cyclical Variations

A satisfactory method for the direct measurement of cyclical variations is not available. The main reason for this is that although these variations may be recurrent, these are seldom found to be of similar pattern having same period and amplitude of oscillations. Moreover, in most of the cases these variations are so intermixed with random variations that it is very difficult, if not impossible, to separate them.

The cyclical variations are often obtained, indirectly, as a residue after the elimination of other components. Various steps of the method are as given below:

1. Compute the trend values (T) and the seasonal indices (S) by appropriate methods. Here S is obtained as a fraction rather than a percentage.
2. Divide Y-values by the product of trend and seasonal index. This ratio would consist of cyclical and random component.
3. If there are no random variations in the time series, the cyclical variations are given by the step (2) above. Otherwise the random variations should be smoothened out by computing moving averages of C.R. values with appropriate period. Weighted moving average with suitable weights may also be used, if necessary, for this purpose.

7.5.2 Seasonal Variations

If the time series data are in terms of annual figures, the seasonal variations are absent. These variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis. As discussed earlier, the seasonal variations are of periodic nature with period equal to one year. These variations reflect the annual repetitive pattern of the economic or business activity of any society. The main objectives of measuring seasonal variations are:

1. To understand their pattern.
2. To use them for short-term forecasting or planning.
3. To compare the pattern of seasonal variations of two or more time series in a given period or of the same series in different periods.
4. To eliminate the seasonal variations from the data. This process is known as deseasonalisation of data.

Causes of Seasonal Variations

The main causes of seasonal variations are as follows:

- ***Climatic Conditions:*** The changes in climatic conditions affect the value of time series variable and the resulting changes are known as seasonal variations. For example, the sale of woolen garments is generally at its peak in the month of November because of the beginning of winter season. Similarly, timely rainfall may increase agricultural output, prices of agricultural commodities are lowest during their harvesting season, etc., reflect the effect of climatic conditions on the value of time series variable.
- ***Customs and Traditions:*** The customs and traditions of the people also give rise to the seasonal variations in time series. For example, the sale of garments and ornaments may be highest during the marriage season, sale of sweets during Diwali, etc., are variations that are the results of customs and traditions of the people. It should be noted here that both of the causes, mentioned above, occur regularly and are often repeated after a gap of less than or equal to one year.

Example: Assuming that trend and cyclical variations are absent, compute the seasonal index for each month of the following data of sales (in ₹ '000) of a company.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011	46	45	44	46	45	47	46	43	40	40	41	45
2012	45	44	43	46	46	45	47	42	43	42	43	44
2013	42	41	40	44	45	45	46	43	41	40	42	45

Solution:

Calculation Table

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011	46	45	44	46	45	47	46	43	40	40	41	45
2012	45	44	43	46	46	45	47	42	43	42	43	44
2013	42	41	40	44	45	45	46	43	41	40	42	45
Total	133	130	127	136	136	137	139	128	124	122	126	134
A_i	44.3	43.3	42.3	45.3	45.3	45.7	46.3	42.7	41.3	40.7	42.0	44.7
S.I.	101.4	99.1	96.8	103.7	103.7	104.6	105.9	97.7	94.5	93.1	96.1	102.3

In the above table, A_i denotes the average and S.I. the seasonal index for a particular month of various years. To calculate the seasonal index, we compute grand average G, given by $G = \frac{\sum A_i}{12} = \frac{524}{12} = 43.7$. Then the seasonal index for a particular month is given by $S.I. = \frac{A_i}{G} \times 100$.

Further, S.I. = 1198.9 ≠ 1200. Thus, we have to adjust these values such that their total is 1200. This can be done by multiplying each figure by. The resulting figures are the adjusted seasonal indices.

Remarks: The totals equal to 1200, in case of monthly indices and 400, in case of quarterly indices, indicate that the ups and downs in the time series, due to seasons, neutralise themselves within that year. It is because of this that the annual data are free from seasonal component.

Example: Compute the seasonal index from the following data by the method of simple averages.

Year	Quarter	Y	Year	Quarter	Y	Year	Quarter	Y
2008	I	106	2010	I	90	2012	I	80
	II	124		II	112		II	104
	III	104		III	101		III	95
	IV	90		IV	85		IV	83
2009	I	84	2011	I	76	2013	I	104
	II	114		II	94		II	112
	III	107		III	91		III	102
	IV	88		IV	76		IV	84

Solution:

Calculation of Seasonal Indices

Years	1st Qr	2nd Qr	3rd Qr	4th Qr
2008	106	124	104	90
2009	84	114	107	88
2010	90	112	101	85
2011	76	94	91	76
2012	80	104	95	83
2013	104	112	102	84
Total	540	660	600	506
A_i	90	110	100	84.33
$\frac{A_i}{G} \times 100$	93.67	114.49	104.07	87.77

We have $G = \frac{\sum A_i}{4} = \frac{384.33}{4} = 96.08$.

Further, since the sum of terms in the last row of the table is 400, no adjustment is needed. These terms are the seasonal indices of respective quarters.

Merits and Demerits

This is a simple method of measuring seasonal variations which is based on the unrealistic assumption that the trend and cyclical variations are absent from the data. However, you will see later that this method, being a part of the other methods of measuring seasonal variations, is very useful.

7.5.3 Ratio to Trend Method

This method is used when cyclical variations are absent from the data, i.e., the time series variable Y consists of trend, seasonal and random components.

Using symbols, you can write Y = T.S.R.

Various steps in the computation of seasonal indices are:

(i) Obtain the trend values for each month or quarter, etc., by the method of least squares.

(ii) Divide the original values by the corresponding trend values. This would eliminate trend values from the data. To get figures in percentages, the quotients are multiplied by 100.

Thus, we have $\frac{Y}{T} \times 100 = \frac{T.S.R.}{T} \times 100 = S.R.100$

(iii) Finally, the random component is eliminated by the method of simple averages.

Example: Assuming that the trend is linear, calculate seasonal indices by the ratio to moving average method from the following data:

Quarterly Output of Coal in 4 Years (in thousand tonnes)

Years	I	II	III	IV
2010	65	58	56	61
2011	68	63	63	67
2012	70	59	56	52
2013	60	55	51	58

Solution:

By adding the values of all the quarters of a year, you can obtain annual output for each of the four years. Fit a linear trend to the data and obtain trend values for each quarter.

Years	Output	$X = 2(t - 1983.5)$	XY	X^2
2010	240	−3	−720	9
2011	261	−1	−261	1
2012	237	1	237	1
2013	224	3	672	9
Total	**962**	**0**	**-72**	**20**

From the above table, we get $a = \frac{962}{4} = 240.5$ and $b = \frac{-72}{20} = -3.6$

Thus, the trend line is Y = 240.5 – 3.6X,

Origin : 1st January 2012, unit of X : 6 months.

The quarterly trend equation is given by $Y = \frac{240.5}{4} - \frac{3.6}{8}X$

or Y = 60.13 – 0.45X, Origin : 1st January 2012, unit of X : 1 quarter (i.e., 3 months).

Shifting origin to 15th Feb. 2012, we get

Y = 60.13 – 0.45(X + 1/2) = 59.9 – 0.45X, origin I-quarter, unit of X = 1 quarter.

The table of quarterly values is given by

Years	I	II	III	IV
2010	63.50	63.05	62.60	62.15
2011	61.70	61.25	60.80	60.35
2012	59.90	59.45	59.00	58.55
2013	58.10	57.65	57.20	56.75

The table of Ratio to Trend Values, i.e. $\frac{Y}{T} \times 100$

Years	I	II	III	IV
2010	102.36	91.99	89.46	98.15
2011	110.21	102.86	103.62	111.02
2012	116.86	99.24	94.92	88.81
2013	103.27	95.40	89.16	102.20
Total	432.70	389.49	377.16	400.18
Average	108.18	97.37	94.29	100.05
S.I.	108.20	97.40	94.32	100.08

Note: Grand Average, $G = \frac{399.89}{4} = 99.97$

Merits and Demerits

It is an objective method of measuring seasonal variations. However, it is very complicated and doesn't work if cyclical variations are present.

7.5.4 Ratio to Moving Average Method

The ratio to moving average is the most commonly used method of measuring seasonal variations. This method assumes the presence of all the four components of a time series. Various steps in the computation of seasonal indices are as follows:

1. Compute the moving averages with period equal to the period of seasonal variations. This would eliminate the seasonal component and minimise the effect of random component. The resulting moving averages would consist of trend, cyclical and random components.
2. The original values, for each quarter (or month) are divided by the respective moving average figures and the ratio is expressed as a percentage, i.e., $\frac{Y}{M.A.} = \frac{TCSR}{TCR'} = SR''$, where R' and R'' denote the changed random components.
3. Finally, the random component R'' is eliminated by the method of simple averages.

Example: Given the following quarterly sale figures, in thousand of rupees, for the year 2010-2013, find the specific seasonal indices by the method of moving averages.

Years	I	II	III	IV
2010	34	33	34	37
2011	37	35	37	39
2012	39	37	38	40
2013	42	41	42	44

Solution:

Calculation of Ratio to Moving Averages

Year/Quarter	Sales	4 - Period Moving Total	Centred Total	4 Period M	$\frac{Y}{M}\times 100$
2010 I	34				
II	33 →				
		138 →	279	34.9	97.4
III	34 →				
		141 →	284	35.5	104.2
IV	37 →				
		143 →	289	36.1	102.5
2011 I	37 →				
		146 →	294	36.8	95.1
II	35 →				
		148 →	298	37.3	99.2
III	37 →				
		150 →	302	37.8	103.2
IV	39 →				
		152 →	305	38.1	102.4
2012 I	39 →				
		153 →	307	38.4	96.4
II	37 →				
		154 →	311	38.9	97.7
III	38 →				
		157 →	318	39.8	100.5
IV	40 →				
		161 →	326	40.8	102.9
2013 I	42 →				
		165 →	334	41.8	98.1
II	41 →				
		169			
III	42				
IV	44				

Calculation of Seasonal Indices

Years	I	II	III	IV
2010	–	–	97.4	104.2
2011	102.5	95.1	99.2	103.2
2012	102.4	96.4	97.7	100.5
2013	102.9	98.1	–	–
Total	307.8	289.6	294.3	307.9
A_i	102.6	96.5	98.1	102.6
S.I.	102.7	96.5	98.1	102.7

Note that the Grand Average $G = \frac{399.8}{4} = 99.95$. Also check that the sum of indices is 400.

Merits and Demerits

This method assumes that all the four components of a time series are present and, therefore, widely used for measuring seasonal variations. However, the seasonal variations are not completely eliminated if the cycles of these variations are not of regular nature. Further, some information is always lost at the ends of the time series.

7.5.5 Link Relatives Method

This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern. The link relatives are percentages of the current period (quarter or month) as compared with previous period. With the computation of link relatives and their average, the effect of cyclical and random component is minimised. Further, the trend gets eliminated in the process of adjustment of chained relatives.

The following steps are involved in the computation of seasonal indices by this method:

1. Compute the link relative (L.R.) of each period by dividing the figure of that period with the figure of previous period. For example, link relative of 3rd quarter

$$= \frac{\text{figure of 3rd quarter}}{\text{figure of 2nd quarter}} \times 100$$

2. Obtain the average of link relatives of a given quarter (or month) of various years. A.M. or M_d can be used for this purpose. Theoretically, the later is preferable because the former gives undue importance to extreme items.

3. These averages are converted into chained relatives by assuming the chained relative of the first quarter (or month) equal to 100. The chained relative (C.R.) for the current period (quarter or month)

$$= \frac{C.R. \text{ of the previous period} \times L.R. \text{ of the current period}}{100}$$

4. Compute the C.R. of first quarter (or month) on the basis of the last quarter (or month). This is given by

$$\frac{\text{C.R. of last quarter (or month)} \times \text{average L.R. of 1st quarter (or month)}}{100}$$

This value, in general, be different from 100 due to long term trend in the data. The chained relatives, obtained above, are to be adjusted for the effect of this trend. The adjustment factor is

$$d = \frac{1}{4}[\text{New } C.R. \text{ for 1st quarter} - 100] \text{ for quarterly data}$$

And $d = \frac{1}{12}[\text{New } C.R. \text{ for 1st month} - 100]$ for monthly data

On the assumption that the trend is linear, d, 2d, 3d, etc., is respectively subtracted from the 2nd, 3rd, 4th, etc., quarter (or month).

5. Express the adjusted chained relatives as a percentage of their average to obtain seasonal indices.

6. Make sure that the sum of these indices is 400 for quarterly data and 1200 for monthly data.

Example: Determine the seasonal indices from the following data by the method of link relatives.

Years	1st Qr	2nd Qr	3rd Qr	4th Qr
2009	26	19	15	10
2010	36	29	23	22
2011	40	25	20	15
2012	46	26	20	18
2013	42	28	24	21

Solution:

Calculation Table

Years	I	II	III	IV
2009	-	73.1	78.9	66.7
2010	360.0	80.5	79.3	95.7
2011	181.8	62.5	80.0	75.0
2012	306.7	56.5	76.9	90.0
2013	233.3	66.7	85.7	87.5
Total	1081.8	339.3	400.8	414.9
Mean	270.5	67.9	80.2	83.0
C.R.	100.0	67.9	54.5	45.2
C.R.(adjusted)	100.0	62.3	43.3	28.4
S.I.	170.9	106.5	74.0	48.6

The chained relative (C.R.) of the 1st quarter on the basis of C.R. of the 4th quarter $= \frac{270.5 \times 45.2}{100} = 122.3$

The trend adjustment factor $d = \frac{1}{4}(122.3 - 100) = 5.6$

Thus, the adjusted C.R. of 1st quarter = 100

and for 2nd = 67.9 – 1× 5.6 = 62.3

for 3rd = 54.5 – 2 × 5.6 = 43.3

for 4th = 45.2 – 3 × 5.6 = 28.4

The grand average of adjusted C.R., $G = \frac{100 + 62.3 + 43.3 + 28.4}{4} = 58.5$

Merits and Demerits

This method is less complicated than the ratio to moving average and the ratio to trend methods. However, this method is based upon the assumption of a linear trend which may not always hold true.

7.6 RANDOM OR IRREGULAR VARIATIONS

As the name suggests, these variations do not reveal any regular pattern of movements. These variations are sometimes called residual or random components. Random variations are that component of a time series which cannot be explained in terms of any of the components discussed so far. This component is obtained as a residue after the elimination of trend, seasonal and cyclical components and hence is often termed as residual component.

These movements are exceedingly difficult to dissociate quantitatively from the business cycle. These variations, though accidental in nature, can cause a continual change in the trends, seasonal and cyclical oscillations during the forthcoming period. Floods, fires, earthquakes, revolutions, epidemics, strikes, etc., are the root causes of such irregularities.

Many analysts prefer to subdivide the irregular variation into episodic and residual variations. Episodic fluctuations are unpredictable, but they can be identified. The initial impact on the economy of a major strike or a war can be identified, but a strike or war cannot be predicted. After the episodic fluctuations have been removed, the

remaining variation is called the residual variation. The residual fluctuations, often called chance fluctuations, are unpredictable, and they cannot be identified. Of course, neither episodic nor residual variation can be projected into the future. The common denominator of every random factor it that does not come about as a result of the ordinary operation of the business system and does not recur in any meaningful manner.

7.7 DECOMPOSITION OF TIME SERIES

Time series decomposition involves thinking of a series as a combination of level, trend, seasonality and noise components. Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting.

Decomposition is primarily used for time series analysis, and as an analysis tool it can be used to inform forecasting models on your problem. It provides a structured way of thinking about a time series forecasting problem, both generally in terms of modeling complexity and specifically in terms of how to best capture each of these components in a given model.

Each of these components are something you may need to think about and address during data preparation, model selection and model tuning. You may address it explicitly in terms of modeling the trend and subtracting it from your data, or implicitly by providing enough history for an algorithm to model a trend if it may exist. You may or may not be able to cleanly or perfectly break down your specific time series as an additive or multiplicative model.

Real-world problems are messy and noisy. There may be additive and multiplicative components. There may be an increasing trend followed by a decreasing trend. There may be non-repeating cycles mixed in with the repeating seasonality components.

Time series decomposition involves separating a time series into several distinct components. There are three components that are typically of interest:

1. T_t, a deterministic, non-seasonal secular trend component. This component is sometimes restricted to being a linear trend, though higher-degree polynomials are also used.
2. S_t, a deterministic seasonal component with known periodicity. This component captures level shifts that repeat systematically within the same period (e.g., month or quarter) between successive years. It is often considered to be a nuisance component, and seasonal adjustment is a process for eliminating it.
3. I_t, a stochastic irregular component. This component is not necessarily a white noise process. It can exhibit autocorrelation and cycles of unpredictable duration. For this reason, it is often thought to contain information about the business cycle, and is usually the most interesting component.

There are three functional forms that are most often used for representing a time series y_t as a function of its trend, seasonal and irregular components:

Additive decomposition, where

$$y_t = T_t + S_t + I_t$$

This is the classical decomposition. It is appropriate when there is no exponential growth in the series, and the amplitude of the seasonal component remains constant over time. For identifiability from the trend component, the seasonal and irregular components are assumed to fluctuate around zero.

Multiplicative decomposition, where

$$Y_t = T_t S_t I_t$$

This decomposition is appropriate when there is exponential growth in the series, and the amplitude of the seasonal component grows with the level of the series. For identifiability from the trend component, the seasonal and irregular components are assumed to fluctuate around one.

Log-additive decomposition, where

$$\log y_t = T_t + S_t + I_t$$

This is an alternative to the multiplicative decomposition. If the original series has a multiplicative decomposition, then the logged series has an additive decomposition. Using the logs can be preferable when the time series contains many small observations. For identifiability from the trend component, the seasonal and irregular components are assumed to fluctuate around zero.

You can estimate the trend and seasonal components by using filters (moving averages) or parametric regression models. Given estimates $\hat{T}_t$ and $\hat{S}_t$, the irregular component is estimated as

$$\hat{I}_t = y_t - \hat{T}_t - \hat{S}_t$$

Using the additive decomposition, and

$$\hat{I}_t = \frac{y_t}{\left(\hat{T}_t \hat{S}_t\right)}$$

using the multiplicative decomposition.

The series

$$y_t - \hat{T}_t$$

(or $y_t / \hat{T}_t$ using the multiplicative decomposition) is called a detrended series.

Similarly, the series $y_t - \hat{S}_t \left(\text{or } y_t / \hat{S}_t\right)$ is called a deseasonalized series.

Check Your Progress

Fill in the blanks:

1. A series of observations, on a variable, recorded after successive intervals of time is called _____.
2. Only _____ trend can be determined by selected points method.
3. The time series data are divided into equal parts and the _____ of the values of each part is computed.
4. The moving average of a group is always shown at the _____ of its period.
5. _____ method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern.
6. The oscillatory movements are termed as _____ if their period of oscillation is equal to one year.

7.8 LET US SUM UP

- A series of observations, on a variable, recorded after successive intervals of time is called a time series.
- The data on the population of a nation is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.
- The analysis of time series implies its decomposition into various factors that affect the value of its variable in a given period.
- It is a quantitative and objective evaluation of the effects of various factors on the activity under consideration.
- Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time.
- Trend values of two or more time series can be used for their comparison.
- Oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation.
- The main objective of measuring seasonal variations is to eliminate the effect of seasonal variations from the data.
- Random variations are usually short-term variations but sometimes their effect may be so intense that the value of trend may get permanently affected.

7.9 UNIT END ACTIVITY

With which component of time series would you associate each of the following statement and why?

(i) An era of prosperity

(ii) Heavy sales on the occasion of Deepawali

(iii) Constantly rising demand for sugar in India

(iv) Price hike in petroleum products due to Iran-Iraq war.

7.10 KEYWORDS

Time Series: A series of observations, on a variable, recorded after successive intervals of time is called a time series.

Trend Analysis: Trend is a long term movement in a time series. It is the underlying direction (an upward or downward tendency) and rate of change in a time series, when allowance has been made for the other components.

Least Squares: It is used to approximately solve over determined systems, i.e. systems of equations in which there are more equations than unknowns.

Graphical Method: Technique used to find graphically the breakeven point and highlight the cost-volume-profit relationships over a wide range of activity. The graphical method requires preparation of a break-even chart.

Seasonal Variations: Seasonal Variation is a component of a time series which is defined as the repetitive and predictable movement around the trend line in one year or less.

Cyclical Variations: A time series, where the time interval between successive observations is less than or equal to one year, may have the effects of both the seasonal and cyclical variations. However, the seasonal variations are absent if the time interval between successive observations is greater than one year.

7.11 QUESTIONS FOR DISCUSSION

1. What is a time series?
2. What are its main components?
3. Explain the objectives of time series analysis.
4. What are the various variations in time series?
5. State the methods of measurement of trend.
6. How would you study the seasonal variations in any time series?
7. Distinguish between secular trend and periodic variations.
8. State the various components of a time series.
9. Give some of the examples of time series analysis.
10. Write some of the application areas where time series analysis is used.
11. Under what situations time series arises?
12. What is meant by smoothing?
13. What are the kinds of smoothing in time series analysis do you know?
14. What are cyclical variations?
15. What are irregular variations?
16. "All periodic variations are not necessarily seasonal". Comment.
17. "A time series is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals". Comment.
18. How would you measure trend in a time series data by the method of least squares? Explain your answer with an example.
19. Explain the method of moving average for the determination of trend in a time series data. What are its merits and demerits?
20. Discuss the underlying assumptions of additive and multiplicative models in a time series analysis. Which of these is more popular in practice and why?
21. Distinguish between the ratio to trend and the ratio to moving average methods of measuring seasonal variations. Which method is more general and why?
22. Fit a straight line trend to the following data on steel production (in M. tonnes). Predict the value for 2014.

Years	:	2007	2008	2009	2010	2011	2012	2013
Production	:	80	84	90	93	98	100	104

23. Fit a straight line trend by method of least squares to the following data on earnings (₹ lakh) of a firm. (a) Assuming that the same trend continues, what would be the predicted earnings for the year 2014?

(b) Convert this equation into a monthly equation with January, 2012 as origin and estimate the values for November, 2011 and April, 2012.

Years	:	2005	2006	2007	2008	2009	2010	2011	2012
Earnings	:	38	40	65	72	79	60	87	95

24. Given below are the figures of production of a sugar factory in '000 tonnes

Years	:	2007	2008	2009	2010	2011	2012	2013
Production	:	77	88	94	85	91	98	90

(i) Fit a straight-line trend by method of least squares.

(ii) Calculate trend values and plot observed values and trend values on a graph.

25. Compute the trend line by the method of least squares from the data on profits (in ₹ '000) of a firm, given below:

Years	:	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Profits	:	110	125	115	135	150	165	155	175	180	200

26. Fit a linear trend to the following data, on average monthly output, with origin at mid-point of the year 2009. Convert this into a monthly trend equation. Estimate the average output for June and August, 2009.

Years	:	2005	2006	2007	2008	2009	2010	2011	2012	2013
Output	:	6.3	7.4	9.3	7.4	8.3	10.6	9.0	8.7	7.9

27. Draw a free hand curve showing trend of the following data:

Years	Output (in tonnes)	Years	Output (in tonnes)
2003	115	2008	120
2004	120	2009	130
2005	123	2010	138
2006	125	2011	145
2007	118	2012	150

28. What are the three components of time series decomposition?

Check Your Progress: Model Answer

1. Time series
2. Linear
3. Arithmetic mean
4. Centre
5. Link Relatives
6. Seasonal Variations

7.12 REFERENCE & SUGGESTED READINGS

Levin, R. I. and Rubin, D. S., *Statistics for Management*, Prentice Hall Inc.

Goon, Gupta and Dasgupta, *Fundamentals of Statistics*, Vol. I & II, World Press Pvt. Ltd.

Mathai and Rathie, *Probability and Statistics*, Macmillan.

Arora, P. N., Arora, S. and Arora, S., *Comprehensive Statistical Methods*, S. Chand.

Weiss, *Introductory Statistics*, Pearson Education.

Doane, D. P. and Seward, L. E., *Applied Statistics in Business and Economics*, Tata McGraw Hill.

Kothari. C. R., *Quantitative Techniques*, Vikas Publishing House.

BLOCK - 4

UNIT - VIII

CORRELATION

CONTENTS

8.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Discuss the meaning and definitions of correlation
- Describe the scope and method of correlation analysis

- Explain the various types of correlation
- Analyze the practical application of correlation

8.1 INTRODUCTION

We often encounter the situations, where data appears as pairs of figures relating to two variables, for example, price and demand of commodity, money supply and inflation, industrial growth and GDP, advertising expenditure and market share, etc. A correlation considers the joint variation of two measurements with no distinction as independent and dependent variables. It is the measure of linear relationship between them. In correlation, we do not restrict or set values of any measurement and observe then as they vary to different levels. It only gives indication whether the two variables move together in linearly.

8.2 MEANING AND DEFINITIONS OF CORRELATION

Correlation is a degree of linear association between two random variables. In these two variables, we do not differentiate them as dependent and independent variables. It may be the case that one is the cause and other is an effect i.e. independent and dependent variables respectively. On the other hand, both may be dependent variables on a third variable. In some cases, there may not be any cause-effect relationship at all. Therefore, if we do not consider and study the underlying economic or physical relationship, correlation may sometimes give absurd results. For example, take a case of global average temperature and Indian population. Both are increasing over past 50 years but obviously not related.

Examples of correlation problems are found in the study of the relationship between IQ and aggregate percentage marks obtained in mathematics examination or blood pressure and metabolism. In these examples, both variables are observed as they naturally occur, since neither variable can be fixed at predetermined levels.

Correlation is an analysis of the degree to which two or more variables fluctuate with reference to each other. Correlation is expressed by a coefficient ranging between –1 and +1.

Positive (+ve) sign indicates movement of the variables in the same direction. For example, variation of the fertilizers used on a farm and yield, observes a positive relationship within technological limits. Whereas negative (–ve) coefficient indicates movement of the variables in the opposite directions, i.e. when one variable decreases, other increases. For example, variation of price and demand of a commodity have inverse relationship. Absence of correlation is indicated if the coefficient is close to zero. Value of the coefficient close to +1 denotes a very strong linear relationship.

The study of correlation helps managers in following ways:

1. To identify relationship of various factors and decision variables.
2. To estimate value of one variable for a given value of other if both are correlated. For example, estimating sales for a given advertising and promotion expenditure.
3. To understand economic behaviour and market forces.
4. To reduce uncertainty in decision-making to a large extent.

Various experts have defined correlation in their own words and their definitions, broadly speaking, imply that correlation is the degree of association between two or more variables.

Some important definitions of correlation are given below:

"If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."

– L.R. Connor

"Correlation is an analysis of covariation between two or more variables."

– A.M. Tuttle

"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

– Croxton and Cowden

"Correlation analysis attempts to determine the 'degree of relationship' between variables."

– Ya Lun Chou

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

8.3 SCOPE OF CORRELATION ANALYSIS

The existence of correlation between two (or more) variables only implies that these variables (i) either tend to increase or decrease together or (ii) an increase (or decrease) in one is accompanied by the corresponding decrease (or increase) in the other.

The questions of the type, whether changes in a variable are due to changes in the other, i.e., whether a cause and effect type relationship exists between them, are not answered by the study of correlation analysis.

If there is a correlation between two variables, it may be due to any of the following situations:

- ***One of the variable may be affecting the other:*** A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of tea or vice-versa. In order to know this, you need to have some additional information apart from the study of correlation. For example if, on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.
- ***The two variables may act upon each other:*** Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent. For example, if you have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat. For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.
- ***The two variables may be acted upon by the outside influences:*** In this case, you might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them. For

example, the demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

- ***A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance):*** This is another situation of spurious correlation. Given the data on any two variables, you may obtain a high value of correlation coefficient when in fact they do not have any relationship. For example, a high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality.

In business, correlation analysis often helps manager to take decisions by estimating the effects of changing the values of the decision variables like promotion, advertising, price and production processes, on the objective parameters like costs, sales, market share, consumer satisfaction and competitive price. The decision becomes more objective by removing subjectivity to certain extent. However, it must be understood that the correlation analysis only tells us about the two or more variables in a data fluctuate together or not. It does not necessarily be due cause and effect relationship. To know if the fluctuations in one of the variables indeed affect other or not, one has to be established with logical understanding of the business environment.

Some of the correlations could be completely nonsense relations like increase in jobs in I.T. and reduction production of wheat over past 3 years in India, or share market bull run of 2004 to 2007 and increase in suicides by farmers in India. There are many reasons to get such spurious correlations. Hence, before you use correlation analysis you must check few factors responsible for the apparent relationship. Firstly, the fluctuation may be a chance coincidence. In this case, we could look at the data over different periods and also study if one factor affects the other through third factor that we have not considered. Secondly, even when correlation exists the logical analysis may tell us that one variable is independent and other dependent on it. For example, surface temperature of the Pacific Ocean (Al Niño) affects monsoons in India but monsoons do not affect temperatures of the Pacific Ocean. Thirdly, in some cases, both variables under study may be fluctuating together due to a variation in the third variables.

Both variables under correlation analysis may be dependent variables and hence not mutually correlated.

In such a case, manager cannot vary one of them and expect other variable to vary. For example, correlation in increase in share prices and stronger rupee against dollar may be due to increase in Foreign Direct Investment (FDI). In this case, expecting to control falling share prices through selling dollars by the Reserve Bank is incorrect. To control these two variables you need to control FDI.

If the falling share prices are due to market sentiments or overheated market, controlling FDI may not help.

Thus, the manager needs to analyze the problem in business environment before he/she can apply the correlation analysis in decision-making.

8.4 METHODS OF CALCULATING CORRELATION

Simple linear correlation is a statistical tool applied in many business situations to find the degree to which two variables vary linearly to one another. Although in many situations even if there are more than two variables involved, two of them may be dominant. In such a case, correlation analysis between these two variables helps us to measure the degree of association between these two variables.

For example, demand of a particular product depends on number of factors. However, association of demand with price may be dominant.

Correlation analysis may also be necessary to eliminate a variable which shows low or hardly any correlation with the variable of our interest. In statistics, there are number of measures to describe degree of association between variables.

These are Karl Pearson's Correlation Coefficient, Spearman's rank correlation coefficient, coefficient of determination, Yule's coefficient of association, coefficient of colligation, etc.

There are different methods which help us to find out whether the variables are related or not.

- Scatter Diagram Method
- Karl Pearson's Coefficient of Correlation
- Rank Method
- Concurrent Deviation Method

We shall discuss these methods one by one.

8.5 SCATTER DIAGRAM

Scatter diagram is the most fundamental graph plotted to show relationship between two variables. It is a simple way to represent bivariate distribution. Bivariate distribution is the distribution of two random variables. Two variables are plotted one against each of the X and Y axis. Thus, every data pair of (x_i, y_j) is represented by a point on the graph, x being abscissa and y being the ordinate of the point. From a scatter diagram, we can find if there is any relationship between the x and y, and if yes, what type of relationship. Scatter diagram thus, indicates nature and strength of the correlation.

The pattern of points obtained by plotting the observed points are knows as scatter diagram.

It gives us two types of information.

1. Whether the variables are related or not.
2. If so, what kind of relationship or estimating equation that describes the relationship.

If the dots cluster around a line, the correlation is called linear correlation. If the dots cluster around a curve, the correlation is called a non-linear or curve linear correlation.

Scatter diagram is drawn to visualize the relationship between two variables. The values of more important variable are plotted on the X-axis while the values of the variable are plotted on the Y-axis. On the graph, dots are plotted to represent different pairs of data. When dots are plotted to represent all the pairs, we get a scatter diagram. The way the dots scatter gives an indication of the kind of relationship which exists between the two variables. While drawing scatter diagram, it is not necessary to take at the point of sign the zero values of X and Y variables, but the minimum values of the variables considered may be taken.

When there is a positive correlation between the variables, the dots on the scatter diagram run from left hand bottom to the right hand upper corner. In case of perfect positive correlation, all the dots will lie on a straight line.

When a negative correlation exists between the variables, dots on the scatter diagram run from the upper left hand corner to the bottom right hand corner. In case of perfect negative correlation, all the dots lie on a straight line.

If a scatter diagram is drawn and no path is formed, there is no correlation.

Let the bivariate data be denoted by (X_i, Y_i), where i = 1, 2 ..... n. In order to have some idea about the extent of association between variables X and Y, each pair (X_i, Y_i), i = 1, 2.....n, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points (also known as dots of the diagram) may cluster around a straight line or a curve or may not show any tendency of association. Various possible situations are shown with the help of following diagrams:

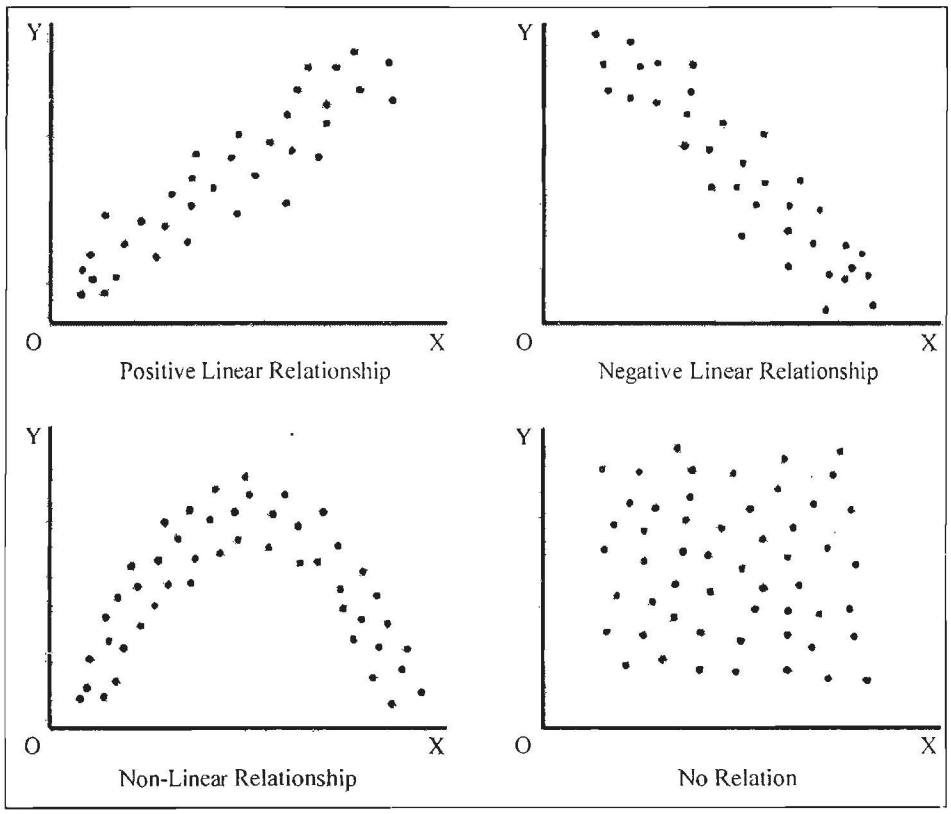


Figure 8.1: Scatter Diagram

If all the points or dots lie exactly on a straight line or a curve, the association between the variables is said to be perfect. This is shown below:

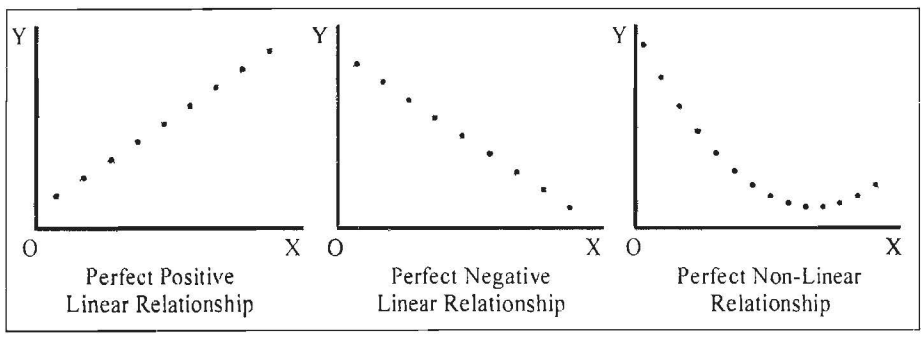


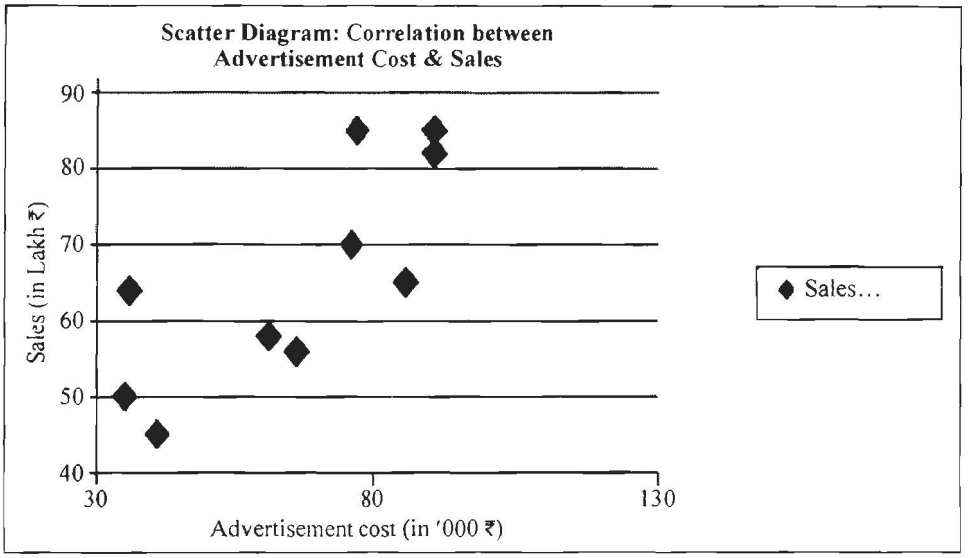
Figure 8.2: Association between the Variables

A scatter diagram of the data helps in having a visual idea about the nature of association between two variables. If the points cluster along a straight line, the association between variables is linear. Further, if the points cluster along a curve, the corresponding association is non-linear or curvilinear. Finally, if the points neither cluster along a straight line nor along a curve, there is absence of any association between the variables. It is also obvious from the above figure that when low (high) values of X are associated with low (high) value of Y, the association between them is said to be positive. Contrary to this, when low (high) values of X are associated with high (low) values of Y, the association between them is said to be negative. This lesson deals only with linear association between the two variables X and Y. We shall measure the degree of linear association by the Karl Pearson's formula for the coefficient of linear correlation.

Example: Figures on advertisement expenditure (X) and Sales (Y) of a firm for the last ten years are given below. Draw a scatter diagram.

Advertisement cost (in '000 ₹)	40	65	60	90	85	75	35	90	34	76
Sales (in Lakh ₹)	45	56	58	82	65	70	64	85	50	85

Solution:



A scatter diagram gives two very useful types of information. First, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get idea of what kind of relationship (linear or non-linear) would describe the relationship. Correlation examines the first question of determining whether an association exists between the two variables, and if it does, to what extent. Regression examines the second question of establishing an appropriate relation between the variables.

8.6 CO-VARIANCE METHOD – THE KARL PEARSON'S CORRELATION COEFFICIENT

The correlation coefficient measures the degree of association between two variables X and Y.

Let us assume, again, that we have data on two variables X and Y denoted by the pairs (X_i, Y_i), i = 1, 2, ... n. Further, let the scatter diagram of the data be as shown in Figure 8.3.

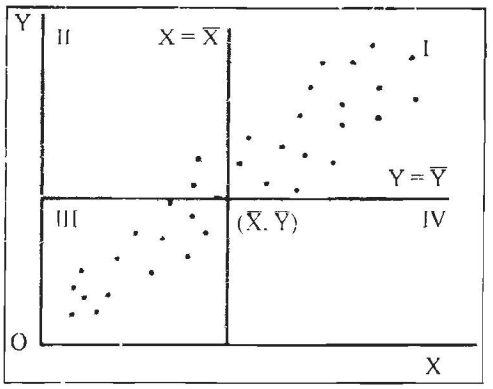


Figure 8.3: Scatter Diagram of the Data

Let $\bar{X}$ and Y be the arithmetic means of X and Y respectively. Draw two lines X = $\bar{X}$ and Y = $\bar{Y}$ on the scatter diagram. These two lines, intersect at the point ($\bar{X}$, $\bar{Y}$) and are mutually perpendicular, divide the whole diagram into four parts, termed as I, II, III and IV quadrants, as shown.

As mentioned earlier, the correlation between X and Y will be positive if low (high) values of X are associated with low (high) values of Y. In terms of the above figure, we can say that when values of X that are greater (less) than $\bar{X}$ are generally associated with values of Y that are greater (less) than $\bar{Y}$, the correlation between X and Y will be positive. This implies that there will be a general tendency of points to concentrate in I and III quadrants. Similarly, when correlation between X and Y is negative, the point of the scatter diagram will have a general tendency to concentrate in II and IV quadrants.

Further, if we consider deviations of values from their means, i.e., $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$, you may note that:

1. Both $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ will be positive for all points in quadrant I.
2. $(X_i - \bar{X})$ will be negative and $(Y_i - \bar{Y})$ will be positive for all points in quadrant II.
3. Both $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ will be negative for all points in quadrant III.
4. $(X_i - \bar{X})$ will be positive and $(Y_i - \bar{Y})$ will be negative for all points in quadrant IV.

It is obvious from the above that the product of deviations, i.e., $(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive for points in quadrants I and III and negative for points in quadrants II and IV.

Since, for positive correlation, the points will tend to concentrate more in I and III quadrants than in II and IV. the sum of positive products of deviations will outweigh the sum of negative products of deviations. Thus, $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive for all the n observations.

Similarly, when correlation is negative, the points will tend to concentrate more in II and IV quadrants than in I and III. Thus, the sum of negative products of deviations will outweigh the sum of positive products and hence $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ will be negative for all the n observations.

Further, if there is no correlation, the sum of positive products of deviations will be equal to the sum of negative products of deviations such that $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ will be equal to zero.

On the basis of the above, we can consider $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ as an absolute measure of correlation. This measure, like other absolute measures of dispersion, skewness, etc., will depend upon (i) the number of observations and (ii) the units of measurements of the variables.

In order to avoid its dependence on the number of observations, we take its average, i.e., $\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$. This term is called covariance in statistics and is denoted as Cov(*X*,*Y*).

To eliminate the effect of units of measurement of the variables, the covariance term is divided by the product of the standard deviation of X and the standard deviation of Y. The resulting expression is known as the Karl Pearson's coefficient of linear correlation or the product moment correlation coefficient or simply the coefficient of correlation, between X and Y.

Karl Pearson's formula for correlation coefficient is given as,

$$r = \frac{Covx.\text{cov}\, y}{\sigma_X \sigma_Y}$$

$$r = \frac{\frac{1}{n}\sum(X - \bar{X})(Y - \bar{Y})}{\sigma_X \sigma_Y} \quad \ldots (1)$$

or $$r_{XY} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum(Y_i - \bar{Y})^2}} \quad \ldots (2)$$

Cancelling $\frac{1}{n}$ from the numerator and the denominator, we get

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \quad \ldots (3)$$

Consider $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i - \bar{X})Y_i - \bar{Y}\sum(X_i - \bar{X})$

$= \sum X_i Y_i - \bar{X}\sum Y_i$ (second term is zero)

$= \sum X_i Y_i - n\bar{X}\bar{Y} \quad \left(\sum Y_i = n\bar{Y}\right)$

Similarly, we can write $\sum(X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$

and $\sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$

Substituting these values in equation (3), we have

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left[\sum X_i^2 - n\bar{X}^2\right]}\sqrt{\left[\sum Y_i^2 - n\bar{Y}^2\right]}} \quad \ldots (4)$$

$$r_{XY} = \frac{\sum X_i Y_i - n \cdot \frac{\sum X_i}{n} \times \frac{\sum Y_i}{n}}{\sqrt{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2}\sqrt{\sum Y_i^2 - n\left(\frac{\sum Y_i}{n}\right)^2}}$$

$$= \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sqrt{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}\sqrt{\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}}} \quad \ldots (5)$$

On multiplication of numerator and denominator by n, we can write

$$r_{XY} = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}} \quad \ldots (6)$$

Further, if we assume $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$, equation (2), given above, can be written as

$$r_{XY} = \frac{\frac{1}{n}\sum x_i y_i}{\sqrt{\frac{1}{n}\sum x_i^2}\sqrt{\frac{1}{n}\sum y_i^2}} \quad \ldots (7)$$

or $$r_{XY} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \quad \ldots (8)$$

or $$r_{XY} = \frac{1}{n}\frac{\sum x_i y_i}{\sigma_x \sigma_y} \quad \ldots (9)$$

Equations (5) or (6) are often used for the calculation of correlation from raw data, while the use of the remaining equations depends upon the forms in which the data are available. For example, if standard deviations of X and Y are given, equation (9) may be appropriate.

Example: Calculate the Karl Pearson's coefficient of correlation from the following pairs of values:

Values of X:	12	9	8	10	11	13	7
Values of Y:	14	8	6	9	11	12	3

Solution:

The formula for Karl Pearson's coefficient of correlation is

$$r_{XY} = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}}$$

The values of different terms, given in the formula, are calculated from the following table:

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
12	14	168	144	196
9	8	72	81	64
8	6	48	64	36
10	9	90	100	81
11	11	121	121	121
13	12	156	169	144
7	3	21	49	9
70	**63**	**676**	**728**	**651**

Here n = 7 (no. of pairs of observations)

$$r_{XY} = \frac{7\times 676 - 70\times 63}{\sqrt{7\times 728-(70)^2}\sqrt{7\times 651-(63)^2}} = 0.949$$

Example: Calculate the Karl Pearson's coefficient of correlation between X and Y from the following data:

No. of pairs of observations n = 8,

$$\sum(X_i-\bar{X})^2 = 184,\ \sum(Y_i-\bar{Y})^2 = 148,$$

$$\sum(X_i-\bar{X})(Y_i-\bar{Y}) = 164,$$

$$\bar{X} = 11 \text{ and } \bar{Y} = 10$$

Solution:

Using the formula, $r_{XY} = \dfrac{\sum(X-\bar{X}_i)(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\sqrt{\sum(Y_i-\bar{Y})^2}}$ we get

$$r_{XY} = \frac{164}{\sqrt{184}\sqrt{148}} = 0.99$$

Example:

1. The covariance between the length and weight of five items is 6 and their standard deviations are 2.45 and 2.61 respectively. Find the coefficient of correlation between length and weight.
2. The Karl Pearson's coefficient of correlation and covariance between two variables X and Y is – 0.85 and – 15 respectively. If variance of Y is 9, find the standard deviation of X.

Solution:

1. Substituting the given values in formula (1) for correlation, we get

$$r_{XY} = \frac{6}{2.45\times 2.61} = 0.94$$

2. Substituting the given values in the formula of correlation, we get

$$-0.85 = \frac{-15}{\sigma_X \times 3} \text{ or } \sigma_X = 5.88$$

8.6.1 Properties of Coefficient of Correlation

1. The coefficient of correlation is independent of the change of origin and scale of measurements.

   In order to prove this property, we change origin and scale of both the variables X and Y.

   Let $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$, where the constants A and B refer to change of origin and the constants h and k refer to change of scale. We can write

   $$X_i = A + hu_i, \quad \therefore \quad \bar{X} = A + h\bar{u}$$

   Thus, you have $X_i - \bar{X} = A + hu_i - A - h\bar{u} = h(u_i - \bar{u})$

   Similarly, $Y_i = B + kv_i, \quad \therefore \quad \bar{Y} = B + k\bar{v}$

   Thus, $Y_i - \bar{Y} = B + kv_i - B - k\bar{v} = k(v_i - \bar{v})$

   The formula for the coefficient of correlation between X and Y is

   $$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2}\sqrt{\sum (Y_i - \bar{Y})^2}}$$

   Substituting the values of $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$, we get

   $$r_{XY} = \frac{\sum h(u_i - \bar{u})k(v_i - \bar{v})}{\sqrt{\sum h^2 (u_i - \bar{u})^2}\sqrt{\sum k^2 (v_i - \bar{v})^2}} = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum (u_i - \bar{u})^2}\sqrt{\sum (v_i - \bar{v})^2}}$$

   $\therefore \quad r_{XY} = r_{uv}$

   This shows that correlation between X and Y is equal to correlation between u and v, where u and v are the variables obtained by change of origin and scale of the variables X and Y respectively.

   This property is very useful in the simplification of computations of correlation. On the basis of this property, you can write a short-cut formula for the computation of r_{XY}:

   $$r_{XY} = \frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}} \quad \ldots (10)$$

2. The coefficient of correlation lies between –1 and + 1.

   To prove this property, you may define

   $$x'_i = \frac{X_i - \bar{X}}{\sigma_X} \quad \text{and} \quad y'_i = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

$$\therefore \quad x_i'^2 = \frac{(X_i - \bar{X})^2}{\sigma_X^2} \text{ and } y_i'^2 = \frac{(Y_i - \bar{Y})^2}{\sigma_Y^2}$$

$$\text{or} \quad \sum x_i'^2 = \frac{\sum (X_i - \bar{X})^2}{\sigma_X^2} \text{ and } \sum y_i'^2 = \frac{\sum (Y_i - \bar{Y})^2}{\sigma_Y^2}$$

From these summations, we can write

$$\text{Also, } r = \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X \sigma_Y} = \frac{1}{n} \cdot \sum \left(\frac{X_i - \bar{X}}{\sigma_X}\right)\left(\frac{Y_i - Y}{\sigma_Y}\right) = \frac{1}{n}\sum x_i' y_i'$$

Consider the sum $x_i' + y_i'$. The square of this sum is always a non-negative number, i.e., $(x_i' + y_i')^2 \geq 0$.

Taking sum over all the observations and dividing by n, we get

$$\frac{1}{n}\sum (x_i' + y_i')^2 \geq 0 \quad \text{or} \quad \frac{1}{n}\sum (x_i'^2 + y_i'^2 + 2x_i' y_i') \geq 0$$

$$\text{or} \quad \frac{1}{n}\sum x_i'^2 + \frac{1}{n}\sum y_i'^2 + \frac{2}{n}\sum x_i' y_i' \geq 0$$

$$\text{or} \quad 1 + 1 + 2r \geq 0 \text{ or } 2 + 2r \geq 0 \text{ or } r \geq -1 \qquad \ldots(11)$$

Further, consider the difference $x_i' - y_i'$. The square of this difference is also non-negative, i.e., $(x_i' - y_i')^2 \geq 0$.

Taking sum over all the observations and dividing by n, we get

$$\frac{1}{n}\sum (x_i' - y_i')^2 \geq 0 \quad \text{or} \quad \frac{1}{n}\sum (x_i'^2 + y_i'^2 - 2x_i' y_i') \geq 0$$

$$\text{or} \quad \frac{1}{n}\sum x_i'^2 + \frac{1}{n}\sum y_i'^2 - \frac{2}{n}\sum x_i' y_i' \geq 0$$

$$\text{or} \quad 1 + 1 - 2r \geq 0 \text{ or } 2 - 2r \geq 0 \text{ or } r \leq 1 \qquad \ldots(12)$$

Combining the inequalities (11) and (12), we get $-1 \leq r \leq 1$. Hence r lies between -1 and $+1$.

3. If X and Y are independent they are uncorrelated, but the converse is not true.

If X and Y are independent, it implies that they do not reveal any tendency of simultaneous movement either in same or in opposite directions. The dots of the scatter diagram will be uniformly spread in all the four quadrants. Therefore, $\sum (X_i - \bar{X})(Y_i - \bar{Y})$ or Cov(X, Y) will be equal to zero and hence, $r_{XY} = 0$. Thus, if X and Y are independent, they are uncorrelated.

The converse of this property implies that if $r_{XY} = 0$, then X and Y may not necessarily be independent. To prove this, we consider the following data:

X	1	2	3	4	5	6	7
Y	9	4	1	0	1	4	9

Here $\Sigma X_i = 28$, $\Sigma Y_i = 28$ and $\Sigma X_i Y_i = 112$.

$$Cov(X.Y)=\frac{1}{n}\left[\sum X_iY_i-\frac{(\sum X_i)(\sum Y_i)}{n}\right]=\frac{1}{7}\left[112-\frac{28\times28}{7}\right]=0.$$

Thus, $r_{XY} = 0$

A close examination of the given data would reveal that although $r_{XY} = 0$, but X and Y are not independent. In fact, they are related by the mathematical relation $Y = (X - 4)^2$.

Example: Calculate the Karl Pearson's coefficient of correlation from the following data:

Height of fathers (inches)	:	66	68	69	72	65	59	62	67	61	71
Height of sons (inches)	:	65	64	67	69	64	60	59	68	60	64

Solution:

Note: When there is no common factor, we can take h = k = 1 and define $u_i = X_i - A$ and $v_i = Y_i - B$.

Calculation of r

Height of fathers (X_i)	**Height of sons (Y_i)**	$u_i = X_i - 65$	$v_i = Y_i - 64$	u_iv_i	u_i^2	v_i^2
66	65	1	1	1	1	1
68	64	3	0	0	9	0
69	67	4	3	12	16	9
72	69	7	5	35	49	25
65	64	0	0	0	0	0
59	60	−6	−4	24	36	16
62	59	−3	−5	15	9	25
67	68	2	4	8	4	16
61	60	−4	−4	16	16	16
71	64	6	0	0	36	0
Total		**10**	**0**	**111**	**176**	**108**

Here n = 10. Using formula (10) for correlation, we get

$$=\frac{10\times111-10\times0}{\sqrt{10\times176-(10)^2}\sqrt{10\times108-0^2}}=0.83$$

Example: Calculate the Karl Pearson's coefficient of correlation from the following data:

(i) Sum of deviations of X values = 5

(ii) Sum of deviations of Y values = 4

(iii) Sum of squares of deviations of X values = 40

(iv) Sum of squares of deviations of Y values = 50

(v) Sum of the product of deviations of X and Y values = 32

(vi) No. of pairs of observations = 10

Solution:

Let $u_i = X_i - A$ and $v_i = Y_i - B$ be the deviations of X and Y values. We are given $\Sigma u_i = 5$, $\Sigma v_i = 4$, $\Sigma u_i^2 = 40$, $\Sigma v_i^2 = 50$, $\Sigma u_i v_i = 32$ and n = 10.

Substituting these values in formula (10), we get

$$r_{XY} = \frac{10 \times 32 - 5 \times 4}{\sqrt{10 \times 40 - 5^2}\sqrt{10 \times 50 - 4^2}} = 0.704$$

Example: Given the following, calculate the coefficient of correlation:

(i) Sum of squares of deviations of X values from mean = 136

(ii) Sum of squares of deviations of Y values from mean = 138

(iii) Sum of products of deviations of X and Y values from their means = 122.

Solution:

Using formula (3) for correlation, we get $r = \frac{122}{\sqrt{136}\sqrt{138}} = 0.89$

Example: Calculate the coefficient of correlation between age group and rate of mortality from the following data:

Age group	:	0-20	20-40	40-60	60-80	80-100
Rate of Mortality	:	350	280	540	760	900

Solution:

Since class intervals are given for age, their mid-values shall be used for the calculation of r.

Table for Calculation of r

Age group	M.V. (X)	Rate of Mort.(Y)	$u_i = \frac{X_i - 50}{20}$	$v_i = \frac{Y_i - 540}{10}$	$u_i v_i$	u_i^2	v_i^2
0-20	10	350	−2	−19	38	4	361
20-40	30	280	−1	−26	26	1	676
40-60	50	540	0	0	0	0	0
60-80	70	760	1	22	22	1	484
80-100	90	900	2	36	72	4	1296
Total			**0**	**13**	**158**	**10**	**2817**

Here n = 5. Using the formula (10) for correlation, we get

$$r_{XY} = \frac{5 \times 158 - 0 \times 13}{\sqrt{5 \times 10 - 0^2}\sqrt{5 \times 2817 - 13^2}} = 0.95$$

Example: Deviations from assumed average of the two series are given below:

Deviations, X series: –10, –6, –4, –1, 0, + 2, + 1, + 5, + 7, + 11

Deviations, Y series: –8, –5, + 4, –2, –4, 0, + 2, 0, –2, + 4

Find out Karl Pearson's coefficient of correlation.

Solution:

Here the values of $u_i = X_i - A$ and $v_i = X_i - B$ are given.

Table for Calculation of r

u_i	−10	−6	−4	−1	0	2	1	5	7	11	**5**
v_i	−8	−5	4	−2	−4	0	2	0	−2	4	**−11**
$u_i v_i$	80	30	−16	2	0	0	2	0	−14	44	**128**
u_i^2	100	36	16	1	0	4	1	25	49	121	**353**
v_i^2	64	25	16	4	16	0	4	0	4	16	**149**

Here n = 10.

$$r_{XY} = \frac{10 \times 128 - 5 \times (-11)}{\sqrt{10 \times 353 - 5^2}\sqrt{10 \times 149 - 11^2}} = 0.609$$

Example: From the following table, find the missing values and calculate the coefficient of correlation by Karl Pearson's method:

X	:	6	2	10	4	?
Y	:	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

Solution:

The missing value in X-series = 5 × 6 − (6 + 2 + 10 + 4) = 30 − 22 = 8

The missing value in Y-series = 5 × 8 − (9 + 11 + 8 + 7) = 40 − 35 = 5

Table for Calculation of r

X	Y	$X - \bar{X}$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
6	9	0	1	0	0	1
2	11	−4	3	−12	16	9
10	5	4	−3	−12	16	9
4	8	−2	0	0	4	0
8	7	2	−1	−2	4	1
Total				**−26**	**40**	**20**

Using formula (3) for correlation, we get $r = \frac{-26}{\sqrt{40}\sqrt{20}} = -0.92$

Example: Calculate Karl Pearson's coefficient of correlation for the following series:

Price (in ₹)	:	10	11	12	13	14	15	16	17	18	19
Demand (in kgs)	:	420	410	400	310	280	260	240	210	210	200

Solution:

Table for Calculation of r

Price (X)	Demand (Y)	$u = X - 14$	$v = \frac{Y-310}{10}$	uv	u^2	v^2
10	420	−4	11	−44	16	121
11	410	−3	10	−30	9	100
12	400	−2	9	−18	4	81
13	310	−1	0	0	1	0
14	280	0	−3	0	0	9
15	260	1	−5	−5	1	25
16	240	2	−7	−14	4	49
17	210	3	−10	−30	9	100
18	210	4	−10	−40	16	100
19	200	5	−11	−55	25	121
Total		**5**	**−16**	**−236**	**85**	**706**

$$r = \frac{-10 \times 236 + 5 \times 16}{\sqrt{10 \times 85 - 25}\sqrt{10 \times 706 - 256}} = -0.96$$

Example: A computer while calculating the correlation coefficient between two variables, X and Y, obtained the following results:

$n = 25,\ \Sigma X = 125,\ \Sigma X^2 = 650,\ \Sigma Y = 100,\ \Sigma Y^2 = 460,\ \Sigma XY = 508.$

It was, however, discovered later at the time of checking that it had copied down two pairs of observations as $\begin{array}{c|c} X & Y \\ \hline 6 & 14 \\ 8 & 6 \end{array}$ in place of the correct pairs $\begin{array}{c|c} X & Y \\ \hline 8 & 12 \\ 6 & 8 \end{array}$. Obtain the correct value of r.

Solution:

First we have to correct the values of $\Sigma X, \Sigma X^2$ .....etc.

Corrected = 125 – (6 + 8) + (8 + 6) = 125

Corrected ^{2} = 650 – (36 + 64) + (64 + 36) = 650

Corrected = 100 – (14 + 6) + (12 + 8) = 100

Corrected ^{2} = 460 – (196 + 36) + (144 + 64) = 436

Corrected = 508 – (84 + 48) + (96 + 48) = 520

Corrected ΣXY = 508 – (84 + 48) + (96 + 48) = 520

$$r = \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}} = 0.67$$

8.6.2 Merits and Limitations of Coefficient of Correlation

The only merit of Karl Pearson's coefficient of correlation is that it is the most popular method for expressing the degree and direction of linear association between the two

variables in terms of a pure number, independent of units of the variables. This measure, however, suffers from certain limitations, given below:

- Coefficient of correlation r does not give any idea about the existence of cause and effect relationship between the variables. It is possible that a high value of r is obtained although none of them seem to be directly affecting the other. Hence, any interpretation of r should be done very carefully.
- It is only a measure of the degree of linear relationship between two variables. If the relationship is not linear, the calculation of r does not have any meaning.
- Its value is unduly affected by extreme items.
- If the data are not uniformly spread in the relevant quadrants, the value of r may give a misleading interpretation of the degree of relationship between the two variables. For example, if there are some values having concentration around a point in first quadrant and there is similar type of concentration in third quadrant, the value of r will be very high although there may be no linear relation between the variables.

8.7 AUTOCORRELATION

The concept of the autocorrelation is similar to the correlation. However, instead of finding correlation between two variables, you may check the correlation between the values of a same variable at different time lag. For example, let us compare i.e. observation at time t with i.e. observation at time $(t + 1)$. The variable is described as time lagged by one period or 'one time lag'. Similarly you can have the observations with 2 or 3 time lags. In general, you can compare observations in series with k time lags. Autocorrelation indicates how the values of the same variable at two different time periods relate to each other. This technique is extremely useful in every functional area in business to study seasonal variation of data and primarily for forecast e.g., demand, prices, manpower availability, spare consumption, inventory variation, fund flow in market, etc. particularly if it has any periodic relationship. This method is useful to identify any cyclic relationship in the data. We could identify such relationship by just looking at the scatter diagram. However, it may not provide an accurate estimate. It would just be a qualitative assessment. If the autocorrelation coefficient is positive or negative, it implies cyclic pattern of the period corresponding to the time lag. On the other hand, a near zero autocorrelation indicates the absence of cyclic pattern.

If the data consists of monthly figures, a twelve-month time lag autocorrelation will indicate yearly correlation with month on month relationship. If the twelve-month time lag autocorrelation coefficient is positive in this case, it implies seasonal pattern of twelve months duration. If the data consists of quarterly figures, four time lag autocorrelations will indicate yearly correlation with quarter on quarter relationship. This is very popular in industry for finding out quarter-on-quarter relationship for sales, profits, etc. Similarly, we can also find monthly, weekly, half yearly, etc. patterns.

If the autocorrelation coefficient is near zero, it indicates absence of seasonality or cyclic pattern. Of course for calculating autocorrelation coefficient you need to know the periodicity that we need to test. There is no clear method for determining the period. Hence, in business, it is advisable to first draw a scatter diagram, identify approximate period in the patterns and then calculate the autocorrelation coefficient. If there is a trend in the data, then autocorrelation coefficient with one time lag will have positive value. In case of completely random data, all autocorrelation coefficients for

different time lags will be near zero. A plot of autocorrelation coefficients, for various time lags indicates the underlying time series.

Procedure for the auto correlation is simple. First, you have to write the series of values of the variable. Then you should shift the same data through number of time periods you need to check the correlation. Thus you may construct series with one, two, etc. time lags. In case of two time lags variable construction, the same variable values with two periods apart relate to each other. If you construct the data with time lag of twelve periods apart, it gives year on year correlation i.e. data of a particular month is correlated for different years. If in such a case, auto correlation coefficient is positive, it implies that there is a seasonal pattern of twelve month duration. On the other hand, near zero auto correlation indicates the absence of seasonal pattern. Point to note that we need to guess time lag from the scatter diagram and then construct the data to calculate auto correlation coefficient. Otherwise you need to check autocorrelations at different period.

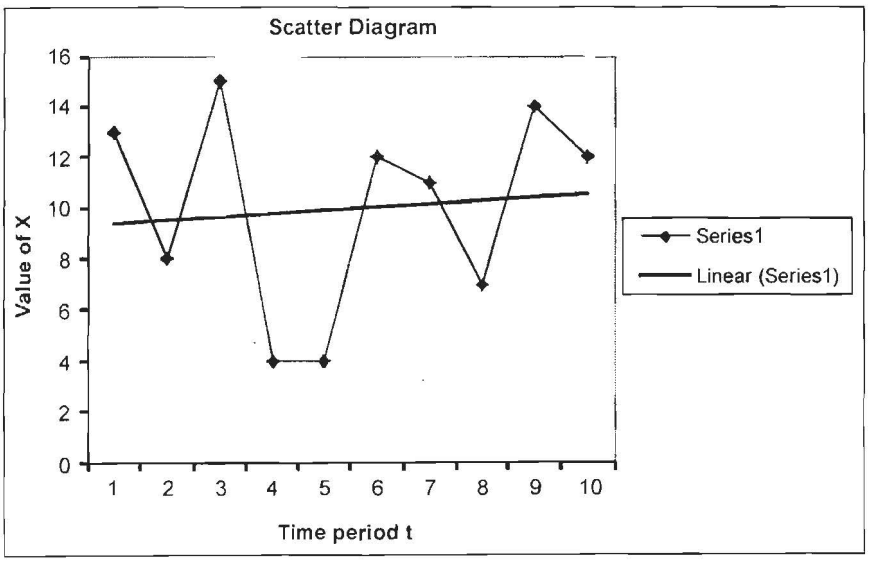
Auto correlation provides very valuable information to a manager regarding the underlying pattern of time series. This helps in identifying seasonality and length of the season.

The example below shows how to construct the auto correlation data for one through five time lag periods.

Example: Plot scatter diagram and construct the data table for one through five time lag periods.

t	1	2	3	4	5	6	7	8	9	10
$X = x_t$	13	8	15	4	4	12	11	7	14	12

Solution:



For scatter diagram, you may plot graph with t on X axis and values of X on Y axis as shown below. The data points have been joined as line diagram to show the pattern. Trend line is also added.

Now we show the table constructing the data for one through five time lag periods.

T	X Original Value	X_1 One time lag	X_2 Two time lag	X_3 Three time lag	X_4 Four time lag	X_5 Five time lag
1	13	8	15	4	4	12
2	8	15	4	4	12	11
3	15	4	4	12	11	7
4	4	4	12	11	7	14
5	4	12	11	7	14	12
6	12	11	7	14	12	-
7	11	7	14	12	-	-
8	7	14	12	-	-	-
9	14	12	-	-	-	-
10	12	-	-	-	-	-

8.7.1 Coefficient of Autocorrelation

Coefficient for autocorrelation at time lag k is given as:

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{X})(x_{t+k} - \bar{X})}{\sum_{t=1}^{n}(x_t - \bar{X})^2}$$

We will demonstrate the process with the example below.

Example: Compute the first five autocorrelations (i.e. up to time lag 5 periods) for the time series given below:

t	1	2	3	4	5	6	7	8	9	10
$X = x_t$	13	8	15	4	4	12	11	7	14	12

Solution:

The calculations are shown below. We have used the data constructed in previous example.

$$\bar{X} = \sum_{i=1}^{n} x_i = \frac{100}{10} = 10$$

t	1	2	3	4	5	6	7	8	9	10	**Total**
x_t	13	8	15	4	4	12	11	7	14	12	100
$x_t - \bar{X}$	3	−2	5	−6	−6	2	1	−3	4	2	
$(x_t - \bar{X})^2$	9	4	25	36	36	4	1	9	16	4	144
x_{t+1}	8	15	4	4	12	11	7	14	12		
$x_{t+1} - \bar{X}$	−2	5	−6	−6	2	1	−3	4	2		
$(x_t - \bar{X})(x_{t+1} - \bar{X})$	−6	−10	−30	36	−12	2	−3	−12	8		−27
x_{t+2}	15	4	4	12	11	7	14	12			
$x_{t+2} - \bar{X}$	5	−6	−6	2	1	−3	4	2			

Contd...

$(x_t - \bar{X})(x_{t+2} - \bar{X})$	15	12	−30	−12	−6	−6	4	−6			−29
x_{t+3}	4	4	12	11	7	14	12				
$x_{t+3} - \bar{X}$	−6	−6	2	1	−3	4	2				
$(x_t - \bar{X})(x_{t+3} - \bar{X})$	−18	12	10	−6	18	8	2				26
x_{t+4}	4	12	11	7	14	12					
$x_{t+4} - \bar{X}$	−6	2	1	−3	4	2					
$(x_t - \bar{X})(x_{t+4} - \bar{X})$	−18	−4	5	18	−24	4					−19
x_{t+5}	12	11	7	14	12						
$x_{t+5} - \bar{X}$	2	1	−3	4	2						
$(x_t - \bar{X})(x_{t+5} - \bar{X})$	6	−2	−15	−24	−12						−47

Using the formula, $r_k = \dfrac{\sum_{t=1}^{n-k}(x_t - \bar{X})(x_{t+k} - \bar{X})}{\sum_{t=1}^{n}(x_t - \bar{X})^2}$

$$r_1 = \frac{-27}{144} = -0.1875$$

$$r_2 = \frac{-29}{144} = -0.2014$$

$$r_3 = \frac{26}{144} = 0.1805$$

$$r_4 = \frac{-19}{144} = -0.1319$$

$$r_5 = \frac{-47}{144} = -0.326$$

It can be seen from the result that there is hardly any cyclic pattern. We can't comment on seasonality since we don't know the periodicity of the data. If it is monthly, in any case, we can't find seasonality since the data is only for 10 periods.

8.8 PRACTICAL APPLICATION OF CORRELATION

The primary purpose of correlation is to establish an association between any two random variables. The presence of association does not imply causation, but the existence of causation certainly implies association. Statistical evidence can only establish the presence or absence of association between variables.

Whether causation exists or not depends merely on reasoning. However, you must be on the guard against spurious or nonsense correlation that may be observed between totally unrelated variables before regression analysis.

Correlation is also used in factor analysis wherein attempts are made to resolve a large set of measured variables in terms of relatively few categories, known as factors. The results could be useful in following three ways:

1. To reveal the underlying or latent factors that determines the relationship between the observed data.

2. To make evident relationship between data that had been obscured before such analysis.
3. To provide a classification scheme when data scored on various rating scales have to be grouped together.

Another major application of correlation is in forecasting with the help of time series models. In past data, one has to identify the trend, seasonality and random pattern in the data before an appropriate forecasting model can be built.

8.9 PEARMAN'S RANK CORRELATION METHOD

This is a crude method of computing correlation between two characteristics. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks. This method is often used in the following circumstances:

1. When the quantitative measurements of the characteristics are not possible, e.g., the results of a beauty contest where various individuals can only be ranked.
2. Even when the characteristic is measurable, it is desirable to avoid such measurements due to shortage of time, money, complexities of calculations due to large data, etc.
3. When the given data consist of some extreme observations, the value of Karl Pearson's coefficient is likely to be unduly affected. In such a situation, the computation of the rank correlation is preferred because it will give less importance to the extreme observations.
4. It is used as a measure of the degree of association in situations where the nature of population, from which data are collected, is not known.

The coefficient of correlation obtained on the basis of ranks is called 'Spearman's Rank Correlation' or simply the 'Rank Correlation'.

Quite often the data is available in the form of some ranking for different variables. Also there are occasions where it is difficult to measure the cause-effect variables. For example, while selecting a candidate, there are number of factors on which the experts base their assessment. It is not possible to measure many of these parameters in physical units e.g. sincerity, loyalty, integrity, tactfulness, initiative, etc. Similar is the case during beauty contests. However, in these cases, the experts may rank the candidates. It is then necessary to find out whether the two sets of ranks are in agreement with each other. This is measured by Rank Correlation Coefficient. The purpose of computing a correlation coefficient in such situations is to determine the extent to which the two sets of ranking are in agreement. The coefficient that is determined from these ranks is known as Spearman's rank coefficient, r_s.

This is defined by the following formula:

$$r_S = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Where, n = Number of observation pairs

$d_i = X_i - Y_i$

X_i = Values of variable X and

Y_i = Values of variable Y

8.9.1 Rank Correlation when Ranks are Given

Example: Ranks obtained by a set of ten students in a mathematics test (variable X) and a physics test (variable Y) are shown below:

Rank for Variable X	1	2	3	4	5	6	7	8	9	10
Rank for Variable Y	3	1	4	2	6	9	8	10	5	7

To determine the coefficient of rank correlation, r_S

Solution:

Computations of Spearman's Rank Correlation as shown below:

Individual	Rank in Maths ($X = x_i$)	Rank in Physics ($Y = y_i$)	$d_i = x_i - y_i$	d_i^2
1	1	3	+2	4
2	2	1	–1	1
3	3	4	+1	1
4	4	2	-2	4
5	5	6	+1	1
6	6	9	+3	9
7	7	8	+1	1
8	8	10	+2	4
9	9	5	–4	16
10	10	7	–3	9
Total				**50**

Now, n = 10, $\sum_{i=1}^{n} d_i^2 = 50$

Using the formula,

$$r_S = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 50}{10(100-1)} = 0.697$$

We can say that there is a high degree of correlation between the performance in mathematics and physics.

8.9.2 Rank Correlation when Ranks are not Given

Example: Find the rank correlation coefficient for the following data:

X :	75	88	95	70	60	80	81	50
Y :	120	134	150	115	110	140	142	100

Solution:

Let R1 and R2 denotes the ranks in X and Y respectively.

X	Y	R1	R2	d = R1 – R2	d^2
75	120	5	5	0	0
88	134	2	4	– 2	4
95	150	1	1	0	0

Contd...

70	115	6	6	0	0
60	110	7	7	0	0
80	140	4	3	1	1
81	142	3	2	1	1
50	100	8	8	0	0
					6

Coefficient of Correlation $P = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 6}{8(64-1)} = +.93$

In this method, the biggest item gets the first rank, the next biggest second rank and so on.

8.9.3 Rank Correlation when Equal Ranks Given

When two or more items have the same rank, a correction has to be applied to $\sum d_i^2$.

For example, if the ranks of X are 1, 2, 3, 3, 5,.... showing that there are two items with the same 3rd rank and fourth rank is skipped, then instead of writing 3, we write 3½ for both. Thus the sum of these ranks which is 7 (3+4= 3½+3½= 7) remains same keeping the mean of ranks unaffected. But in such cases the standard deviation is affected. Therefore, correction is required for the Rank Correlation Coefficient. For this, $\sum d_i^2$ is increased by $\frac{(m^3-m)}{12}$ for each tie, where m is number of items in each tie.

We must remember that if there are more than one group of items with common rank, this correction factor is to be added that many times once for each group.

Example: Twelve salesmen are ranked for efficiency and length of service as below:

Salesman	A	B	C	D	E	F	G	H	I	J	K	L
Efficiency (X)	1	2	3	4	4	4	7	8	9	10	11	12
Length of Service (Y)	2	1	5	3	9	7	7	6	4	11	10	11

Find the value of Spearman's Rank Coefficient.

Solution:

Computations of Spearman's Rank Correlation as shown below:

Individual	**Efficiency ($X = x_i$)**	**Length of Service ($Y = y_i$)**	$d_i = x_i - y_i$	d_i^2
A	1	2	-1	1
B	2	1	1	1
C	3	5	-2	4
D	(4+5+6)/3 = 5	3	2	4
E	(4+5+6)/3 = 5	9	-4	16
F	(4+5+6)/3 = 5	(7+8)/2 = 7.5	-2.5	6.25
G	7	(7+8)/2 = 7.5	-0.5	0.25
H	8	6	2	4
I	9	4	5	25

Contd...

J	10	(11+12)/2 = 11.5	-1.5	2.25
K	11	10	1	1
L	12	(11+12)/2 = 11.5	0.5	0.25
Total				**65**

Now, n = 12, $\sum_{i=1}^{n} d_i^2 = 65$

Using the formula,

$$r_S = 1 - \frac{6 \times \left\{ \sum_{i=1}^{n} d_i^2 + \frac{1}{12} \times (3^3 - 3) + \frac{1}{12} \times (2^3 - 2) + \frac{1}{12} \times (2^3 - 2) \right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times \{65 + 2 + 0.5 + 0.5\}}{12(144 - 1)} = 0.762$$

We can conclude that there is a high degree of correlation between efficiency and length of service.

8.10 CORRELATION COEFFICIENT USING CONCURRENT DEVIATION

This is the easiest method to find the correlation between two variables. Although the method is effective in giving the direction of the correlation as positive or negative but fails to give the accurate strength of the correlation. In this method, we check the fluctuation in each data series as increasing (+), or decreasing (–) or equal values. Then we count the number of items that increase or decrease or remains equal concurrently and denote as c. The correlation coefficient is then calculated as,

$$r = \pm\sqrt{\pm\left(\frac{2 \times c - n}{n}\right)}$$

Where, n = Total number of pairs

c = Number of concurrent changes

Example: The data of advertisement expenditure (X) and sales (Y) of a company for past 10 year period is given below. Determine the correlation coefficient between these variables and comment the correlation.

X	50	50	50	40	30	20	20	15	10	5
Y	700	650	600	500	450	400	300	250	210	200

Solution:

S. No.	X	Deviation Sign	Y	Deviation Sign	Concurrent Deviation
1.	50	.....	700	.....	.....
2.	50	=	650	-	-
3.	50	=	600	-	-
4.	40	-	500	-	+
5.	30	-	450	-	+

Contd...

6.	20	-	400	-	+
7.	20	=	300	-	-
8.	15	-	250	-	+
9.	10	-	210	-	+
10.	5	-	200	-	+
Total $\sum$					**6**

Therefore,

$$r = \pm\sqrt{\pm\left(\frac{2\times c - n}{n}\right)} = +\sqrt{+\left(\frac{2\times 6 - 9}{9}\right)} = 0.577$$

The result indicates that there is positive correlation between advertisement expenditure (X) and sales (Y).

1. Sign ± is selected to make the value of $\left(\frac{2\times c - n}{n}\right)$ positive. The same sign is used outside the radical.
2. This method does not give strength of correlation. The method is ad hoc and used only to reduce the efforts of tedious calculations.

8.11 TYPES OF CORRELATION

Types of correlation that need to be differentiated before using the correlation coefficient for managerial decision-making are given below.

8.11.1 Positive or Negative Correlation

In positive correlation, both factors increase or decrease together. When we say a perfect correlation, the scatter diagram will show a linear (straight line) plot with all points falling on straight line. If you take appropriate scale, the straight line inclination can be adjusted to 45°, although it is not necessary as long as inclination is not 0° or 90° where there is no correlation at all because value of one variable changes without any change in the value of other variable. In case of negative correlation, when one variable increases the other decrease and visa versa. If the scatter diagram shows the points distributed closely around an imaginary line, we say it is high degree of correlation.

If we can hardly see any unique imaginary line around which the observations are scattered, we say correlation does not exist.

Even in case of imaginary line being parallel to one of the axes we say no correlation exists between the variables. If the imaginary line is a straight line we say the correlation is linear.

8.11.2 Simple or Multiple Correlation

In simple correlation, the variation is between only two variables under study and the variation is hardly influenced by any external factor. In other words, if one of the variables remains same, there won't be any change in other variable. For example, variation in sales against price change in case of a price sensitive product under stable market conditions shows a negative correlation. In multiple correlation, more than two variables affect one another.

We need to study correlation between all the pairs that are affecting each other and study extent to which they have the influence.

8.11.3 Partial or Total Correlation

In case of multiple correlation analysis, there are two approaches to study the correlation. In case of partial correlation, we study variation of two variables and excluding the effects of other variables by keeping them under controlled condition. In case of 'total correlation' study, you may allow all relevant variables to vary with respect to each other and find the combined effect. With few variables, it is feasible to study 'total correlation'.

As number of variables increase, it becomes impractical to study the 'total correlation'.

8.11.4 Linear and Non-linear Correlation

The manager must be careful in analyzing the correlation using coefficients because most of the coefficients are based on assumption of linearity. Hence plotting a scatter diagram is good practice.

Scatter diagram not only tell us about linearity or non-linearity but also whether the data is cyclic.

When values of two variables have a constant rate of change it is linear correlation. In such a case, the differential (derivative) of relationship is constant with the graph of the data being a straight line. In case on non-linear correlation, the rate of variation changes as values increase or decrease. The non-linear relationship could be approximated to a polynomial (parabolic, cubic etc.), exponential sinusoidal, etc. In such cases, using the correlation coefficients based on linear assumption will be misleading unless used over a very short data range.

Using computers, we could analyze a non-linear correlation to a certain extent, with some simplified assumption.

Check Your Progress

Fill in the blanks:

1. _____ sign indicates movement of the variables in the same direction.
2. _____ is the most fundamental graph plotted to show relationship between two variables.
3. If a scatter diagram is drawn and no path is formed, there is no _____.
4. Correlation is also used in factor analysis wherein attempts are made to resolve a large set of measured variables in terms of relatively few categories known as _____.
5. The coefficient of correlation obtained on the basis of ranks is called _____.
6. In case of _____ study, you may allow all relevant variables to vary with respect to each other and find the combined effect.

8.12 LET US SUM UP

- When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.

- So far we have considered distributions relating to single characteristics. Such distributions are known as univariate distribution.
- Correlation is a degree of linear association between two random variables. In these two variables, we do not differentiate them as dependent and independent variables. It may be the case that one is the cause and other is an effect i.e. independent and dependent variables respectively. On the other hand, both may be dependent variables on a third variable.
- In business, correlation analysis often helps manager to take decisions by estimating the effects of changing the values of the decision variables like promotion, advertising, price and production processes, on the objective parameters like costs, sales, market share, consumer satisfaction and competitive price. The decision becomes more objective by removing subjectivity to certain extent.
- The correlation coefficient r may assume values between -1 and 1. The sign indicates whether the association is direct (+ve) or inverse (-ve). A numerical value of r equal to unity indicates perfect association while a value of zero indicates no association.
- The correlation is said to be positive when the increase (decrease) in the value of one variable is accompanied by an increase (decrease) in the value of other variable also. Negative or inverse correlation refers to the movement of the variables in opposite direction. Correlation is said to be negative, if an increase (decrease) in the value of one variable is accompanied by a decrease (increase) in the value of other.
- In simple correlation, the variation is between only two variables under study and the variation is hardly influenced by any external factor. In other words, if one of the variables remains same, there won't be any change in other variable.
- In case of multiple correlation analysis, there are two approaches to study the correlation. In case of partial correlation, we study variation of two variables and excluding the effects of other variables by keeping them under controlled condition.
- When the amount of change in one variable tends to keep a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. But if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then the correlation is said to be non-linear.
- Correlation analysis may also be necessary to eliminate a variable which shows low or hardly any correlation with the variable of our interest. In statistics, there are number of measures to describe degree of association between variables. These are Karl Pearson's Correlation Coefficient, Spearman's rank correlation coefficient, coefficient of determination, Yule's coefficient of association, coefficient of colligation, etc.
- The correlation coefficient measures the degree of association between two variables X and Y.
- When various units under consideration are observed simultaneously, with regard to two characteristics, we get a Bivariate Distribution. For example, the simultaneous study of the heights and weights of students of a college.
- For such data also, we can compute mean, variance, skewness, etc. for each individual characteristics.

- In bivariate distribution, we are interested in knowing whether there exists some relationship between two characteristics or in other words, how far the two variables, corresponding to two characteristics, tend to move together in same or opposite directions i.e. how far they are associated.

8.13 UNIT END ACTIVITY

Collect a random sample of 20 stones, for each stone measure its:

(a) Maximum dimension

(b) Minimum dimension

(c) Weight

Does there appear any connection between a) and b), b) and c) or c) and a)?

8.14 KEYWORDS

Correlation: When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.

Correlation Analysis: Correlation analysis attempts to determine the 'degree of relationship' between variables.

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

Dots of the Diagram: Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points is also known as dots of the diagram.

Scatter Diagram: Let the bivariate data be denoted by (X_i, Y_i), where i = 1, 2 ..... n. In order to have some idea about the extent of association between variables X and Y, each pair (X_i, Y_i), i = 1, 2.....n, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Spearman's Rank Correlation: This is a crude method of computing correlation between two characteristics. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks.

Univariate Distribution: Distributions, relating to a single characteristic, are known as univariate distribution.

Exponential Trend: The general form of an exponential trend is Y = a.bt, where a and b are constants.

Least Square Methods: This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, is minimized.

8.15 QUESTIONS FOR DISCUSSION

1. Define correlation between two variables.
2. Distinguish between positive and negative correlation. Illustrate by using diagrams.
3. Write down an expression for the Karl Pearson's coefficient of linear correlation. Why is it termed as the coefficient of linear correlation? Explain.

4. Write short notes on scatter diagram.

5. Compute Karl Pearson's coefficient of correlation from the following data:

X :	8	11	15	10	12	16
Y :	6	9	11	7	9	12

6. Calculate Karl Pearson's coefficient of correlation between the marks obtained by 10 students in economics and statistics.

Roll No. :	1	2	3	4	5	6	7	8	9	10
Marks in eco. :	23	27	28	29	30	31	33	35	36	39
Marks in Stat. :	18	22	23	24	25	26	28	29	30	32

7. Find Karl Pearson's coefficient of correlation from the following data and interpret its value.

Wages (₹) :	100	101	103	102	100	99	97	98	96	95
Cost of Living (₹) :	98	99	99	97	95	92	95	94	90	91

8. Coefficient of rank correlation and the sum of squares of differences in corresponding ranks are 0.9021 and 28 respectively. Determine the number of pairs of observations.

9. What are the advantages of studying correlation?

10. Under what conditions will Spearman's formula and Karl Pearson's formula give equal results?

11. Show that the coefficient of correlation, r, is independent of change of origin and scale.

12. Prove that the coefficient of correlation lies between - 1 and +1.

13. "If two variables are independent the correlation between them is zero, but the converse is not always true". Explain the meaning of this statement.

14. Calculate coefficient of correlation between X and Y as per the data given below:

X	14	16	20	22	28	30	34	40	45
Y	97	89	68	65	56	50	37	18	12

15. Find out the coefficient of correlation from the following data:

X :	300	350	400	450	500	550	600	650	700
Y :	1600	1500	1400	1300	1200	1100	1000	900	800

16. Calculate the coefficient of correlation from the following data and interpret the result.

$\sum XY = 8425, \bar{X} = 28.5, \bar{Y} = 28.0, \sigma_X = 10.5, \sigma_Y = 5.6$, and $n = 10$

17. Draw a scatter diagram of the following data and indicate whether the correlation between the variables is positive or negative.

Height (inches) :	62	72	70	60	67	70	64	65	60	70
Weight (lbs.) :	50	65	63	52	56	60	59	58	54	65

18. Calculate Pearson's coefficient of correlation for the following data:

Price (₹)	:	22	24	26	28	30	32	34	36	38	40
Demand (Tonnes)	:	60	58	58	50	48	48	48	42	36	32

Check Your Progress: Model Answer

1. Positive (+ve)
2. Scatter diagram
3. Correlation
4. Factors
5. Spearman's Rank Correlation or Rank Correlation
6. Total correlation

8.16 REFERENCE & SUGGESTED READINGS

- Keller, G. (2020). **Statistics for Management and Economics** (11th ed.). Cengage Learning. ISBN: 9780357108251
- Render, B., Stair, R. M., Hanna, M. E., & Hale, T. S. (2021). **Quantitative Analysis for Management** (13th ed.). Pearson. ISBN: 9780134543162
- McClave, J. T., Benson, P. G., &Sincich, T. (2018). **Statistics for Business and Economics** (13th ed.). Pearson. ISBN: 9780134506594
- Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). **Statistical Techniques in Business and Economics** (17th ed.). McGraw-Hill Education. ISBN: 9781259666360
- Siegel, A. F. (2019). **Practical Business Statistics** (7th ed.). Academic Press. ISBN: 9780128131350
- Groebner, D. F., Shannon, P. W., Fry, P. C., & Smith, K. D. (2022). **Business Statistics: A Decision-Making Approach** (11th ed.). Pearson. ISBN: 9780136681503

UNIT-IX

REGRESSION ANALYSIS

CONTENTS

9.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Define the term regression
- Describe regression analysis
- Explain lines of regression
- Analyze the estimation of regression line

- Understand the concept of multiple regression
- Explain multiple regression equation with three variables
- Discuss regression equation in terms of correlation coefficients

9.1 INTRODUCTION

As we have seen, correlation analysis indicates whether two variables fluctuate with any relationship or not. Regression provides us a measure of the relationship and also facilitates to predict one variable for a value of other variable. Thus, unlike correlation analysis, in regression analysis, one variable is independent and other dependent. This relationship need not be a cause-effect relationship. This must be checked by applying business knowledge. Thus, the regression analysis only gives a mathematical measure of average relationship between two variables. This is one of the most commonly used (and abused) statistical tool for predictions or forecast of economic and business information for decision-making. Like in correlation, regression analysis can also be studied as 'simple and multiple', 'total and partial', 'linear and non-linear', etc. depending upon the type of data and method we use for regression analysis.

9.2 MEANING AND DEFINITION OF REGRESSION

The word Regression implies 'going back or falling back to mean or average value but in most application of regression, we do not use regression in this sense. We use it for the forecasting purpose or to understand underlying mathematical relationship.

Although correlation and regression both attempt to establish whether relationship exists between two or more variables or not, these two techniques differ in approach. If we only want to know the degree and direction of relationship we use correlation analysis. But if we want to forecast or predict the values we need regression analysis. In correlation, there is no distinction between independent and dependent variables. But for regression analysis, we need to specify independent and dependent variables clearly. In case of correlation, we are only interested in finding whether the relationship exists. Hence, the measuring error is only to establish confidence in our analysis. However, in regression our analysis itself is based on the concept of minimizing the errors.

If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1, 2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation in statistics. Since the coefficient of correlation is measure of the degree of linear association of the variables, we shall discuss only linear regression equation. This does not, however, imply the non-existence of non-linear regression equations.

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. As pointed out earlier, the nature of a regression equation is different from the nature of a mathematical equation, e.g., if $Y = 10 + 2X$ is a mathematical equation then it implies that Y is exactly equal to 20 when $X = 5$. However, if $Y = 10 + 2X$ is a regression equation, then $Y = 20$ is an average value of Y when $X = 5$.

According to Morris Myers Blair, "regression is the measure of the average relationship between two or more variables in terms of the original units of the data."

Regression is any statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other variable (usually denoted by X). The relation between selected values of x and

observed values of y (from which the most probable value of y can be predicted for any value of x) is Regression.

The term regression was first introduced by Sir Francis Galton in 1877.

In his study of the relationship between heights of fathers and sons, he found that tall fathers were likely to have tall sons and vice-versa. However, the mean height of sons of tall fathers was lower than the mean height of their fathers and the mean height of sons of short fathers was higher than the mean height of their fathers. In this way, a tendency of the human race to regress or to return to a normal height was observed. Sir Francis Galton referred this tendency of returning to the mean height of all men as regression in his research paper, "Regression towards mediocrity in hereditary stature".

The term 'Regression', originated in this particular context, is now used in various fields of study, even though there may be no existence of any regressive tendency.

9.2.1 Simple Linear Regression

The most commonly used form of regression is linear regression, and the most common type of linear regression is called ordinary least squares regression. Linear regression uses the values from an existing data set consisting of measurements of the values of two variables, X and Y, to develop a model that is useful for predicting the value of the dependent variable, Y for given values of X.

This model is used if we have bivariate distribution i.e. only two variables are considered and the 'best fit' curve is approximated to a straight line. This describes the liner relationship between two variables. Although it appears to be too simplistic, in many business situations, it is adequate. At least, initial study can be based on this model for any decision-making situation. Then we could either use other models of some Ad hoc methods to cater for the complexity of the business situation. If the system is found to have many non-random components we may have to discard this model and use some other model. This model assumes the errors are purely due to randomness and all non-random fluctuations are captured by our 'best fit' curve. Thus we can use the regression analysis for prediction of dependent variable for a given value of independent variable or for controlling the independent variable to get the desired results or to explain relationship for reliable predictions.

9.3 ELEMENTS OF A REGRESSION EQUATION

The regression equation is written as

$$Y = a + bX + e$$

- Y is the value of the Dependent variable (Y), what is being predicted or explained
- a or Alpha, a constant; equals the value of Y when the value of X=0
- b or Beta, the coefficient of X; the slope of the regression line; how much Y changes for each one-unit change in X.
- X is the value of the Independent variable (X), what is predicting or explaining the value of Y
- e is the error term; the error in predicting the value of Y, given the value of X (it is not displayed in most regression equations).

For example, say we know what the average speed is of cars on the freeway when we have 2 highway patrols deployed (average speed=75 mph) or 10 highway patrols deployed (average speed=35 mph). But what will be the average speed of cars on the freeway when we deploy 5 highway patrols?

Average Speed on Freeway (Y)	Number of Patrol Cars Deployed (X)
75	2
35	10

From our known data, we can use the regression formula (calculations not shown) to compute the values of and obtain the following equation:

Y= 85 + (-5) X, where

Y is the average speed of cars on the freeway.

a=85, or the average speed when X=0.

b=(-5), the impact on Y of each additional patrol car deployed.

X is the number of patrol cars deployed.

That is, the average speed of cars on the freeway when there are no highway patrols working (X=0) will be 85 mph. For each additional highway patrol car working, the average speed will drop by 5 mph. For five patrols (X = 5), Y = 85 + (– 5) (5) = 85 – 25 = 60 mph

There may be some variations on how regression equations are written in the literature. For example, you may sometimes see the dependent variable term (Y) written with a little "hat" (^) on it, or called Y-hat. This refers to the predicted value of Y. The plain Y refers to observed values of Y in the data set used to calculate the regression equation.

You may see the symbols for alpha (a) and beta (b) written in Greek letters, or you may see them written in English letters. The coefficient of the independent variable may have a subscript, as may the term for X, for example, b_1X_1 (this is common in multiple regression).

9.4 APPLICABILITY AND USES OF REGRESSION ANALYSIS

Regression analysis is one of the most popular and commonly used statistical tools in business. With availability of computer packages, it has simplified the use. However, one must be careful before using this tool as it gives only mathematical measure based on available data. It does not check whether the cause effect relationship really exists and if it exists which is dependent and which is dependent variable. Regression analysis helps in the following way:

- It provides mathematical relationship between two or more variables. This mathematical relationship can then be used for further analysis and treatment of information using more complex techniques.
- Since most of the business analysis and decisions are based on cause-effect relationships, regression analysis is highly valuable tool to provide mathematical model for this relationship.
- Most wide use of regression analysis is of course estimation and forecast.
- Regression analysis is also used in establishing the theories based on relationships of various parameters. Some of the common examples are demand and supply, money supply and expenditure, inflation and interest rates, promotion expenditure and sales, productivity and profitability, health of workers and absenteeism, etc.

We need to have statistical model that will extract information from the given data to establish the regression relationship between independent and dependent relationship. The model should capture systematic behaviour of data.

The non-systematic behaviour cannot be captured and called as errors. The error is due to random component that cannot be predicted as well as the component not adequately considered in statistical model. Good statistical model captures the entire systematic component leaving only random errors.

In any model, we attempt to capture everything which is systematic in data. Random errors cannot be captured in any case. Assuming the random errors are 'Normally distributed' we can specify the confidence level and interval of random errors. Thus, our estimates are more reliable.

If the variables in a bivariate distribution are correlated, the points in scatter diagram approximately cluster around some curve. If the curve is straight line we call it as linear regression. Otherwise, it is curvilinear regression. The equation of the curve which is closest to the observations is called the 'best fit'.

The best fit is calculated as per Legender's principle of least sum squares of deviations of the observed data points from the corresponding values on the 'best fit' curve. This is called s minimum squared error criteria. It may be noted that the deviation (error) can be measured in X direction or Y direction. Accordingly we will get two 'best fit' curves. If we measure deviation in Y direction, i.e. for a given value of data point (x_i, y_i), then we measure corresponding y value on 'beast fit' curve and then take the value of deviation in y, we call it as regression of Y on X. In the other case, if we measure deviations in X direction we call it as regression of X and Y.

9.5 ESTIMATION OF REGRESSION LINE

For a bivariate data (X_i, Y_i), i = 1, 2, ... n, you can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of Y on X. If, on the other hand, Y is used for estimating the average values of X, the relation will be called regression of X on Y. For a bivariate data, there will always be two lines of regression. It will be shown later that these two lines are different, i.e., one cannot be derived from the other by mere transfer of terms, because the derivation of each line is dependent on a different set of assumptions.

9.5.1 Line of Regression of Y on X

The general form of the line of regression of Y on X is $Y_{Ci} = a + bX_i$, where Y_{Ci} denotes the average or predicted or calculated value of Y for a given value of $X = X_i$. This line has two constants, a and b. The constant a is defined as the average value of Y when X = 0. Geometrically, it is the intercept of the line on Y-axis. Further, the constant b, gives the average rate of change of Y per unit change in X, is known as the regression coefficient.

The above line is known if the values of a and b are known. These values are estimated from the observed data (X_i, Y_i), i = 1, 2, ... n.

Note: It is important to distinguish between Y_{Ci} and Y_i. Whereas Y_i is the observed value, Y_{Ci} is a value calculated from the regression equation.

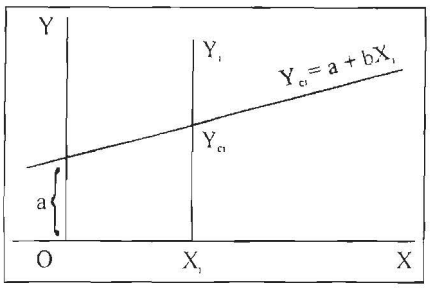


Figure 9.1: Line of Regression Y on X

Using the regression $Y_{Ci} = a + bX_i$, we can obtain Y_{C1}, Y_{C2}, ... Y_{Cn} corresponding to the X values X_1, X_2, ... X_n respectively. The difference between the observed and calculated value for a particular value of X say X_i is called error in estimation of the i^{th} observation on the assumption of a particular line of regression. There will be similar type of errors for all the n observations. We denote by $e_i = Y_i - Y_{Ci}$ (i = 1, 2, ...n), the error in estimation of the i th observation. As is obvious from Figure 9.1, e_i will be positive if the observed point lies above the line and will be negative if the observed point lies below the line. Therefore, in order to obtain a figure of total error, e_i's are squared and added. Let S denote the sum of squares of these errors, i.e., $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - Y_{Ci})^2$.

The regression line can, alternatively, be written as a deviation of Y_i from Y_{ci} i.e. $Y_i - Y_{ci} = e_i$ or $Y_i = Y_{ci} + e_i$ or $Y_i = a + bX_i + e_i$. The component $a + bX_i$ is known as the deterministic component and e_i is random component.

The value of S will be different for different lines of regression. A different line of regression means a different pair of constants a and b. Thus, S is a function of a and b. We want to find such values of a and b so that S is minimum. This method of finding the values of a and b is known as the Method of Least Squares.

Rewrite the above equation as $S = (Y_i - a - bX_i)^2$ ($Y_{Ci} = a + bX_i$).

The necessary conditions for minima of S are

(i) $\frac{\partial S}{\partial a} = 0$ and (ii) $\frac{\partial S}{\partial b} = 0$, where $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial b}$ are the partial derivatives of S w.r.t. a and b respectively.

Now $\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n} (Y_i - a - bX_i) = 0$

or $\sum_{i=1}^{n} (Y_i - a - bX_i) = \sum_{i=1}^{n} Y_i - na - b\sum_{i=1}^{n} X_i = 0$

or $\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i$ ...(1)

Also, $\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n} (Y_i - a - bX_i)(-X_i) = 0$

or $-2\sum_{i=1}^{n} (X_i Y_i - aX_i - bX_i^2) = \sum_{i=1}^{n} (X_i Y_i - aX_i - bX_i^2) = 0$

or $\sum_{i=1}^{n} X_i Y_i - a\sum_{i=1}^{n} X_i - b\sum_{i=1}^{n} X_i^2 = 0$

or $\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2$ ...(2)

Equations (1) and (2) are a system of two simultaneous equations in two unknowns a and b, which can be solved for the values of these unknowns. These equations are also known as normal equations for the estimation of a and b. Substituting these values of a and b in the regression equation $Y_{Ci} = a + bX_i$, we get the estimated line of regression of Y on X.

Expressions for the Estimation of a and b

Dividing both sides of the equation (1) by n, we have

$$\frac{\sum Y_i}{n} = \frac{na}{n} + \frac{b\sum X_i}{n} \quad \text{or} \quad \bar{Y} = a + b\bar{X} \qquad ...(3)$$

This shows that the line of regression $Y_{Ci} = a + bX_i$ passes through the point $(\bar{X}, \bar{Y})$.

From equation (3), we have

$$a = \bar{Y} - b\bar{X} \qquad ...(4)$$

Substituting this value of a in equation (2), we have

$$\sum X_i Y_i = (\bar{Y} - b\bar{X})\sum X_i + b\sum X_i^2$$

$$= \bar{Y}\sum X_i - b\bar{X}\sum X_i + b\sum X_i^2$$

or $\sum X_i Y_i - n\bar{X}\bar{Y} = b\left(\sum X_i^2 - n\bar{X}^2\right)$

or $b = \dfrac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$ ...(5)

Also, $\sum X_i Y_i - n\bar{X}\bar{Y} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$

And $\sum X_i^2 - n\bar{X}^2 = \sum (X_i - \bar{X})^2$

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \qquad ...(6)$$

or $b = \dfrac{\sum x_i y_i}{\sum x_i^2}$ ...(7)

where x_i and y_i are deviations of values from their arithmetic mean.

Dividing numerator and denominator of equation (6) by n, we have

$$b = \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum (X_i - \bar{X})^2} = \frac{Cov(X,Y)}{\sigma_X^2} \qquad ...(8)$$

The expression for b, which is convenient for use in computational work, can be written from equation (5) is given below:

$$b = \frac{\sum X_i Y_i - n\frac{\sum X_i}{n}\cdot\frac{\sum Y_i}{n}}{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2} = \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

Multiplying numerator and denominator by n, we have

$$b = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2} \qquad \ldots(9)$$

To write the shortcut formula for b, we shall show that it is independent of change of origin but not of change of scale.

As in case of coefficient of correlation we define

$$u_i = \frac{X_i - A}{h} \quad \text{and} \quad v_i = \frac{Y_i - B}{k}$$

Or $X_i = A + hu_i$ and $Y_i = B + kv_i$

$\therefore$ $\bar{X} = A + h\bar{u}$ and $\bar{Y} = B + k\bar{v}$

also $\left(X_i - \bar{X}\right) = h\left(u_i - \bar{u}\right)$ and $Y_i - \bar{Y} = k\left(v_i - \bar{v}\right)$

Substituting these values in equation (6), we have

$$b = \frac{hk\sum\left(u_i - \bar{u}\right)\left(v_i - \bar{v}\right)}{h^2\sum\left(u_i - \bar{u}\right)^2} = \frac{k\sum\left(u_i - \bar{u}\right)\left(v_i - \bar{v}\right)}{h\sum\left(u_i - \bar{u}\right)^2}$$

$$= \frac{k}{h}\left[\frac{n\sum u_i v_i - \left(\sum u_i\right)\left(\sum v_i\right)}{n\sum u_i^2 - \left(\sum u_i\right)^2}\right] \qquad \ldots(10)$$

(***Note:*** If h = k they will cancel each other)

Consider equation (8), $b = \frac{Cov\left(X,Y\right)}{\sigma_X^2}$

Writing $Cov\left(X,Y\right) = r.\sigma_X\sigma_Y$, we have

$$b = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r \times \frac{\sigma_Y}{\sigma_X}$$

The line of regression of Y on X, i.e $Y_{Ci} = a + bX_i$ can also be written as

$$Y_{Ci} = \bar{Y} - b\bar{X} + bX_i \quad \text{or} \quad Y_{Ci} - \bar{Y} = b\left(X_i - \bar{X}\right) \qquad \ldots(11)$$

or $$\left(Y_{Ci} - \bar{Y}\right) = r \times \frac{\sigma_Y}{\sigma_X}\left(X_i - \bar{X}\right) \qquad \ldots(12)$$

9.5.2 Line of Regression of X on Y

The general form of the line of regression of X on Y is $X_{Ci} = c + dY_i$, where X_{Ci} denotes the predicted or calculated or estimated value of X for a given value of $Y = Y_i$ and c and d are constants. d is known as the regression coefficient of regression of X on Y.

In this case, we have to calculate the value of c and d so that

$S' = \Sigma (X_i - X_{Ci})^2$ is minimised.

As in the previous section, the normal equations for the estimation of c and d are

$$\Sigma X_i = nc + d\ \Sigma Y_i \qquad \ldots(13)$$

and $$\Sigma X_i Y_i = c\ \Sigma Y_i + d\ \Sigma Y_i^2 \qquad \ldots(14)$$

Dividing both sides of equation (13) by n, we have

$$\bar{X} = c + d\bar{Y}$$

This shows that the line of regression also passes through the point $(\bar{X}, \bar{Y})$. Since both the lines of regression passes through the point $(\bar{X}, \bar{Y})$, therefore $(\bar{X}, \bar{Y})$ is their point of intersection.

We can write $c = \bar{X} - d\bar{Y}$

As before, the various expressions for d can be directly written, as given below:

$$d = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum Y_i^2 - n\bar{Y}^2} \qquad \ldots(16)$$

or $$d = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} \qquad \ldots(17)$$

or $$d = \frac{\sum x_i y_i}{\sum y_i^2} \qquad \ldots(18)$$

$$= \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum (Y_i - \bar{Y})^2} = \frac{Cov(X,Y)}{\sigma_Y^2} \qquad \ldots(19)$$

Also $$d = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n\sum Y_i^2 - (\sum Y_i)^2} \qquad \ldots(20)$$

This expression is useful for calculating the value of d. Another short-cut formula for the calculation of d is given by

$$d = \frac{h}{k}\left[\frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{n\sum v_i^2 - (\sum v_i)^2}\right] \qquad \ldots(21)$$

where $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$

Consider equation (19)

$$d = \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{r\sigma_X\sigma_Y}{\sigma_Y^2} = r \cdot \frac{\sigma_X}{\sigma_Y} \qquad \ldots(22)$$

Substituting the value of c from equation (15) into line of regression of X on Y, we have

$$X_{Ci} = \bar{X} - d\bar{Y} + dY_i \quad \text{or} \quad \left(X_{Ci} - \bar{X}\right) = d\left(Y_i - \bar{Y}\right) \qquad \ldots(23)$$

or $$\left(X_{Ci} - \bar{X}\right) = r \cdot \frac{\sigma_X}{\sigma_Y}\left(Y_i - \bar{Y}\right) \qquad \ldots(24)$$

It should be noted here that the two lines of regression are different because these have been obtained in entirely two different ways. In case of regression of Y on X, it is assumed that the values of X are given and the values of Y are estimated by minimising $\Sigma(Y_i - Y_{Ci})^2$ while in case of regression of X on Y, the values of Y are assumed to be given and the values of X are estimated by minimising $\Sigma(X_i - X_{Ci})^2$. Since these two lines have been estimated on the basis of different assumptions, they are not reversible, i.e., it is not possible to obtain one line from the other by mere transfer of terms. There is, however, one situation when these two lines will coincide.

From the study of correlation, we may recall that when r = ±1, there is perfect correlation between the variables and all the points lie on a straight line. Therefore, both the lines of regression coincide and hence they are also reversible in this case. By substituting r = ±1 in equation (12) or (24), it can be shown that the lines of regression in both the cases become

$$\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right) = \pm\left(\frac{X_i - \bar{X}}{\sigma_X}\right)$$

Further when r = 0, equation (12) becomes $Y_{Ci} = \bar{Y}$ and equation (24) becomes $X_{Ci} = \bar{X}$. These are the equations of lines parallel to X-axis and Y-axis respectively. These lines also intersect at the point $\left(\bar{X}, \bar{Y}\right)$ and are mutually perpendicular at this point.

Correlation Coefficient and the Two Regression Coefficients

Since $b = r \times \frac{\sigma_Y}{\sigma_X}$ and $d = r \times \frac{\sigma_X}{\sigma_Y}$, we have

$b.d = r\frac{\sigma_Y}{\sigma_X} \times r\frac{\sigma_X}{\sigma_Y} = r^2$ or $r = \sqrt{b.d}$. This shows that correlation coefficient is the geometric mean of the two regression coefficients.

The following points should be kept in mind about the coefficient of correlation and the regression coefficients:

1. Since $r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, $b = \frac{Cov(X,Y)}{\sigma_X^2}$ and $d = \frac{Cov(X,Y)}{\sigma_Y^2}$, therefore the sign of r, b and d will always be same and this will depend upon the sign of Cov (X, Y).
2. Since bd = r^2 and $0 \le r^2 \le 1$, therefore either both b and d are less than unity or if one of them is greater than unity, the other must be less than unity such that $0 \le b.d \le 1$ is always true.

Example: Obtain the two regression equations and find correlation coefficient between X and Y from the following data:

X : 10 9 7 8 11

Y : 6 3 2 4 5

Solution:

Calculation Table

X	Y	XY	X^2	Y^2
10	6	60	100	36
9	3	27	81	9
7	2	14	49	4
8	4	32	64	16
11	5	55	121	25
45	**20**	**188**	**415**	**90**

(a) ***Regression of Y on X***

$$b = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} = \frac{5\times 188 - 45\times 20}{5\times 415 - (45)^2} = 0.8$$

Also, $\bar{X} = \frac{45}{5} = 9$ and $\bar{Y} = \frac{20}{5} = 4$

Now $a = \bar{Y} - b\bar{X} = 4 - 0.8 \times 9 = -3.2$

$\therefore$ Regression of Y on X is $Y_c = -3.2 + 0.8X$

(b) ***Regression of X on Y***

$$d = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum Y^2 - \left(\sum Y\right)^2} = \frac{5\times 188 - 45\times 20}{5\times 90 - (20)^2} = 0.8$$

Also, $= 9 - 0.8 \times 4 = 5.8$

$\therefore$ The regression of X on Y is $X_c = 5.8 + 0.8Y$

(c) ***Coefficient of correlation*** $r = \sqrt{b.d} = \sqrt{0.8\times 0.8} = 0.8$

Example: From the data given below, find:

(a) The two regression equations.

(b) The coefficient of correlation between marks in economics and statistics.

(c) The most likely marks in statistics when marks in economics are 30.

Marks in Eco. : 25 28 35 32 31 36 29 38 34 32

Marks in Stat. : 43 46 49 41 36 32 31 30 33 39

Solution:

Calculation Table

Marks in Eco. (X)	Marks in Stat. (Y)	u = X – 31	v = Y – 41	uv	u^2	v^2
25	43	–6	2	–12	36	4
28	46	–3	5	–15	9	25
35	49	4	8	32	16	64
32	41	1	0	0	1	0
31	36	0	–5	0	0	25
36	32	5	–9	–45	25	81
29	31	–2	–10	20	4	100
38	30	7	–11	–77	49	121
34	33	3	–8	–24	9	64
32	39	1	–2	–2	1	4
Total		**10**	**– 30**	**– 123**	**150**	**488**

From the table, we have

$$\bar{X} = 31 + \frac{10}{10} = 32 \text{ and } \bar{Y} = 41 - \frac{30}{10} = 38.$$

(a) ***The lines of regression***

(i) *Regression of Y on X*

$$b = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2} = \frac{-1230 + 300}{1500 - 100} = -0.66$$

$$= 38 + 0.66 \times 32 = 59.26$$

∴ Regression equation is

$$Y_C = 59.26 - 0.66X$$

(ii) *Regression of X on Y*

$$d = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum v^2 - (\sum v)^2} = \frac{-1230 + 300}{4880 - 900} = -0.23$$

$$c = \bar{X} - d\bar{Y} = 32 + 0.23 \times 38 = 40.88$$

∴ Regression equation is

$$X_C = 40.88 - 0.23Y$$

(b) ***Coefficient of correlation***

$$r = \sqrt{b \times d} = -\sqrt{-0.66 \times -0.23} = -0.39$$

Note that r, b and d are of same sign.

(c) Since we have to estimate marks in statistics denoted by Y, therefore, regression of Y on X will be used. The most likely marks in statistics when marks in economics are 30, is given by

$$Y_C = 59.26 - 0.66 \times 30 = 39.33$$

Example: Obtain the two lines of regression from the following data and estimate the blood pressure when age is 50 years. Can we also estimate the blood pressure of a person aged 20 years on the basis of this regression equation? Discuss.

Age (X) (in years) : 56 42 72 39 63 47 52 49 40 42 68 60

Blood Pressure (Y) : 127 112 140 118 129 116 130 125 115 120 135 133

Solution:

Calculation Table

X	Y	u = X − 52	v = Y − 125	uv	u^2	v^2
56	127	4	2	8	16	4
42	112	−10	−13	130	100	169
72	140	20	15	300	400	225
39	118	−13	− 7	91	169	49
63	129	11	4	44	121	16
47	116	−5	−9	45	25	81
52	130	0	5	0	0	25
49	125	−3	0	0	9	0
40	115	−12	−10	120	144	100
42	120	−10	−5	50	100	25
68	135	16	10	160	256	100
60	133	8	8	64	64	64
Total		**6**	**0**	**1012**	**1404**	**858**

From the table, we have

$$\bar{X} = 52 + \frac{6}{12} = 52.5 \quad \text{and} \quad \bar{Y} = 125$$

(a) ***Regression of Y on X***

$$b = \frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum u^2 - \left(\sum u\right)^2} = \frac{12 \times 1012 - 6 \times 0}{12 \times 1404 - (6)^2} = 0.72$$

Also $a = \bar{Y} - b\bar{X} = 125 - 0.72 \times 52.5 = 87.2$

∴ The line of regression of Blood pressure (Y) on Age (X) is

$$Y_C = 87.2 + 0.72X$$

(b) ***Regression of X on Y***

$$d = \frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum v^2 - \left(\sum v\right)^2} = \frac{12 \times 1012 - 6 \times 0}{12 \times 858 - 0} = 1.18$$

Also $c = \bar{X} - d\bar{Y} = 52.5 - 1.18 \times 125 = -95$

∴ Line of regression of Age (X) on Blood pressure (Y) is

$$X_C = -95 + 1.18Y$$

(c) (i) To estimate blood pressure (Y) for a given age, X = 50 years, you should use regression of Y on X

$\therefore \quad Y_C = 87.2 + 0.72 \times .50 = 123.2$

(ii) The estimate of blood pressure when age is 20 years

$Y_C = 87.2 + 0.72 \times .20 = 101.6$

It should be noted here that this estimate is wrong because the blood pressure of a normal person cannot be less than 110.

This result reflects the limitations of regression analysis with regard to estimation or prediction.

It is important to note that the prediction, based on regression line, should be done only for those values of the variable that are not very far from the range of the observed data, used to derive the line of regression.

The prediction from a regression line for a value of the variable that is far away from the observed data is likely to give inconsistent results like the one obtained above.

9.5.3 Regression Coefficient in a Bivariate Frequency Distribution

As in case of calculation of correlation coefficient, we can directly write the formula for the two regression coefficients for a bivariate frequency distribution as given below:

$$b = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{N\sum f_iX_i^2 - \left(\sum f_iX_i\right)^2}$$

or, if we define $u_i = \frac{X_i - A}{h}$ and $v_j = \frac{Y_j - B}{k}$,

$$b = \frac{k}{h}\left[\frac{N\sum\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{N\sum f_iu_i^2 - \left(\sum f_iu_i\right)^2}\right]$$

Similarly, $$d = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{N\sum f_j'Y_i^2 - \left(\sum f_j'Y_j\right)^2}$$

or $$d = \frac{h}{k}\left[\frac{N\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{N\sum f_j'v_i^2 - \left(\sum f_j'v_j\right)^2}\right]$$

Example: By calculating the two regression coefficients obtain the two regression lines from the following data:

Y → / X ↓	0-5	5-10	10-15
0-10	2	5	7
10-20	1	3	2
20-30	8	4	0

Solution:

The mid points of X-values are 5, 15, 25.

Let $u = \frac{X-15}{10}$, $\therefore$ Corresponding u-values become – 1, 0, 1

Similarly, the mid-points of Y-values are 2.5, 7.5, 12.5

Let $v = \frac{Y-7.5}{5}$, $\therefore$ Corresponding v-values become – 1, 0, 1

Calculation Table

u \ v	–1	0	1	f_i	$f_i u_i$	$f_i u_i^2$	$f_{ij} u_i v_j$
–1	2 [2]	5 [0]	7 [–7]	14	–14	14	–5
0	1 [0]	3 [0]	2 [0]	6	0	0	0
1	8 [–8]	4 [0]	0 [0]	12	12	12	–8
f_j'	11	12	9	32	–2	26	–13
$f_j' v_j'$	–11	0	9	–2			
$f_j' v_j'^2$	11	0	9	20			

From the table N = 32 (total frequency)

(a) ***Regression of Y on X***

Regression Coefficient (here h = 10 and k = 5)

$$b = \left[\frac{-32 \times 13 - 2 \times 2}{32 \times 26 - 4}\right] \times \frac{5}{10} = \frac{-416-4}{832-4} \times \frac{1}{2} = -0.25$$

Also, $\bar{X} = 15 + \frac{10(-2)}{32} = 14.73$ and $\bar{Y} = 7.5 + \frac{5(-2)}{32} = 7.19$

$\therefore a = \bar{Y} - b\bar{X} = 7.19 + 0.25 \times 14.73 = 10.87$

Hence, the regression of Y on X becomes $Y_C = 10.87 - 0.25X$

(b) ***Regression of X on Y***

Regression coefficient $d = \left[\frac{-420}{32 \times 20 - 4}\right] \times \frac{10}{5} = -1.32$

Also, $c = \bar{X} - d\bar{Y} = 14.73 + 1.32 \times 7.19 = 24.22$

Hence, the regression of X on Y becomes $X_C = 24.22 - 1.32Y$

9.5.4 Coefficient of Determination

We recall that in the line of regression $Y_C = a + bX$, X is used to estimate the value of Y. Further, the estimate of Y, independently of X, is given by a constant. Let this constant be A. Thus, we can write $Y_C = A$.

Given the observations $Y_1, Y_2, \ldots\ldots Y_n$, A will be the best estimate of Y if $S = \sum_{i=1}^{n}(Y_i - A)^2$ is minimum.

The necessary condition for minimum of S is $\frac{\partial S}{\partial A} = 0$.

. i.e., $2\sum(Y_i - A) = 0$ or $\sum Y_i - nA = 0$ or $A = \bar{Y}$

$\therefore$ The best estimate (an estimate having minimum sum of squares of errors) of Y, independently of X, is given by $Y_C = \bar{Y}$.

If X and Y are independent variables, the two lines of regression are $Y_C = \bar{Y}$ and $Y_C = \bar{Y}$.

Very often, when we use X for the estimation of Y, we are interested in knowing how far the use of X enables us to explain the variations in Y values from or, in other words, how much of the variations in Y, from , are being explained by the regression equation $Y_{Ci} = a + bX_i$? To answer this question, we write

$Y_i - \bar{Y} = Y_i - Y_{Ci} + Y_{Ci} - \bar{Y}$ (Subtracting and adding Y_{Ci})

or $Y_i - \bar{Y} = (Y_i - Y_{Ci}) + (Y_{Ci} - \bar{Y})$

Squaring both sides and taking sum over all the observations, we have

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - Y_{Ci})^2 + \sum(Y_{Ci} - \bar{Y})^2 + 2\sum(Y_i - Y_{Ci})(Y_{Ci} - \bar{Y}) \quad \ldots(1)$$

Consider the product term

$$2\sum(Y_i - Y_{Ci})(Y_{Ci} - \bar{Y}) = 2\sum\left[\{Y_i - \bar{Y} - b(X_i - \bar{X})\}\{b(X_i - \bar{X})\}\right]$$

$$= 2b\sum(Y_i - \bar{Y})(X_i - \bar{X}) - 2b^2\sum(X_i - \bar{X})^2$$

$$= 2b^2\sum(X_i - \bar{X})^2 - 2b^2\sum(X_i - \bar{X})^2 = 0$$

Thus, equation (1) becomes

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - Y_{Ci})^2 + \sum(Y_{Ci} - \bar{Y})^2 \quad \ldots(2)$$

You may observed that $Y_{Ci} - \bar{Y}$ is the deviation of the estimated value from $\bar{Y}$. This deviation has occurred because X and Y are related by the regression equation $Y_{Ci} = a + bX_i$, so that the estimate of Y is Y_{Ci} when $X = X_i$. Similar type of deviations would occur for other values of X. Thus, the magnitude of the term $\sum(Y_{Ci} - \bar{Y})^2$ gives the strength of the relationship, $Y_{Ci} = a + bX_i$, between X and Y or, equivalently, the variations in Y that are explained by the regression equation.

The other term $Y_i - Y_{Ci}$ gives the deviation of i^{th} observed value from the regression line and thus the magnitude of the term $\sum(Y_i - Y_{Ci})^2$ gives the variations in Y about the line of regression. These variations are also known as unexplained variations in Y.

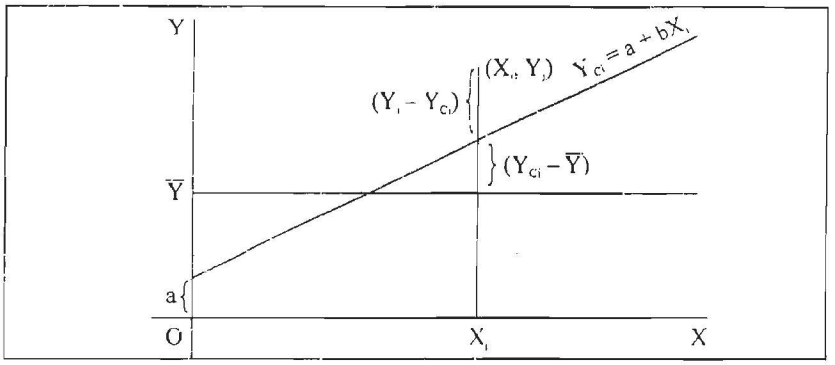


Figure 9.2

Adding the two types of variations, we get the magnitude of total variations in Y. Thus, equation (2) can also be written as

Total variations in Y = Unexplained variations in Y + Explained variations in Y.

Dividing both sides of equation (2) by $\sum(Y_i - \bar{Y})^2$, we have

$$1 = \frac{\sum(Y_i - Y_{Ci})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum(Y_{Ci} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \quad \ldots(3)$$

or 1 = Proportion of unexplained variations + Proportion of variations explained by the regression equation.

The proportion of variation explained by regression equation is called the coefficient of determination.

Thus, the coefficient of determination $= \frac{\sum(Y_{Ci} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$

$$= \frac{b^2\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\left[\sum(X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum(X_i - \bar{X})^2\sum(Y_i - \bar{Y})^2} = r^2$$

This result shows that the coefficient of determination is equal to the square of the coefficient of correlation, i.e., r^2 gives the proportion of variations explained by each regression equation.

It should be obvious from the above that it is desirable to calculate the coefficient of correlation prior to the fitting of a regression line. If r^2 is high enough, the fitted line will explain a greater proportion of the variations in the dependent variable. A low value of r^2 would, however, indicate that the proposed fitting of regression would not be of much use.

The expression for the coefficient of determination for regression of X on Y can be written in a similar way. Here we can write $r^2 = \frac{\sum(X_{Ci} - \bar{X})^2}{\sum(X_i - \bar{X})^2}$.

9.5.5 Coefficient of Non-determination

The proportion of unexplained variations is also termed as the coefficient of non-determination. It is denoted by k^2, where $k^2 = (1 - r^2)$. The square root of k^2 is termed as the coefficient of alienation, i.e., $k = \sqrt{(1-r^2)}$.

Example: Comment on the following statements:

(i) The two regression coefficients of bivariate data are 0.7 and 1.4.

(ii) A correlation coefficient r = 0.8, between the two variables X and Y, implies a relationship twice as close as r = 0.4.

Solution:

(i) This statement implies that $r^2 = 0.7 \times 1.4 = 0.98$, i.e., a linear regression fitted to the data would explain 98% of the variations in the dependent variable.

(ii) The given statement is wrong. Since r = 0.8 implies that a regression fitted to the data would explain 64% of the variations in the dependent variable while r = 0.4 implies that the proportion of such variations is only 16%. Thus, r = 0.8 implies a relation that is four times as close as r = 0.4.

Example: The correlation coefficient between two variables is found to be 0.8. Explain the meaning of this statement.

Solution:

The given statement implies that:

(i) Two variables are highly correlated.

(ii) There is positive association between them, i.e., an increase in value of one is accompanied by the increase in value of the other and vice-versa.

(iii) A linear regression fitted to the data would explain 64% of the variations in the dependent variable.

9.5.6 Mean of the Estimated Values

We may recall that Y_C and X_C are the estimated values from the regressions of Y on X and X on Y respectively.

Consider the regression equation $Y_{Ci} - \bar{Y} = b(X_i - \bar{X})$.

Taking sum over all the observations, we get

$$\Sigma(Y_{Ci} - \bar{Y}) - b\Sigma(X_i - \bar{X}) = 0$$

$$\Rightarrow \quad \Sigma Y_{Ci} - n\bar{Y} = 0 \quad or \quad \frac{\Sigma Y_{Ci}}{n} = \bar{Y}_C = \bar{Y} \qquad \ldots(1)$$

Similarly, it can be shown that $\bar{X}_C = \bar{X}$.

This implies that the mean of the estimated values is also equal to the mean of the observed values.

9.5.7 Mean and Variance of 'e_i' Values

(i) ***Mean of e_i values***

We know that $e_i = Y_i - Y_{Ci}$

Taking sum over all the observations, we have

$$\sum e_i = \sum (Y_i - Y_{Ci}) = \sum Y_i - \sum Y_{Ci} = 0 \text{ [from equation (1)]}$$

$\therefore$ Mean of e_i values is equal to zero.

(ii) ***Variance of e_i values***

The variance of e_i values, in case of regression of Y on X, is given by

$$S^2_{Y.X} = \frac{1}{n}\sum (e_i - 0)^2 = \frac{1}{n}\sum (Y_i - Y_{Ci})^2 \qquad \ldots (2)$$

[Note that $\sum (Y_i - Y_{Ci})^2$ is the magnitude of unexplained variation in Y]

$$S^2_{Y.X} = \frac{1}{n}\sum \left[(Y_i - \bar{Y}) - b(X_i - \bar{X}) \right]^2$$

$$= \frac{\sum (Y_i - \bar{Y})^2}{n} + \frac{b^2 \sum (X_i - \bar{X})^2}{n} - \frac{2b \sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$= \sigma_Y^2 + b^2\sigma_X^2 - 2b \cdot b\sigma_X^2 = \sigma_Y^2 - b^2\sigma_X^2$$

$$= \sigma_Y^2 - r^2\sigma_Y^2 = \sigma_Y^2 (1 - r^2)$$

Similarly, it can be shown that the mean of e'_i ($= X_i - X_{Ci}$) values, in case of regression of X on Y, is also equal to zero. Further, their variance, i.e.,

$$S^2_{X.Y} = \sigma_X^2 (1 - r^2)$$

Alternatively equation (2) can be written as

$$S^2_{Y.X} = \frac{1}{n}\sum (Y_i - Y_{ci}) Y_i - \frac{1}{n}\left[\sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i \right]$$

Similarly, we can write

$$\Sigma^2_{X.Y} = \frac{1}{n}\left[\sum X_i^2 - c\sum X_i - d\sum X_i Y_i \right]$$

The above expressions for the variance are based on the following:

$$(Y_i - Y_{ci})^2 = \sum (Y_i - Y_{ci})(Y_i - Y_{ci})$$

$$= \sum (Y_i - Y_{ci}) Y_i - \sum (Y_i - Y_{ci}) Y_{ci}$$

It can be shown that the last term is zero.

$$\sum (Y_i - Y_{ci}) Y_{ci} = \sum [(Y_i - \bar{Y}) - b(X_i - \bar{X})][+ b(X_i - \bar{X})]$$

$$= \sum (Y_i - \bar{Y}) - b\sum (X_i - \bar{X}) + b\sum (X_i - \bar{X})(Y_i - \bar{Y}) - b^2 \sum (X_i - \bar{X})^2$$

$$= 0 - 0 + b^2 \sum (X_i - \bar{X})^2 - b^2 \sum (X_i - \bar{X})^2 = 0$$

Standard Error of the Estimate

The standard error of the estimate of regression is given by the positive square root of the variance of e_i values.

The standard error of the estimate of regression of Y on X or simply the standard error of the estimate of Y is given as, $S_{Y.X} = \sigma_Y \sqrt{1-r^2}$.

Similarly, $S_{Y.X} = \sigma_Y \sqrt{1-r^2}$ is the standard error of the estimate X.

According to the theory of estimation, an unbiased estimate of the variance of e_i values is given by

$$s^2_{Y.X} = \frac{\sum e_i^2}{n-2} = \frac{n}{n-2} \times \frac{\sum e_i^2}{n} = \frac{n}{n-2} \times \sigma_Y^2 \left(1-r^2\right)$$

$\therefore$ The standard errors of the estimate of Y and that of X are written as

$s_{X.Y} = \sigma_Y \sqrt{\frac{n}{(n-2)}\left(1-r^2\right)}$ and $s_{X.Y} = \sigma_X \sqrt{\frac{n}{(n-2)}\left(1-r^2\right)}$ respectively.

Note that difference between these standard errors tend to be equal to the standard errors for large values of n. In practice, the value of $n > 30$ may be treated as large.

Example: From the following data, compute (i) the coefficient of correlation between X and Y, (ii) the standard error of the estimate of Y:

$\sum x^2 = 24 \quad \sum y^2 = 42 \quad \sum xy = 30 \quad N = 10$, where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

Solution:

The coefficient of correlation between X and Y is given by

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{30}{\sqrt{24}\sqrt{42}} = 0.94$$

The standard error of the estimate of Y is given by ($n < 30$)

$$s_{Y.X} = \sqrt{\frac{\left(1-r^2\right)\sum y^2}{n-2}} = \sqrt{\frac{\left(1-0.94^2\right)\times 42}{8}} = 0.79$$

Example: For 100 items, it is given that the regression equations of Y on X and X on Y are $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ respectively. Compute the arithmetic means of X and Y and the coefficient of determination. If the standard deviation of X is given to be 3, compute the standard error of the estimate of Y.

Solution:

(a) ***Means of X and Y:*** Since the lines of regression pass through the point $\left(\bar{X}, \bar{Y}\right)$, the simultaneous solution of the given regression equations would give the mean values of X and Y as $\bar{X} = 13, \bar{Y} = 17$.

(b) ***Coefficient of determination:*** We assume that $8X - 10Y + 66 = 0$ is the regression of Y on X and $40X - 18Y = 214$ is the regression of X on Y. Thus, the respective regression coefficients b and d are given by $\frac{8}{10}$ and $\frac{18}{40}$.

$\therefore$ The coefficient of determination $r^2 = b.d = \frac{8}{10} \times \frac{18}{40} = 0.36$

(c) ***Standard error of the estimate of Y:*** We know that $\sigma_{Y\,X} = \sigma_Y \sqrt{1 - r^2}$. To find s_Y we use the relation $b = r \times \frac{\sigma_Y}{\sigma_X}$

Also $r^2 = \frac{9}{25}$ $\therefore$ $r = \frac{3}{5}$ Thus, $\sigma_Y = \frac{b.\sigma_X}{r} = \frac{8}{10} \times \frac{5}{3} \times 3 = 4$

Hence, $\sigma_{Y\,X} = 4\sqrt{1 - 0.36} = 3.2$

9.6 MULTIPLE REGRESSION

In case of two variable relationships, we ignore the possibility of variations in dependent variables due to two or more independent variables. We have considered a simple model of first order with one independent variable. But in a real business situation, we may have number of independent variables e.g. demand of commodity may be dependent on price, requirement of consumers, availability of money with consumers, alternative products available in the market, change in consumption pattern, etc. In such case, we may have to use relation like,

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \in$$

We could develop 'multiple correlations' and multiple linear regression that consider simultaneously linear regression relationships among three or more variables. As regards to mathematical treatment using Manual calculations, we will restrict to three variables.

In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.

For more than three variables, we can use computer packages or MS excel because mathematical relationships become too complex. Also, once we understand the concept of multiple regressions, it is efficient to use computers for solving the problems.

9.7 MULTIPLE REGRESSION EQUATION WITH THREE VARIABLES

Let there be three variable X_1, X_2 and X_3. Let X_1 be dependent variable and X_2 and X_3 independent variables. If we consider linear relationship between X_1 and X_2 as well as X_1 and X_3, the locus of X_1, X_2 and X_3 is a plane. Regression equation establishes the plane that defines the average relationship. In general, we can write the relation as,

$$\hat{x}_1 = \alpha + \beta_{12\bullet 3} x_2 + \beta_{13\bullet 2} x_3 \qquad \ldots(1)$$

$\hat{x}_1$ is estimate of X_1 based of x_2 and x_3 along with the regression equation. α is X_1 intercept and $\beta_{12\bullet 3}$ and $\beta_{13\bullet 2}$ are regression coefficients based of data on population. Since we do not know the complete data we can only calculate the estimates of $\alpha, \beta_{12\bullet 3}$ and $\beta_{13\bullet 2}$. These we denote as 'a, $b_{12\bullet 3}$, $b_{13\bullet 2}$' respectively. $b_{12\bullet 3}$ is the partial regression coefficient (estimate) of X_1 on X_2 keeping X_3 as constant. Whereas, $b_{13\bullet 2}$ is the partial regression coefficient of X_1 on X_3 keeping X_2 as constant. Thus $b_{12\bullet 3}$ and

$b_{13\bullet 2}$ represent the contributions (rate) that x_2 and x_3 make $\hat{x}$ respectively. We can calculate the equation of the regression plane as,

$$\hat{x}_1 = a + b_{12\bullet 3}x_2 + b_{13\bullet 2}x_3 \quad \ldots(2)$$

In this case, these regression coefficients only indicate isolated (partial) influence. These do not take into account what is the interaction between X_2 and X_3. There are cases where two independent variables influence each other and none may independently have any correlation with dependent variable. Yet, the measurement of 'multiple relationship' may reveal a high degree and correlation.

Now we need to find the values of $a, b_{12\bullet 3}$ and $b_{13\bullet 2}$. Once again we use least square fit that we have used for linear and nonlinear regression equations of two variables. First we calculate the Sum Square (SS) of deviations (errors) of x_1 (the actual value or observed value) from $\hat{x}$ (the estimated value using the regression equation). This is given by,

$$SSE = \sum (x_1 - \hat{x}_1)^2 = \sum (x_1 - a - b_{12\bullet 3}x_2 - b_{13\bullet 2}x_3)^2$$

To minimize the SSE its partial derivatives with respective to $a, b_{12\bullet 3}, b_{13\bullet 2}$ must be zero. By equating partial derivatives to zero and simplification we get "Normal Equations" as:

$$\sum x_1 = a \times n + b_{12\bullet 3}\sum x_2 + b_{13\bullet 2}\sum x_3 \quad \ldots(3)$$

$$\sum x_1 x_2 = a\sum x_2 + b_{12\bullet 3}\sum x_2^{\,2} + b_{13\bullet 2}\sum x_2 x_3 \quad \ldots(4)$$

$$\sum x_1 x_3 = a\sum x_3 + b_{12\bullet 3}\sum x_2 x_3 + b_{13\bullet 2}\sum x_3^{\,2} \quad \ldots(5)$$

These three linear equations can be solved if we know data on set of corresponding values of x_1, x_2 and x_3. Solving we get the values of a, $b_{12\bullet 3}$ and $b_{13\bullet 2}$ and hence the regression equation (of the plane) as, $\hat{x}_1 = a + b_{12\bullet 3}x_2 + b_{13\bullet 2}x_3$

The equation (3) can be divided by simplified and simplified as,

$$\bar{X}_1 = a + b_{12\bullet 3} \times \bar{X}_2 + b_{13\bullet 2} \times \bar{X}_3 \quad \ldots(6)$$

Subtracting equation (6) from (2) we can rewrite the regression equation as,

$$(\hat{x} - \bar{X}_1) = b_{12\bullet 3}(x_2 - \bar{X}_2) + b_{13\bullet 2}(x_3 - \bar{X}_3) \quad \ldots(7)$$

If we know $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3$, then we need to find $b_{12\bullet 3}$ and $b_{13\bullet 2}$ so as to get the regression equation. From normal equations and measuring x_1, x_2 and x_3 from $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3$ without loss of generality i.e. shifting the origin to a point $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$ (We have seen that change of origin does not change regression coefficients). We get,

$$b_{12\bullet 3} = \frac{\sum x_1 x_2 \sum x_3^{\,2} - (\sum x_1 x_3)(\sum x_2 x_3)}{\sum x_1^{\,2} \sum x_3^{\,2} - (\sum x_2 x_3)^2} \quad \ldots(8)$$

$$b_{13\bullet 2} = \frac{\sum x_1 x_3 \sum x_2^{\,2} - (\sum x_1 x_2)(\sum x_2 x_3)}{\sum x_1^{\,2} \sum x_2^{\,2} - (\sum x_2 x_3)^2} \quad \ldots(9)$$

Note that in this formula x_1, x_2 and x_3 are measured from $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3$ and not from origin.

Using MS Excel for Multiple Regression

It is very simple to find multiple regression equation and analysis of multiple regression using MS Excel. This requires use of 'Data Analysis Pak' in 'Tools' menu. (If you don't have it in 'Tools' menu, you can use 'Insert' → 'Add In' menu to add 'Data Analysis Pak'). The procedure is as follows:

(i) Open an MS Excel worksheet. Enter the data of x_1, x_2 and x_3 in three adjacent columns (or rows).

(ii) Select any cell and use **'Data Analysis Pak'** tool **'Regression'** from the drop down menu as **[Tools→Data Analysis...→Regression→OK]** if you have Excel 97-2003 or from Quick Access Tool Bar as **[Data→Data Analysis→Regression→OK]**.

(iii) Now the regression menu box will appear. Follow the box and enter **'Input Y Range:'** as column of values of x_1. We can either enter cell range or 'drag select' the range of values with mouse. We can also select the heading or label cell.

(iv) **'Input X Range:'** as both columns containing values of x_2 and x_3. Note that these values should be in adjacent columns.

(v) Also select other choices as required. Check the box for labels if label cells are selected in input. Select the **'Constant is zero'** box **if the intercept is at origin** (i.e. $a = 0$ or $x_1 = 0$ for $x_2 = x_3 = 0$). Select output range as per your choice where you want output to be displayed. (You can also get Residue Plot, Line Fit Plot, Normal Probability plot, etc. It is recommended that you try these on your own. These are very useful for your guidelines during decision-making). Then select **'OK'**.

(vi) We get the result (output) as Summary Output, ANOVA and coefficient analysis (Besides the residue plots if you have asked for).

This is very quick and effective way of solving multiple regression problems. We can use it for 'Multiple Regression' up to 16 independent variables.

Example: Given below is the data of 10 people from a group on Body Metabolic Rate in Kcal/day (BMR), Body Mass Index in Kg/m^2 and Age in Years. Find the multiple regression relationship (regression plane), comment on the relationship, and find estimate of BMR for a person of age 32 and BMI 22.48.

Sl. No.	BMR	BMI	Age
1.	1459.30	20.43	21
2.	1451.60	20.49	23
3.	1352.20	18.08	25
4.	1535.80	22.43	27
5.	1581.70	22.76	29
6.	1470.60	20.96	30
7.	1569.20	25.09	37
8.	1493.80	20.70	41
9.	1470.60	22.67	49
10.	1157.50	19.48	19
11.	1474.60	22.51	21
12.	1597.00	25.63	25
13.	1581.10	22.55	26
14.	1482.20	21.10	38
15.	1459.00	20.81	46

Solution:

Using Manual Calculations

(i) The calculations are shown in the following table.

Sl. No.	BMR= x_1	BMI= x_2	Age= x_3	x_1x_2	x_1x_3	x_2x_3	x_2^2	x_3^2
1.	1459.3	20.43	21	29813.50	30645.3	429.03	417.38	441
2.	1451.6	20.49	23	29743.28	33386.8	471.27	419.84	529
3.	1352.2	18.08	25	24447.78	33805.0	452.00	326.89	625
4.	1535.8	22.43	27	34447.99	41466.6	605.61	503.10	729
5.	1581.7	22.76	29	35999.49	45869.3	660.04	518.02	841
6.	1470.6	20.96	30	30823.78	44118.0	628.80	439.32	900
7.	1569.2	25.09	37	39371.23	58060.4	928.33	629.51	1369
8.	1493.8	20.70	41	30921.66	61245.8	848.70	428.49	1681
9.	1470.6	22.67	49	33338.50	72059.4	1110.83	513.93	2401
10.	1157.5	19.48	19	22548.10	21992.5	370.12	379.47	361
11.	1474.6	22.51	21	33193.25	30966.6	472.71	506.70	441
12.	1597.0	25.63	25	40931.11	39925.0	640.75	656.90	625
13.	1581.1	22.55	26	35653.81	41108.6	586.30	508.50	676
14.	1482.2	21.10	38	31274.42	56323.6	801.80	445.21	1444
15.	1459.0	20.81	46	30361.79	67114.0	957.26	433.06	2116
Total $\sum$	**22136.2**	**325.69**	**457**	**482869.68**	**678086.9**	**9963.55**	**7126.32**	**15179**

Thus, we get "Normal Equations" as:

$$\sum x_1 = a \times n + b_{12\bullet3}\sum x_2 + b_{13\bullet2}\sum x_3$$

$$\Rightarrow 22136.2 = 15 \times a + 325.69 \times b_{12\bullet3} + 457 \times b_{13\bullet2}$$

$$\sum x_1x_2 = a\sum x_2 + b_{12\bullet3}\sum x_2^2 + b_{13\bullet2}\sum x_2x_3$$

$$\Rightarrow 482869.68 = 325.69 \times a + 7126.32 \times b_{12\bullet3} + 9963.55 \times b_{13\bullet2}$$

$$\sum x_1x_3 = a\sum x_3 + b_{12\bullet3}\sum x_2x_3 + b_{13\bullet2}\sum x_3^2$$

$$\Rightarrow 678086.9 = 457 \times a + 9963.55 \times b_{12\bullet3} + 15179 \times b_{13\bullet2}$$

Solving we get the coefficients of regression equation as,

$a = 566.1323$, $b_{12\bullet3} = 39.59968$ and $b_{13\bullet2} = 1.634565$

Thus the multiple regression relationship is,

$$\hat{x}_1 = 566.1323 + 39.59968 \times x_2 + 1.634565 \times x_3$$

(ii) Now for $x_2 = 22.48$ and $x_3 = 32$, we get

$$\hat{x}_1 = 566.1323 + 39.59968 \times 22.48 + 1.634565 \times 32$$

$$= 1508.6392$$

Using MS Excel

(a) Open an MS Excel worksheet. Enter the data of x_1 the BMR, x_2 the BMI and x_3 the age in three adjacent columns (say dependent variable BMR from B4 to B19 with heading in cell B3, independent variable BMI from C4 to C19 with heading in cell C3 and independent variable Age from D4 to D19 with heading in cell).

(b) Select any cell and use **'Data Analysis Pak'** tool **'Regression'** from the menu as **[Tools → Data Analysis...→ Regression → OK]** or **[Data → Data Analysis → Regression → OK]** depending on the Excel version.

(c) Now the regression menu box will appear. Follow the box and enter **'Input Y Range:'** as **$**B$3:$B$19 (or drag select the cell range), Select **'Input X Range:'** as **$**C$3:$D$19 (or drag select the cell range). Check the box labels. Let the confidence level remain as 95% default value. Don't check the box 'Constant is zero'. Select output range as 'New worksheet ply' where you want output to be displayed. (You can also get Residue Plot, Line Fit Plot, Normal Probability plot, etc. It is recommended that you try these on your own. These are very useful for your guidelines during decision-making). Then select **'OK'**.

(d) We get the resulst (output) as summary output, ANOVA and coefficient analysis (Besides the residue plots you have asked for) on new worksheet ply as follows,

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.751023
R Square	0.564036
Adjusted R Square	0.491375
Standard Error	78.00114
Observations	15

ANOVA

	df	SS	MS	F	Significance F
Regression	2	94458.18	47229.09	7.762608	0.006866
Residual	12	73010.14	6084.178		
Total	14	167468.3			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	566.1084	232.0184	2.439929	0.031162	60.58379	1071.633	60.58379	1071.633
BMI	39.60083	10.67503	3.709669	0.002983	16.34194	62.85972	16.34194	62.85972
Age	1.634527	2.228403	0.733497	0.477341	−3.22075	6.489801	−3.22075	6.489801

We can see that we got the same results as that of manual calculation (see coefficients column).

Thus, the multiple regression relationship is,

$$\hat{x}_1 = 566.1084 + 39.60083 \times x_2 + 1.634527 \times x_3$$

In fact, we get much more information as a result like regression statistics, ANOVA, standard errors and p-value to know the quality of our predictions.

9.8 REGRESSION EQUATION IN TERMS OF CORRELATION COEFFICIENTS

Let r_{12} be the simple linear correlation coefficient between X_1 and X_2. Similarly r_{13} and r_{23} are correlation coefficient between X_1 and X_3, and X_2 and X_3, respectively. These are known as zero order correlation coefficients. Now the variance of X_1 is,

$$\sigma_1^2 = \frac{\sum x_1^2}{n}, \quad \sigma_2^2 = \frac{\sum x_2^2}{n} \quad \text{and} \quad \sigma_3^2 = \frac{\sum x_3^2}{n}$$

Where x_1, x_2 and x_3 are measured from $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3$ without loss of generality.

We know that,

$$r_{12} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \sum x_2^2}} = \frac{\sum x_1 x_2}{n\sigma_1\sigma_2}$$

Or, $$\sum x_1 x_2 = n\sigma_1\sigma_2 r_{12}$$

Similarly,

$$\sum x_1 x_3 = n\sigma_1\sigma_3 r_{13} \text{ and } \sum x_2 x_3 = n\sigma_2\sigma_3 r_{23}$$

Substituting in equation (8) and (9) we get the regression coefficients as,

$$b_{12\bullet3} = \frac{(r_{12} - r_{13}r_{23})\sigma_1}{(1 - r_{23}^2)\sigma_2} \quad \ldots(10)$$

$$b_{13\bullet2} = \frac{(r_{13} - r_{12}r_{23})\sigma_1}{(1 - r_{23}^2)\sigma_3} \quad \ldots(11)$$

Thus we get the regression equation as,

$$\hat{x}_1 = b_{12\bullet3}x_2 + b_{13\bullet2}x_3 = \frac{(r_{12} - r_{13}r_{23})\sigma_1}{(1 - r_{23}^2)\sigma_2} \times x_2 + \frac{(r_{13} - r_{12}r_{23})\sigma_1}{(1 - r_{23}^2)\sigma_3} \times x_3$$

Or, $$\hat{x}_1 = \frac{\sigma_1}{(1 - r_{23}^2)}\left[\frac{(r_{12} - r_{13}r_{23})}{\sigma_2} \times x_2 + \frac{(r_{13} - r_{12}r_{23})}{\sigma_3} \times x_3\right]$$

Also if variable X_3 is uncorrected, i.e. $r_{13} = 0$ and $r_{23} = 0$ the regression equation reduces to,

$$\hat{x}_1 = b_{12}x_2 = r_{12}\frac{\sigma_1}{\sigma_2}x_2$$

This is same as the regression equation for two variables.

When the data points are measured from the origin (0, 0, 0) we need to modify the above equation. In that case rather than calculating $\bar{X}_1, \bar{X}_2, \bar{X}_3$ and shifting the coordinates we can calculate three zero order coefficients of linear correlation as,

$$r_{12} = \frac{\sum (x_1 - \bar{X}_1)(x_2 - \bar{X}_2)}{\sqrt{\sum (x_1 - \bar{X}_1)^2 \sum (x_2 - \bar{X}_2)^2}}$$

Or,
$$r_{12} = \frac{n\sum x_1 x_2 - \sum x_1 \sum x_2}{\sqrt{\left[n\sum x_1^2 - \left(\sum x_1\right)^2\right]\left[n\sum x_2^2 - \left(\sum x_2\right)^2\right]}}$$

We must note that in this case X_1 and X_2 are measured from origin.

Similarly,

$$r_{13} = \frac{n\sum x_1 x_3 - \sum x_1 \sum x_3}{\sqrt{\left[n\sum x_1^2 - \left(\sum x_1\right)^2\right]\left[n\sum x_3^2 - \left(\sum x_3\right)^2\right]}}$$

$$r_{23} = \frac{n\sum x_2 x_3 - \sum x_2 \sum x_3}{\sqrt{\left[n\sum x_2^2 - \left(\sum x_2\right)^2\right]\left[n\sum x_3^2 - \left(\sum x_3\right)^2\right]}}$$

Example: The following constants are obtained from the measurement on length in mm (X_1), volume in cm^3 (X_2) and weight in gm (X_3) of 100 fruits.

$\bar{X}_1 = 55.95$; $\bar{X}_2 = 51.48$; $\bar{X}_3 = 56.03$

$\sigma_1 = 2.26$; $\sigma_2 = 4.39$; $\sigma_3 = 4.41$

$r_{12} = 0.578$; $r_{13} = 0.581$; $r_{23} = 0.974$

Obtain the equation for the plane of regression of weight of the fruits on its length and volume. Also estimate the weight of a fruit whose length is 58 mm and volume is 52.2 cm^3.

Solution:

(i) Now the regression equation is given as,

$$\hat{x}_3 = \frac{\sigma_3}{(1 - r_{12}^2)}\left[\frac{(r_{31} - r_{32}r_{12})}{\sigma_1} \times x_1 + \frac{(r_{32} - r_{31}r_{12})}{\sigma_2} \times x_2\right]$$

Where x_1, x_2 and x_3 are measured from $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3$

Now substituting,

$$\hat{x}_3 = \frac{4.41}{(1 - 0.578^2)}\left[\frac{(0.581 - 0.974 \times 0.578)}{2.26} \times x_1 + \frac{(0.974 - 0.581 \times 0.578)}{4.29} \times x_2\right]$$

$$\hat{x}_3 = 6.62\left[0.008 \times x_1 + 0.149 \times x_2\right] = 0.053x_1 + 0.986x_2$$

Now if we measure x_1, x_2 and x_3 data points the origin (0, 0, 0), the equation becomes,

$$(\hat{x}_3 - \bar{X}_3) = 0.053(x_1 - \bar{X}_1) + 0.986(x_2 - \bar{X}_2)$$

Or, $(\hat{x}_3 - 56.03) = 0.053(x_1 - 55.95) + 0.986(x_2 - 51.48)$

Or, $\hat{x}_3 = 0.053x_1 + 0.986x_2 + 2.305$

(ii) Now for $x_1 = 58$ mm and $x_2 = 52.2$ cm^3

$\hat{x}_3 = 0.053x_1 + 0.986x_2 + 2.305 = 0.053 \times 58 + 0.986 \times 52.2 + 2.305 = 56.848$

$\hat{x}_3$ is the weight of 100 fruits. Hence, the weight of 1 fruit is, 0.56848 gm

9.9 STANDARD ERROR OF ESTIMATE

Measure of the standard error of estimate $\hat{x}_1$ is the deviation of observed value x_1 from the estimated value $\hat{x}_1$. The standard error is defined as,

$$\sigma_{1\bullet 23} = \sqrt{\frac{\sum(x_1 - \hat{x}_1)^2}{n}}$$

In this case, $1\bullet 23$ indicate that 1 is dependent variable and 2 and 3 are independent variables. Thus, it is a root mean of Sum Square of deviation of estimated value from the measured value of X_1, the estimate being obtained from the regression equation of X_1 on X_2 and X_3. With algebraic substitution and simplification, standard error of estimate can be written as,

$$\sigma_{1\bullet 23} = \sigma_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2 \times r_{12} r_{13} r_{23}}{(1 - r_{23}^2)}}$$

Similarly we could find,

$$\sigma_{2\bullet 13} = \sigma_2 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2 \times r_{12} r_{13} r_{23}}{(1 - r_{13}^2)}}$$

$$\sigma_{3\bullet 12} = \sigma_3 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2 \times r_{12} r_{13} r_{23}}{(1 - r_{12}^2)}}$$

The formula and manual calculation with addition of every independent variable becomes complex and time consuming. But with any computer package or MS Excel, it is always a good idea to include as many independent variables that we believe to have some influence on dependent variable, from our business analysis. If correlation analysis shows otherwise, then we could drop them in further analysis.

Example: Calculate the equation of regression of X_1 on X_2 and X_3 and estimate X_1 when $x_2 = 165$ and $x_3 = 175$.

Given, $\bar{X}_1 = 170$, $\bar{X}_2 = 160$, $\bar{X}_3 = 168$, $\sigma_1 = 2.4$, $\sigma_2 = 2.7$, $\sigma_3 = 2.7$

$r_{12} = 0.28$, $r_{13} = 0.49$ and $r_{23} = 0.51$

Solution:

Now, $$b_{12\bullet 3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} = \frac{2.4}{2.7} \times \frac{0.28 - 0.49 \times 0.51}{1 - 0.51^2} = 0.036$$

$$b_{13\bullet 2} = \frac{\sigma_1}{\sigma_3} \times \frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} = \frac{2.4}{2.7} \times \frac{0.49 - 0.28 \times 0.51}{1 - 0.51^2} = 0.417$$

Hence, the regression equation of X_1 on X_2 and X_3 is,

$$x_1 - \bar{X}_1 = b_{12\cdot3} \times (x_2 - \bar{X}_2) + b_{13\cdot2} \times (x_3 - \bar{X}_3)$$

Or, $$(x_1 - 170) = 0.036 \times (x_2 - 160) + 0.417 \times (x_3 - 168)$$

Or, $$x_1 = 0.036x_2 + 0.417x_3 + 94.184$$

Now, at $x_2 = 165$ and $x_3 = 175$

$$x_1 = 0.036 \times x_2 + 0.417 \times x_3 + 94.184 = 173.099$$

Example: Personal manager of a large industrial unit is interested to find a measure that can be used to fix the yearly wages of skilled workers. On an experimental basis, he compiled the data on the length of the service and their yearly wages (in ₹ '000) from a group of 10 randomly selected workers.

Length of Service (Years)	11	7	9	5	8	6	10	12	3	4
Yearly Wages (₹ '000)	14	11	10	9	13	10	14	16	6	7

Obtain the regression equation of wages on length of service.

Solution:

Using Manual Calculations

Method I: First we find out $\bar{X}$ and $\bar{Y}$ with number of observations n = 10. Further calculations are indicated in the table.

Sl. No.	$X = x_i$	$Y = y_i$	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - \bar{X})^2$	$(x_i - \bar{X})(y_i - \bar{Y})$
1.	11	14	3.5	3	12.25	10.5
2.	7	11	– 0.5	0	0.25	0
3.	9	10	1.5	– 1	2.25	– 1.5
4.	5	9	– 2.5	– 2	6.25	5
5.	8	13	0.5	2	0.25	1
6.	6	10	– 1.5	– 1	2.25	1.5
7.	10	14	2.5	3	6.25	7.5
8.	12	16	4.5	5	20.25	22.5
9.	3	6	– 4.5	– 5	20.25	22.5
10.	4	7	– 3.5	– 4	12.25	14
Total	**75**	**110**			**82.5**	**83**

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{75}{10} = 7.5 \qquad \text{and} \qquad \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{110}{10} = 11$$

Now, regression line slope for regression Y on X is,

$$b_{YX} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} = \frac{83}{82.5} = 1.006$$

Hence, the regression equation is,

$$\hat{y} - \bar{Y} = b_{YX}(X - \bar{X}) \Rightarrow \hat{y} = 11 + 1.006 \times (X - 7.5)$$

Or, $$\hat{y} = 3.455 + 1.006X$$

Method II

Sl. No.	$X = x_i$	$Y = y_i$	X^2	Y^2	XY
1.	11	14	121	196	154
2.	7	11	49	121	77
3.	9	10	81	100	90
4.	5	9	25	81	45
5.	8	13	64	169	104
6.	6	10	36	100	60
7.	10	14	100	196	140
8.	12	16	144	256	192
9.	3	6	9	36	18
10.	4	7	16	49	28
Total	**75**	**110**	**645**	**1304**	**908**

Regression equation Y on X is given by,

$$\hat{y} = a + bX$$

We have two normal equations,

$$\sum XY = a\sum X + b\sum X^2 \Rightarrow 908 = 75a + 645b$$

And, $$\sum Y = na + b\sum X \Rightarrow 110 = 10a + 75b$$

Solving these simultaneous equations, we get

$$a = 3.455 \quad \text{and} \quad b = 1.006$$

Hence, the regression equation Y on X is,

$$\hat{y} = 3.455 + 1.006X$$

9.10 DIFFERENCES BETWEEN CORRELATION ANALYSIS AND REGRESSION ANALYSIS

Both the techniques are directed towards a common purpose of establishing the degree and direction of relationship between two or more variables but the methods of doing so are different. The choice of one or the other will depend on the purpose. If the purpose is to know the degree and direction of relationship, correlation is an appropriate tool but if the purpose is to estimate a dependent variable with the substitution of one or more independent variables, the regression analysis shall be more helpful. The point of difference is discussed below:

- ***Degree and Nature of Relationship:*** The correlation coefficient is a measure of degree of co-variability between two variables whereas regression analysis is used to study the nature of relationship between the variables so that we can predict the

value of one on the basis of another. The reliance on the estimates or predictions depends upon the closeness of relationship between the variables.

- ***Cause and Effect Relationship:*** The cause and effect relationship is explained by regression analysis. Correlation is only a tool to ascertain the degree of relationship between two variables and we cannot say that one variable is the cause and other the effect. A high degree of correlation between price and demand for a commodity or at a particular point of time may not suggest which the cause is and which the effect is. However, in regression analysis cause and effect relationship is clearly expressed – one variable is taken as dependent and the other an independent.

Like in correlation, regression analysis can also be studied as 'simple and multiple', 'total and partial', 'linear and nonlinear', etc. depending upon the type of data and method we use for regression analysis. Regression word implies 'going back or falling back to mean or average value' but in most application of regression we do not use regression in this sense. We use it for the forecasting purpose or to understand underlying mathematical relationship.

Although correlation and regression both attempt to establish whether relationship exists between two or more variables or not, these two techniques differ in approach. If we only want to know the degree and direction of relationship we use correlation analysis. But if we want to forecast or predict the values we need regression analysis.

In correlation, there is no distinction between independent and dependent variables. But for regression analysis we need to specify independent and dependent variables clearly. In case of correlation, we are only interested in finding whether the relationship exists. Hence, the measuring error is only to establish confidence in our analysis. However, in regression our analysis itself is based on the concept of minimizing the errors.

Check Your Progress

Fill in the blanks:

1. If we only want to know the degree and direction of relationship we use _____ analysis.
2. The term regression was first introduced by _____ in 1877.
3. The regression equation is written as _____.
4. The non-systematic behaviour cannot be captured and called as _____.
5. The value of _____ will be different for different lines of regression.
6. The proportion of variation explained by regression equation is called the _____.

9.11 LET US SUM UP

- The regression equations are useful for predicting the value of dependent variable for given value of the independent variable.
- The nature of a regression equation is different from the nature of a mathematical equation.
- The term 'Regression', originated in this particular context, is now used in various fields of study, even though there may be no existence of any regressive tendency.
- For a bivariate data (X_i, Y_i), i = 1,2, ... n, we can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X.

- ***Regression of Y on X***

  Regression coefficient $b = \dfrac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$

  Also $b = \dfrac{Cov(X,Y)}{\sigma_X^2} = r \times \dfrac{\sigma_Y}{\sigma_X}$

- Regression of X on Y

  Regression Coefficient $d = \dfrac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2} = \dfrac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$

- Relation of r with b and d

  $$b \times d = r \cdot \frac{\sigma_Y}{\sigma_X} \cdot r \cdot \frac{\sigma_X}{\sigma_Y} = r^2$$

  or $r = \sqrt{b \times d}$

- For multiple regression and non-linear regression models, MS Excel or any other computer package would help in reducing voluminous calculations.
- We have considered a simple model of first order with one independent variable. But in a real business situation we may have number of independent variables e.g. demand of commodity may be dependent on price, requirement of consumers, availability of money with consumers, alternative products available in the market, change in consumption pattern, etc.
- Regression equation establishes the plane that defines the average relationship.
- These three linear equations can be solved if we know data on set of corresponding values of x_1, x_2 and x_3. Solving we get the values of a, $b_{12\cdot3}$ and $b_{13\cdot2}$ and hence the regression equation (of the plane) as, $\hat{x}_1 = a + b_{12\cdot3}x_2 + b_{13\cdot2}x_3$
- It is very simple to find multiple regression equation and analysis of multiple regression using MS Excel. This requires use of 'Data Analysis Pak' in 'Tools' menu. (If you don't have it in 'Tools' menu, you can use 'Insert'→ 'Add In' menu to add 'Data Analysis Pak').
- Let r_{12} be the simple linear correlation coefficient between X_1 and X_2. Similarly r_{13} and r_{23} are correlation coefficient between X_1 and X_3, and X_2 and X_3, respectively. These are known as zero order correlation coefficients. Now the variance of X_1 is,

  $$\sigma_1^2 = \frac{\sum x_1^2}{n}, \quad \sigma_2^2 = \frac{\sum x_2^2}{n} \quad \text{and} \quad \sigma_3^2 = \frac{\sum x_3^2}{n}$$

  Where x_1, x_2 and x_3 are measured from $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3$ without loss of generality.
- Measure of the standard error of estimate $\hat{x}_1$ is the deviation of observed value x_1 from the estimated value $\hat{x}_1$.

9.12 UNIT END ACTIVITY

In a city, held over two days the following performance was recorded in the high jump and long jump. All distances are in meres.

Competitors	A	B	C	D	E	F	G
High Jump (x)	1.90	1.85	1.96	1.88	1.88	?	1.92
Long Jump (y)	6.22	6.24	6.50	6.36	6.32	6.44	?

What performance might have been expected from F in the high jump and G in the long jump if they had completed.

9.13 KEYWORDS

Regression: Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

Regression Analysis: Regression analysis is a branch of statistical theory which is widely used in all the scientific disciplines. It is a basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life.

Intrinsically Linear: Non-linear models that can be transformed to yield linear models are called intrinsically linear.

Coefficient of Determination: It is defined as the ratio of explained variance of the dependent variable to the total variance. It can be shown that this measure is equal to the square of the correlation coefficient.

Coefficient of Alienation: It is square root of coefficient of non-determination.

Variable: A variable is an alphabetic character representing a number, called the **value** of the variable, which is either arbitrary or not fully specified or unknown.

Dependent Variable: The "dependent variable" represents the output or effect, or is tested to see if it is the effect.

Independent Variable: The "independent variables" represent the inputs or causes, or are tested to see if they are the cause.

Multiple Regression: Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).

***Standard error of estimate*:** Measure of the standard error of estimate $\hat{x}_1$ is the deviation of observed value x_1 from the estimated value $\hat{x}_1$.

9.14 QUESTIONS FOR DISCUSSION

1. Define the term regression.
2. Distinguish between correlation and regression.
3. Discuss least square method of fitting regression.
4. What do you understand by linear regression?
5. Why there are two lines of regression?
6. Under what condition(s) can there be only one line?

7. What is the method of least squares?
8. The two regression coefficients obtained by a student are 2.58 and 0.48. Comment.
9. The coefficient of correlation between X and Y is 0.85 and one of the regression coefficient is – 0.21. Comment.
10. The two regression coefficients are 1.5 and 0.6 and the coefficient of correlation is 0.90. Comment.
11. The two regression coefficients are –2.7 and –0.3 and the coefficient of correlation is 0.90. Comment.
12. What are regression coefficients?
13. The two regression coefficients are greater than unity. Comment.
14. From the following data, obtain the two regression equations and calculate the coefficient of correlation there from :

Sales	:	91	97	103	121	67	124	52	73	111	57
Purchases	:	97	75	69	97	70	91	39	61	83	47

15. Two random variables have the following regression equations:

    $3X + 2Y - 26 = 0$ and $6X + Y - 31 = 0$

    Find the means and the correlation coefficient between X and Y. Also find sY if the variance of X is 25.

16. For 10 observations on price (p) and supply (s), the following data (in appropriate units) were obtained:

    $\Sigma p = 70 \quad \Sigma s = 90 \quad \Sigma p^2 = 514 \quad \Sigma s^2 = 852 \quad \Sigma ps = 660 \quad N = 10$

    Obtain (i) the regression line of s on p, and (ii) estimate the supply when price is 16 units.

17. For a given set of bivariate data, the following results were obtained:

    $bY.X = -1.5$ and $bX.Y = -0.2$

    Find the most probable value of Y when $X = 60$

18. The correlation coefficient between the number of railway accidents and the number of babies born in a year was found to be 0.8. Comment
19. In a bivariate distribution, the two regression coefficients are – 0.5 and 0.9. Comment
20. In a bivariate distribution, the two regression coefficients are – 0.98 and – 1.52. Comment
21. Define the regression of Y on X and of X on Y for a bivariate data (X_i, Y_i), i = 1, 2, ... n. What would be the values of the coefficient of correlation if the two regression lines (a) intersect at right angle and (b) coincide?
22. Show that the proportion of variations explained by a regression equation is r^2.
23. What is the relation between Total Sum of Squares (TSS)? Explained Sum of Squares (ESS) and Residual Sum of Squares (RSS). Use this relationship to prove that the coefficient of correlation has a value between –1 and +1.
24. The regression line gives only a 'best estimate' of the quantity in question. We may assess the degree of uncertainty in this estimate by calculating its standard error. Explain.

25. Given a scatter diagram of a bivariate data involving two variables X and Y. Find the conditions of minimisation of and hence derive the normal equations for the linear regression of Y on X. What sum is to be minimised when X is regressed on Y? Write down the normal equation in this case.
26. Explain, fully, the meaning of regression of one variable Y on another variable X. Discuss the method of least squares for fitting a linear regression of the form $Y = a + bX$. Write down the normal equations and show that $b = r \cdot \frac{\sigma_Y}{\sigma_X}$, where the symbols have their usual meaning.
27. Show that the coefficient of correlation is the geometric mean of the two regression coefficients.
28. Show that the two lines of regression obtained by this method are irreversible except when $r = \pm 1$.
29. The correlation coefficient between two variables X and Y is $r = 0.60$. If find the equations of the two regression lines. Plot these equations on a graph.
30. For a certain X and Y series, where X and Y are correlated, the two lines of regression are $5X - 6Y + 90 = 0$ and $15X - 8Y - 130 = 0$.

    Find which line is regression of Y on X and which is of regression of X on Y? Find the means of the two series and the coefficient of correlation between them.
31. What are the formulae for regression coefficients for multiple regression?
32. How will you derive regression coefficients from regression equations in case of multiple regression?
33. How do you analyse multiple regression using MS Excel?
34. How will you express regression equation in terms of correlation coefficients?

Check Your Progress: Model Answer

1. Correlation
2. Sir Francis Galton
3. $Y = a + bX + e$
4. Errors
5. S
6. Coefficient of determination

9.15 REFERENCE & SUGGESTED READINGS

- Bowerman, B. L., O'Connell, R. T., Murphree, E. S., & Orris, J. B. (2018). **Essentials of Business Statistics** (6th ed.). McGraw-Hill Education. ISBN: 9781259549939
- Doane, D. P., & Seward, L. E. (2019). **Applied Statistics in Business and Economics** (6th ed.). McGraw-Hill Education. ISBN: 9781260224035
- Gupta, S. C., & Kapoor, V. K. (2018). **Fundamentals of Applied Statistics** (4th ed.). Sultan Chand & Sons. ISBN: 9788180547967
- Keller, G. (2022). **Business Analytics: A Data-Driven Decision Making Approach** (1st ed.). Cengage Learning. ISBN: 9780357717828
- Evans, J. R. (2020). **Business Analytics: Methods, Models, and Decisions** (2nd ed.). Pearson. ISBN: 9780135231679

BLOCK - 5

UNIT-X

INDEX NUMBER

CONTENTS

Contd...

10.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Define the term index number
- Discuss the features and uses of an index number
- Understand notations and terminologies used in index numbers
- Establish relation between weighted aggregative and weighted arithmetic average of price relatives index numbers
- Make comparison of Laspeyres' and Paasche's index numbers
- Tell about chain base index numbers
- Explain calculation of Inflation
- Discuss the concept of Stock Market Index
- Describe the problems in the construction of index numbers

10.1 INTRODUCTION

An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations. Suppose that we want to compare the average price level of different items of food in 1992 with what it was in 1990. Let the different items of food be wheat, rice, milk, eggs, ghee, sugar, pulses, etc. If the prices of all these items change in the same ratio and in the same direction; assume that prices of all the items have increased by 10% in

1992 as compared with their prices in 1990; then there will be no difficulty in finding out the average change in price level for the group as a whole. Obviously, the average price level of all the items taken as a group will also be 10% higher in 1992 as compared with prices of 1990. However, in real situations, neither the prices of all the items change in the same ratio nor in the same direction, i.e., the prices of some commodities may change to a greater extent as compared to prices of other commodities. Moreover, the price of some commodities may rise while that of others may fall. For such situations, the index numbers are very useful device for measuring the average change in prices or any other characteristics like quantity, value, etc., for the group as a whole.

The data on the population of a nation is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

10.2 DEFINITIONS AND CHARACTERISTICS OF INDEX NUMBERS

Some important definitions of index numbers are given below:

"An index number is a device for comparing the general level of magnitude of a group of distinct, but related, variables in two or more situations."

—Karmel and Polasek

"An index number is a special type of average that provides a measurement of relative changes from time to time or from place to place."

—Wessell, Wilett and Simone

"Index number shows by its variation the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice."

—Edgeworth

"An index number is a single ratio (usually in percentage) which measures the combined (i.e., averaged) change of several variables between two different times, places or situations."

—Tuttle

On the basis of the above definitions, the following characteristics of index numbers are worth mentioning:

- ***Index numbers are specialised averages:*** As we know that an average of data is its representative summary figure. In a similar way, an index number is also an average, often a weighted average, computed for a group. It is called a specialised average because the figures, that are averaged, are not necessarily expressed in homogeneous units.
- ***Index numbers measure the changes for a group which are not capable of being directly measured:*** The examples of such magnitudes are: Price level of a group of items, level of business activity in a market, level of industrial or agricultural output in an economy, etc.
- ***Index numbers are expressed in terms of percentages:*** The changes in magnitude of a group are expressed in terms of percentages which are independent of the units of measurement. This facilitates the comparison of two or more index numbers in different situations. Index numbers are specialized type of averages that are used to measure the changes in characteristics which is not capable of being directly measured. For example, it is not possible to measure business

activity in a direct way, however, relative changes in a business activity can be determined by the direct measurement of changes in some factors that affect it. Similarly, it is not possible to measure, directly, the price level of a group of items, but changes in price level can be measured by using price index numbers.

10.3 USES OF INDEX NUMBERS

The main uses of index numbers are mentioned below:

- ***To measure and compare changes:*** The basic purpose of the construction of an index number is to measure the level of activity of phenomena like price level, cost of living, level of agricultural production, level of business activity, etc. It is because of this reason that sometimes index numbers are termed as barometers of economic activity. It may be mentioned here that a barometer is an instrument which is used to measure atmospheric pressure in physics. The level of an activity can be expressed in terms of index numbers at different points of time or for different places at a particular point of time. These index numbers can be easily compared to determine the trend of the level of an activity over a period of time or with reference to different places.
- ***To help in providing guidelines for framing suitable policies:*** Index numbers are indispensable tools for the management of any government or non-government organisation. For example, the increase in cost of living index is helpful in deciding the amount of additional dearness allowance that should be paid to the workers to compensate them for the rise in prices. In addition to this, index numbers can be used in planning and formulation of various government and business policies.
- ***Price index numbers are used in deflating:*** This is a very important use of price index numbers. These index numbers can be used to adjust monetary figures of various periods for changes in prices. For example, the figure of national income of a country is computed on the basis of the prices of the year in question. Such figures, for various years often known as national income at current prices, do not reveal the real change in the level of production of goods and services. In order to know the real change in national income, these figures must be adjusted for price changes in various years. Such adjustments are possible only by the use of price index numbers and the process of adjustment, in a situation of rising prices, is known as deflating.
- ***To measure purchasing power of money:*** We know that there is inverse relation between the purchasing power of money and the general price level measured in terms of a price index number. Thus, reciprocal of the relevant price index can be taken as a measure of the purchasing power of money.

10.4 CONSTRUCTION OF INDEX NUMBERS

To illustrate the construction of an index number, we reconsider various items of food mentioned earlier.

Let the prices of different items in the two years, 1990 and 1992, be as given below:

Item	Price in 1990 (in ₹/unit)	Price in 1992 (in ₹/unit)
1. Wheat	300/quintal	360/quintal
2. Rice	12/kg.	15/kg.
3. Milk	7/litre	8/litre
4. Eggs	11/dozen	12/dozen
5. Ghee	80/kg.	88/kg.
6. Sugar	9/kg.	10/kg.
7. Pulses	14/kg.	16/kg.

The comparison of price of an item, say wheat, in 1992 with its price in 1990 can be done in two ways, explained below:

1. By taking the difference of prices in the two years, i.e., 360 – 300 = 60, one can say that the price of wheat has gone up by ₹ 60/quintal in 1992 as compared with its price in 1990.

2. By taking the ratio of the two prices, i.e., $\frac{360}{300} = 1.20$ one can say that if the price of wheat in 1990 is taken to be 1, then it has become 1.20 in 1992. A more convenient way of comparing the two prices is to express the price ratio in terms of percentage, i.e. $\frac{360}{300} \times 100 = 120$ known as Price Relative of the item. In our example, price relative of wheat is 120 which can be interpreted as the price of wheat in 1992 when its price in 1990 is taken as 100. Further, the figure 120 indicates that price of wheat has gone up by 120 – 100 = 20% in 1992 as compared with its price in 1990.

The first way of expressing the price change is inconvenient because the change in price depends upon the units in which it is quoted. This problem is taken care of in the second method, where price change is expressed in terms of percentage. An additional advantage of this method is that various price changes, expressed in percentage, are comparable. Further, it is very easy to grasp the 20% increase in price rather than the increase expressed as ₹ 60/quintal.

For the construction of index number, we have to obtain the average price change for the group in 1992, usually termed as the Current Year, as compared with the price of 1990, usually called the Base Year. This comparison can be done in two ways:

1. By taking suitable average of price relatives of different items. The methods of index number construction based on this procedure are termed as Average of Price Relative Methods.

2. By taking ratio of the averages of the prices of different items in each year. These methods are popularly known as Aggregative Methods.

Since the average in each of the above methods can be simple or weighted, these can further be divided as simple or weighted.

Various methods of index number construction can be classified as shown below:

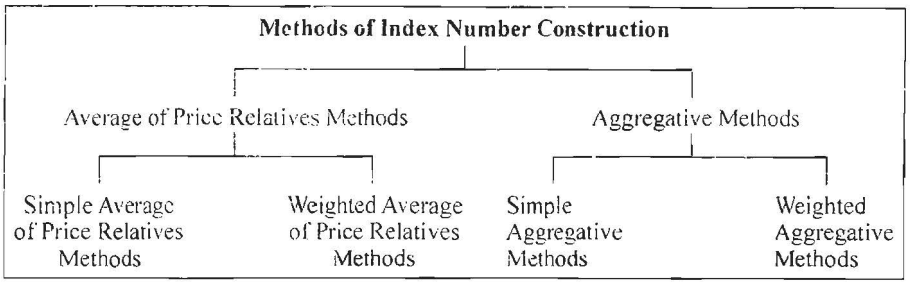


Figure 10.1: Methods of Index Number Construction

In addition to this, a particular method would depend upon the type of average used. Although, geometric mean is more suitable for averaging ratios, arithmetic mean is often preferred because of its simplicity with regard to computations and interpretation.

10.5 NOTATIONS AND TERMINOLOGY

Before writing various formulae of index numbers, it is necessary to introduce certain notations and terminology for convenience.

Base Year: The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Current Year: The year under consideration for which the comparisons are to be computed is called the current year. It is commonly denoted by writing '1' as a subscript of the variable.

Let there be n items in a group which are numbered from 1 to n. Let p_0^i denote the price of the i^{th} item in base year and p_1^i denote its price in current year, where i = 1, 2, ... n. In a similar way, q_0^i and q_1^i will denote the quantities of the i^{th} item in base and current years respectively.

Using these notations, we can write an expression for price relative of the i^{th} item as $P_i = p_{1i}/p_{0i} \times 100$ and quantity relative of the i^{th} item as $Q_i = q_{1i}/q_{0i} \times 100$.

Further, P_{01} will be used to denote the price index number of period '1' as compared with the prices of period '0'. Similarly, Q_{01} and V_{01} would denote the quantity and the value index numbers respectively of period '1' as compared with period '0'.

10.6 PRICE INDEX NUMBERS

The following are the Price Index Numbers:

10.6.1 Simple Average of Price Relatives

1. When arithmetic mean of price relatives is used

   The index number formula is given by

$$P_{01} = \frac{\sum P_i}{n} \quad \text{or} \quad P_{01} = \frac{\sum \frac{p_{1i}}{p_{0i}} \times 100}{n}$$

   Omitting the subscript i, the above formula can also be written as

$$P_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{n}.$$

2. When geometric mean of price relatives is used

The index number formula is given by

$$P_{01} = \left(P_1 \times P_2 \times ..... P_n \right)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} P_i \right)^{\frac{1}{n}} = \text{Antilog} \left[\frac{\sum \log P_i}{n} \right]$$

($\prod$ is used to denote the product of terms.)

Example: Given below are the prices of 5 items in 1985 and 1990. Compute the simple price index number of 1990 taking 1985 as base year. Use (a) arithmetic mean and (b) geometric mean.

Item	Price in1985 (₹/unit)	Price in1990 (₹/unit)
1.	15	20
2.	8	7
3.	200	300
4.	60	110
5.	100	130

Solution:

Calculation Table

Item	Price in 1985 (P_{0i})	Price in 1990 (P_{0i})	Price Relative $P_i = \frac{P_{1i}}{P_{0i}} \times 100$	log P_i
1.	15	20	133.33	2.1249
2.	8	7	87.50	1.9420
3.	200	300	150.00	2.1761
4.	60	110	183.33	2.2632
5.	100	130	130.00	2.1139
Total			**684.16**	**10.6201**

Index number, using A.M., is $P_{01} = \frac{684.16}{5} = 136.83$ and Index number, using G.M., is

$$P_{01} = \text{Antilog} \left[\frac{10.6201}{5} \right] = 133.06$$

10.6.2 Weighted Average of Price Relatives

In the method of simple average of price relatives, all the items are assumed to be of equal importance in the group. However, in most of the real life situations, different items of a group have different degree of importance. In order to take this into account, weighing of different items, in proportion to their degree of importance, becomes necessary.

Let w_i be the weight assigned to the i th item (i = 1, 2, ... n). Thus, the index number, given by the weighted arithmetic mean of price relatives, is $P_{01} = \frac{\sum P_i w_i}{\sum w_i}$.

Similarly, the index number, given by the weighted geometric mean of price relatives can be written as follows:

$$P_{01} = \left[P_1^{w_1} . P_2^{w_2} \quad P_n^{w_n} \right]^{\frac{1}{\sum w_i}} = \left[\prod P_i^{w_i} \right]^{\frac{1}{\sum w_i}} \text{ or } P_{01} = \text{Antilog} \left[\frac{\sum w_i \log P_i}{\sum w_i} \right]$$

Nature of Weights

While taking weighted average of price relatives, the values are often taken as weights. These weights can be the values of base year quantities valued at base year prices, i.e., $p_{0i}q_{0i}$, or the values of current year quantities valued at current year prices, i.e., $p_{1i}q_{1i}$, or the values of current year quantities valued at base year prices, i.e., $p_{0i}q_{1i}$, etc., or any other value.

Example: Construct an index number for 2010 taking 2002 as base for the following data, by using

1. Weighted arithmetic mean of price relatives and
2. Weighted geometric mean of price relatives.

Commodities	Prices in 2002	Prices in 2010	Weights
A	60	100	30
B	20	20	20
C	40	60	24
D	100	120	30
E	120	80	10

Solution:

Calculation Table

Commodities	Prices in 2002 (p_0)	Prices in 2010 (p_1)	P.R. (P) $= \frac{P_1}{P_0} \times 100$	Wts (w)	Pw	log P	w log P
A	60	100	166.67	30	5000.1	2.2219	66.657
B	20	20	100.00	20	2000.0	2.0000	40.000
C	40	60	150.00	24	3600.0	2.1761	52.226
D	100	120	120.00	30	3600.0	2.0792	62.376
E	120	80	66.67	10	666.7	1.8239	18.239
Total				**114**	**14866.8**		**239.48**

$\therefore$ Index number using A.M. is $P_{01} = \frac{14866.8}{114} = 130.41$

and index number using G.M. is $P_{01} = \text{Anti log} \left[\frac{239.498}{114} \right] = 126.15$

10.6.3 Simple Aggregative Method

In this method, the simple arithmetic mean of the prices of all the items of the group for the current as well as for the base year is computed separately. The ratio of current year average to base year average multiplied by 100 gives the required index number.

Using notations, the arithmetic mean of prices of n items in current year is given by $\frac{\sum p_{1i}}{n}$ and the arithmetic mean of prices in base year is given by $\frac{\sum p_{0i}}{n}$.

Simple aggregative price index $P_{01} = \frac{\frac{\sum p_{1i}}{n}}{\frac{\sum p_{0i}}{n}} \times 100 = \frac{\sum p_{1i}}{\sum p_{0i}} \times 100$

Omitting the subscript i, the above index number can also be written as $P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$

Example: The following table gives the prices of six items in the years 2010 and 2011. Use simple aggregative method to find index of 2011 with 2010 as base.

Item	Price in 2010 (₹)	Price in 2011 (₹)
A	40	50
B	60	60
C	20	30
D	50	70
E	80	90
F	100	100

Solution:

Let p_0 be the price in 2010 and p_1 be the price in 2011. Thus, we have

$$\Sigma p_0 = 350 \text{ and } \Sigma p_1 = 400$$

$$\therefore \quad P_{01} = \frac{400}{350} \times 100 = 114.29$$

10.6.4 Weighted Aggregative Method

This index number is defined as the ratio of the weighted arithmetic means of current to base year prices multiplied by 100.

Using the notations, defined earlier, the weighted arithmetic mean of current year prices can be written as $= \frac{\sum p_{1i} w_i}{\sum w_i}$

Similarly, the weighted arithmetic mean of base year prices $= \frac{\sum p_{0i} w_i}{\sum w_i}$

Price Index Number, $P_{01} = \frac{\frac{\sum p_{1i} w_i}{\sum w_i}}{\frac{\sum p_{0i} w_i}{\sum w_i}} \times 100 = \frac{\sum p_{1i} w_i}{\sum p_{0i} w_i} \times 100$

Omitting the subscript, we can also write $P_{01} = \frac{\sum p_1 w}{\sum p_0 w} \times 100$

Nature of Weights

In case of weighted aggregative price index numbers, quantities are often taken as weights. These quantities can be the quantities purchased in base year or in current year or an average of base year and current year quantities or any other quantities. Depending upon the choice of weights, some of the popular formulae for weighted index numbers can be written as follows:

1. ***Laspeyres's Index:*** Laspeyres' price index number uses base year quantities as weights. Thus, we can write

$$P_{01}^{La} = \frac{\sum p_{1i} q_{0i}}{\sum p_{0i} q_{0i}} \times 100 \quad \text{or} \quad P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

2. ***Paasche's Index:*** This index number uses current year quantities as weights. Thus, we can write

$$P_{01}^{Pa} = \frac{\sum p_{1i} q_{1i}}{\sum p_{0i} q_{1i}} \times 100 \quad \text{or} \quad P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

3. ***Fisher's Ideal Index:*** As will be discussed later that the Laspeyres's Index has an upward bias and the Paasche's Index has a downward bias. In view of this, Fisher suggested that an ideal index should be the geometric mean of Laspeyres' and Paasche's indices. Thus, the Fisher's formula can be written as follows:

$$P_{01}^{F} = \sqrt{P_{01}^{La} \times P_{01}^{Pa}} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

If we write $L = \frac{\sum p_1 q_0}{\sum p_0 q_0}$ and $P = \frac{\sum p_1 q_1}{\sum p_0 q_1}$, the Fisher's Ideal Index can also be written as $P_{01} = \sqrt{L \times P} \times 100$.

4. ***Dorbish and Bowley's Index:*** This index number is constructed by taking the arithmetic mean of the Laspeyres's and Paasche's indices.

$$P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 + \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100\right]$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100 = \frac{1}{2}\,[L \times P] \times 100$$

5. ***Marshall and Edgeworth's Index:*** This index number uses arithmetic mean of base and current year quantities.

$$P_{01}^{ME} = \frac{\sum p_1\left(\frac{q_0 + q_1}{2}\right)}{\sum p_0\left(\frac{q_0 + q_1}{2}\right)} \times 100 = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100 = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

6. ***Walsh's Index:*** Geometric mean of base and current year quantities are used as weights in this index number.

$$P_{01}^{Wa} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$$

7. ***Kelly's Fixed Weights Aggregative Index:*** The weights, in this index number, are quantities which may not necessarily relate to base or current year. The weights, once decided, remain fixed for all periods. The main advantage of this index over Laspeyres's index is that weights do not change with change of base year. Using symbols, the Kelly's Index can be written as

$$P_{01}^{Ke} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

Example: Calculate the weighted aggregative price index for 1990 from the following data:

Item	Price in 1971	Price in 1990	Weights
A	8	9.5	5
B	12	12.5	1
C	6.5	9	3
D	4	4.5	6
E	6	7	4
F	2	4	3

Solution:

Calculation Table

Item	Price in 1971 (p_0)	Price in 1990 (p_1)	Weights (w)	p_0w	p_1w
A	8	9.5	5	40.0	47.5
B	12	12.5	1	12.0	12.5
C	6.5	9	3	19.5	27.0
D	4	4.5	6	24.0	27.0
E	6	7	4	24.0	28.0
F	2	4	3	6.0	12.0
Total				**125.5**	**154.0**

Price Index (1971 = 100) $P_{01} = \frac{154.0}{125.5} \times 100 = 122.71$

The term within bracket, i.e., 1971 = 100, indicates that base year is 1971.

Example: For the data given in the following table, compute

Calculation Table

Commodity	p_0	q_0	p_1	q_1
A	10	30	12	50
B	8	15	10	25
C	6	20	6	30
D	4	10	6	20

(i) Laspeyres's Price Index

(ii) Paasche's Price Index

(iii) Fisher's Ideal Index

(iv) Dorbish and Bowley's Price Index

(v) Marshall and Edgeworth's Price Index

(vi) Walsh's Price Index

Solution:

The calculation of various price index numbers are done as given below:

(i) $P_{01}^{La} = \frac{690}{580} = 118.97$

(ii) $P_{01}^{Pa} = \frac{1150}{960} \times 100 = 119.79$

(iii) $P_{01}^{Fi} = \sqrt{\frac{690}{580} \times \frac{1150}{960}} \times 100 = 119.38$

(iv) $P_{01}^{DB} = \frac{1}{2}\left[\frac{690}{580} + \frac{1150}{960}\right] \times 100 = 119.4$

(v) $P_{01}^{ME} = \frac{690+1150}{580+960} \times 100 = 119.48$

(vi) $P_{01}^{Wa} = \frac{890.1}{745.7} \times 100 = 119.36$

Example: Construct index numbers for the following data by taking (a) price of 1975 as base, and (b) average of all the prices as base.

Solution:

Calculation Table

Years	Price	Index (1975 = 100)	Index Base = 168.75*
1975	110	100	$\frac{110}{168.75} \times 100 = 65.2$
1976	120	$\frac{120}{110} \times 100 = 109.1$	$\frac{120}{168.75} \times 100 = 71.1$
1977	160	$\frac{160}{110} \times 100 = 145.5$	$\frac{160}{168.75} \times 100 = 94.8$
1978	150	$\frac{150}{110} \times 100 = 136.4$	$\frac{150}{168.75} \times 100 = 88.9$
1979	180	$\frac{180}{110} \times 100 = 163.6$	$\frac{180}{168.75} \times 100 = 106.7$
1980	200	$\frac{200}{110} \times 100 = 181,.8$	$\frac{200}{168.75} \times 100 = 118.5$
1981	220	$\frac{220}{110} \times 100 = 200.0$	$\frac{220}{168.75} \times 100 = 130.4$
1982	210	$\frac{210}{110} \times 100 = 190.9$	$\frac{210}{168.75} \times 100 = 124.4$

Example: Taking average of the prices (₹/quintal) as base, construct price index numbers for the three years from the following data:

Years	Price of Wheat (₹)	Price of Rice (₹)	Price of Sugar (₹)
1	100	400	600
2	140	520	660
3	180	580	840

Solution:

$$\text{Average price of wheat} = \frac{100+140+180}{3} = \frac{420}{3} = 140$$

$$\text{Average price of rice} = \frac{400+520+580}{3} = \frac{1500}{3} = 500$$

$$\text{Average price of sugar} = \frac{600+660+840}{3} = \frac{2100}{3} = 700$$

Calculation Table

Price Relatives with base = Average price

Wheat	Rice	Sugar	Total	Index No. = $\frac{\text{Total}}{3}$
$\frac{100}{140}\times 100 = 71.4$	$\frac{400}{500}\times 100 = 80$	$\frac{600}{700}\times 100 = 85.7$	237.1	79.0
$\frac{140}{140}\times 100 = 100$	$\frac{520}{500}\times 100 = 104$	$\frac{660}{700}\times 100 = 94.3$	298.3	99.4
$\frac{180}{140}\times 100 = 128.6$	$\frac{580}{500}\times 100 = 116$	$\frac{840}{700}\times 100 = 120$	364.6	121.5

Example: From the following data, construct price index numbers by using the following formulae:

1. Simple aggregative index.
2. Simple arithmetic average of price relatives index.
3. Weighted aggregative index with average quantities as weights.
4. Weighted arithmetic average of price relatives with expenditure in base year as weights.

Items	Price (p_0)	Quantity (q_0)	Price (p_1)	Quantity (q_1)
A	10	12	12	15
B	7	15	5	20
C	5	24	9	20
D	16	5	14	5

Solution:

Calculation Table

Items	p_0	q_0	p_1	q_1	P	p_0q_0(w)	p_0q_1	p_1q_0	p_1q_1	Pw
A	10	12	12	15	120.0	120	150	144	180	14400
B	7	15	5	20	71.4	105	140	75	100	7497
C	5	24	9	20	180.0	120	100	216	180	21600
D	16	5	14	5	87.5	80	80	70	70	7000
	38		**40**		**458.9**	**425**	**470**	**505**	**530**	**50497**

1. Simple aggregative index $P_{01} = \frac{\sum p_1}{\sum p_0}\times 100 = \frac{40}{38}\times 100 = 105.3$

2. Simple A.M. of price relatives index $P_{01} = \frac{\sum P}{n} = \frac{458.3}{4} = 114.6$

3. Required weighted aggregative index

$$P_{01}^{ME} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100 = \frac{505 + 530}{425 + 470} \times 100 = 115.6$$

4. Weighted A.M. of price relatives index $P_{01} = \frac{\sum Pw}{\sum w} = \frac{50497}{425} = 118.8$

10.7 QUANTITY INDEX NUMBERS

A quantity index number measures the change in quantities in current year as compared with a base year. The formulae for quantity index numbers can be directly written from price index numbers simply by interchanging the role of price and quantity. Similar to a price relative, we can define a quantity relative as

$$Q = \frac{q_1}{q_0} \times 100$$

Various formulae for quantity index numbers are as given below:

(i) Simple aggregative index $Q_{01} = \frac{\sum q_1}{\sum q_0} \times 100$

(ii) Simple average of quantity relatives

(a) Taking A.M. $Q_{01} = \frac{\frac{\sum q_1}{\sum q_0} \times 100}{n} = \frac{\sum Q}{n}$

(b) Taking G.M. $Q_{01} = \text{Antilog}\left[\frac{\sum \log Q}{n}\right]$

(iii) Weighted aggregative index

(a) $Q_{01}^{La} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$ (Base year prices are taken as weights)

(b) $Q_{01}^{Pa} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$ (Current year prices are taken as weights)

(c) $Q_{01}^{Fi} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$ Other aggregative formulae can also be written in a similar way.

(iv) Weighted average of quantity relatives

(a) Taking A.M. $Q_{01} = \frac{\sum Qw}{\sum w}$

(b) Taking G.M. $Q_{01} = \text{Antilog}\left[\frac{\sum w \log Q}{\sum w}\right]$

Like weighted average of price relatives, values are taken as weights.

Example: Using Fisher's formula, the quantity index number from the following data:

Article	1974		1976	
	Price (₹)	Value (₹)	Price (₹)	Value (₹)
A	5	50	4	48
B	8	48	7	49
C	6	18	5	20

Solution:

Calculation Table

Article	1974			1976			p_0q_1	p_1q_0
	p_0	V_0	$q_0=\frac{V_0}{p_0}$	p_1	V_1	$q_1=\frac{V_1}{p_1}$		
A	5	50	10	4	48	12	60	40
B	8	48	6	7	49	7	56	42
C	6	18	3	5	20	4	24	15
Total	$\Sigma p_0q_0=116$			$\Sigma p_1q_1=117$			140	97

$$Q_{01}^{F_1}=\sqrt{\frac{\sum q_1p_0}{\sum q_0p_0}\times\frac{\sum q_1p_1}{\sum q_0p_1}}\times 100=\sqrt{\frac{140}{116}\times\frac{117}{97}}=120.65$$

10.8 VALUE INDEX NUMBER

A value index number gives the change in value in current period as compared with base period. The value index, denoted by V_{01}, is given by the formula $V_{01}=\frac{\sum p_1q_1}{\sum p_0q_0}\times 100$. The value index for the data given in example is given by $V_{01}=\frac{117}{116}\times 100=100.86$.

10.9 TESTS OF ADEQUACY OF INDEX NUMBER FORMULAE

A number of formulae have been given for the construction of index numbers. However, each one of them suffers from one or the other type of drawbacks. It was, therefore, suggested that a satisfactory index number formula should satisfy certain mathematical criteria. These mathematical criteria, also known as tests of adequacy of index numbers, are given below:

1. ***Unit Test:*** This test requires that an index number formula should be independent of the units of measurement of the variables, i.e., its value should not change with change in units of measurement. This test is satisfied by all index numbers except simple aggregate formula.
2. ***Time Reversal Test:*** According to Fisher, "The formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as base or putting it in another way, the index number reckoned forward should be reciprocal of the one reckoned backward.

This test is devised on the analogy of a single commodity case. In case of a single commodity, if p_0 is its price in period '0' and p_1 is its price in period '1', then the change in price in period '1' as compared with the price of period '0' is given by $\frac{p_1}{p_0}$. Similarly, change in price in period '0' as compared with price of period '1' is given by $\frac{p_0}{p_1}$ which is reciprocal of $\frac{p_1}{p_0}$.

Generalising this to the case of more than one commodity the index number P_{01} measures the change in price in period '1' as compared with the price of period '0' and the index number P_{10} measures the change in price in period '0' as compared with the price of period '1'. By analogy with one commodity case, the time reversal test requires that the condition $P_{01} = \frac{1}{P_{10}}$ or $P_{01} \times P_{10} = 1$ should be satisfied.

Thus, according to time reversal test, the product of an index number with another index number, obtained by reversing of time, should be equal to unity.

The implication of time reversal test is that a formula which shows a rise of 25% (say) in period '1' as compared with period '0' must show a fall of 20% in period '0' as compared with period '1' because $\frac{125}{100} \times \frac{100}{125} = 1$ or $1.25 \times 0.80 = 1$. Here 1.25 implies a rise of 25% while 0.80 implies a fall of 20%.

This test is not satisfied by the Laspeyres's and Paasche's index numbers. This is obvious from below. We can write the Laspeyres's index as

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \qquad \text{(omitting multiplication by 100)}$$

On reversing the timings, i.e., replacing 1 by 0 and 0 by 1, we have $P_{10} = \frac{\sum p_0 q_1}{\sum p_1 q_1}$.

Since $P_{01} \times P_{10} \neq 1$, the Laspeyres's index does not satisfy time reversal test. Similarly, it can be shown that Paasche's index does not satisfy time reversal test.

This test is satisfied by Fisher's Ideal Index. We write

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \quad \text{and} \quad P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = 1$$

It can also be shown that Marshall-Edgeworth's, Walsh's and Kelly's indices also satisfy the time reversal test.

3. ***Factor Reversal Test:*** This test requires that the product of price index and the corresponding quantity index should be equal to the value index, i.e., $P_{01} \times Q_{01} = V_{01}$.

In the words of Fisher, *"Just as our formula should permit the interchange of two periods without giving inconsistent results, so it ought to permit interchange of prices and quantities without giving inconsistent results, i.e., the two results multiplied together should give the true value ratio."*

This test is also based on the analogy with a single commodity case. For a single commodity, if price change between two periods is $\frac{p_1}{p_0}$ and the quantity change is $\frac{q_1}{q_0}$, then Price Change × Quantity Change = $\frac{p_1}{p_0} \times \frac{q_1}{q_0} = \frac{p_1 q_1}{p_0 q_0}$, which denotes change in value between the two periods.

By analogy with the case of a single commodity, if P_{01} is the price index and Q_{01} is the quantity index between two periods '0' and '1', then we should have

$$P_{01} \times Q_{01} = V_{01}, \text{ where } V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

The implication of factor reversal test is that a formula which shows a price rise of 40% and a quantity fall of 20% (say) in current year as compared with base year should show a 12% increase of value in current year as compared with base year, i.e., 1.40 × 0.80 = 1.12.

This test is neither satisfied by Laspeyres's nor by the Paasche's index number. However, Fisher's ideal index satisfies this test.

To show that the test is satisfied by the Fisher's index, we write

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

Reversing the role of factors, we can write

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

The time reversal and factor reversal tests were suggested by Fisher and, consequently, he gave his ideal formula. This formula is ideal in the sense that it satisfies the above two tests. However, from practical point of view, the ideal index formula is not of much use. The main disadvantages with this index number are: (i) it is extremely cumbersome to calculate and (ii) its interpretation is ambiguous, i.e., we do not know what this index number measures.

In practice, the statisticians continue to rely upon simple, although less exact index number formulae, capable of unambiguous interpretation. From this point of view, the Laspeyres's and Paasche's index numbers are generally used, although, these neither satisfy time reversal nor factor reversal tests.

4. ***Circular Test:*** This test is an extension of the time reversal test. Let there be (t + 1) periods, denoted as 0, 1, 2, ... t respectively and we construct price index numbers P_{01}, P_{12}, ... $P_{t-1,t}$ and P_{t0} for each of the periods 1, 2, ... t and 0 respectively. This test requires that an index number formula should be such that the following condition is satisfied, i.e., $P_{01} \times P_{12} \times ..... \times P_{t-1,t} \times P_{t0} = 1$.

Since, according to time reversal test $P_{0t} \times P_{t0} = 1$, therefore, the above condition can also be written as $P_{01} \times P_{12} \times ..... \times P_{t-1,t} = P_{0t}$.

None of the popular formulae like Laspeyres's, Paasche's and Fisher's satisfy this test. The test is only satisfied by (a) simple aggregative index, (b) weighted

aggregative index with fixed weights (Kelly's index) and (c) simple geometric mean of price relatives index.

Example: Show that the Laspeyres's and Paasche's index numbers satisfy the time reversal and factor reversal tests in the most unlikely situation where either correlation between price and quantity movements is zero or prices (or quantities) of all the commodities change in the same ratio.

Solution:

The two index numbers will be equal under the conditions given in the question.

Thus, we have

$$\frac{P_{01}^{La}}{P_{01}^{Pa}} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} = 1 \qquad \ldots(1)$$

i.e., $P_{01}^{La} \times P_{10}^{La} = 1$

Thus, the Laspeyres's formula satisfies time reversal test.

Further, on considering the reciprocal of equation (1) we can show that the Paasche's formula also satisfies time reversal test.

To show that the Laspeyres's formula satisfies the factor reversal test, we multiply and divide left hand side of equation (1) by $\Sigma p_0 q_0$. Thus, we have

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_0 q_0} = 1$$

On adjusting terms, we can write $P_{01}^{La} \times Q_{01}^{La} = V_{01}$

Similarly, if we multiply and divide left hand side of equation (1) by $\Sigma p_1 q_1$, we can show that the factor reversal test is also satisfied by the Paasche's index.

Example: Using simple arithmetic average of price relatives, prepare two price index numbers from the following data:

(a) A price index for 1977 with 1970 as base.

(b) A price index for 1970 with 1977 as base.

Explain the paradoxical results. Suggest some suitable index number.

Commodity	Unit Price (in ₹) Year	
	1970	1977
A	1.00	0.50
B	0.30	0.60

Solution:

Let us denote 1970 by '0' and 1977 by '1'.

$$P_{01} = \frac{\frac{0.50}{1.00} \times 100 + \frac{0.60}{0.30} \times 100}{2} = 125$$

This shows that prices have gone up by 25% in 1977 as compared to the prices of 1970.

Further, $P_{10} = \dfrac{\dfrac{1.00}{0.50}\times 100 + \dfrac{0.30}{0.60}\times 100}{2} = 125$

This result shows that prices have gone up by 25% in 1970 as compared to the prices of 1977.

We note that the above two results are mutually contradictory. The basic reason for this is the use of arithmetic mean which is not a suitable average for ratios. To avoid this inconsistency, we should use geometric mean. Thus, we get

$$P_{01} = \left[\frac{0.50}{1.00}\times 100 \times \frac{0.60}{0.30}\times 100\right]^{\frac{1}{2}} = 100$$

and $$P_{10} = \left[\frac{1.00}{0.50}\times 100 \times \frac{0.30}{0.60}\times 100\right]^{\frac{1}{2}} = 100$$

Further, the inconsistency can be avoided if we consider simple aggregate formula, i.e.

$$P_{01} = \frac{1.10}{1.30}\times 100 = 84.62 \quad \text{and } P_{10} = \frac{1.30}{1.10}\times 100 = 118.18$$

Thus, the results of the above index numbers are consistent.

Example: A household spends its entire income on two goods A and B. The prices and the quantities purchased in the base and the current year are given below. Calculate the additional income to be given to the household so as to fully compensate it for the price rise, by using both the Laspeyres's and Paasche's index numbers.

Commodity	Basic Period		Current Period	
	Price	Quantity	Price	Quantity
A	30	10	40	8
B	5	20	8	10

Solution:

In order to find extra income needed to compensate for the rise of price, we shall compute price and quantity index numbers using each formula. The multiplication of price and the quantity index numbers will give value index. This index would indicate the additional income required to fully compensate the household for price rise. The subscripts 0 and 1, in the following table denote the base year and current year values.

Calculation Table

Comm.	P_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
A	30	10	40	8	300	400	240	320
B	5	20	8	10	100	160	50	80
Total					**400**	**560**	**290**	**400**

1. Calculation of additional income using Laspeyre's index.

   $P_{01}^{La} = \dfrac{560}{400} = 1.40$ (Note that 100 has been dropped here) and $Q_{01}^{La} = \dfrac{390}{400} = 0.725$,

   $P_{01} \times Q_{01} = 1.40 \times 0.75 = 1.015$

   This result shows that the income of the household should be 1.5% higher in current year. Since income in base year is ₹ 400, the additional income required to compensate is $400 \times \dfrac{1.5}{100} = ₹\ 6$

2. Calculation of additional income using Paasche's index.

$$P_{01}^{Pa} = \frac{400}{290} = 1.379 \quad \text{and} \quad Q_{01}^{Pa} = \frac{400}{560} = 0.714$$

$$\therefore P_{01} \times Q_{01} = 1.379 \times 0.714 = 0.985$$

This result shows that the income of the household should be decreased by 1.5%, i.e., by ₹ 6 in the current year as compared with base year.

Note: The value index is $V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{400}{400}$, which implies that the total expenditure of the household in base and current years are same and accordingly, they do not need any compensation. The reason for this discrepancy is that both the Laspeyres's and the Paasche's index numbers do not satisfy factor reversal test.

10.10 CHAIN BASE INDEX NUMBERS

So far, we have considered index numbers where comparisons of various periods were done with reference to a particular period, termed as base period. Such type of index number series is known as fixed base series. There are several examples of fixed base series like the series of index numbers of industrial production, agricultural production, wholesale prices, etc. The main problem with a fixed base series arises when the base year becomes too distant from the current year. In such a situation, it may happen that commodities which used to be very important in the base year are no longer so in current year. Furthermore, certain new commodities might be in use while some old commodities are dropped in current year. In short, this implies that the relative importance of various items is likely to change and, therefore, the comparison of a particular year with a remote base year may appear to be meaningless. A way out to this problem is to construct Chain Base Index Numbers, where current year is compared with its preceding year.

Similar to price relatives, here we define link relatives. A link relative of a commodity in a particular year is equal to the ratio of this year's price to last year's price multiplied by hundred. Using symbols, the link relative of i^{th} commodity in period t is written as $L_{ti} = \frac{p_{ti}}{p_{t-1i}} \times 100$.

When there are n commodities, the chain base index for period t is given by a suitable average of their link relatives. For example, taking simple arithmetic mean of link relatives we can write the chain base index as

$$P_t^{CB} = \frac{\sum L_t}{n} \times 100 = \frac{\sum \frac{p_t}{p_{t-1}} \times 100}{n} \qquad \ldots (1)$$

We may note here that a chain base index is equal to link relative of a commodity when there is only one commodity.

10.10.1 Chained Index Numbers

The chain base index numbers, obtained above, are as such of not much use because these have been computed with reference to a different base period and hence not comparable with each other. To avoid this difficulty, these are required to be chained to a common base period. The process of chaining is based upon the concept of circular test.

The expression for chained index for period 't' with '0' as base period, denoted as P_{0t}^{Ch}, can be written as

$$P_{01}^{Ch} = \frac{P_1^{CB}}{100} \times \frac{P_2^{CB}}{100} \times \ldots \times \frac{P_t^{CB}}{100} \times 100$$

or $$P_{0t}^{Ch} = \frac{P_t^{CB} \times P_{0(t-1)}^{Ch}}{100} \quad \left(P_{0(t-1)}^{Ch} = \frac{P_1^{CB}}{100} \times \quad \times \frac{P_{t-1}^{CB}}{100} \times 100 \right)$$

$$\text{Chained Index of current yr} = \frac{\text{CBI of current yr} \times \text{Chained Index of previous yr}}{100}$$

$$= \frac{\text{Average of Link Relatives of current yr} \times \text{Chained Index of previous yr}}{100} \quad \ldots (2)$$

Remarks:

1. When there is a single commodity, the chained index will be equal to the fixed base index, as shown below:

$$P_{0t}^{Ch} = \frac{p_1}{p_0} \times \frac{p_2}{p_1} \times \quad \times \frac{p_t}{p_{t-1}} \times 100 = \frac{p_t}{p_0} \times 100 = P_{0t}^{FB}$$

2. Theoretically, the value of chained index should be equal to the fixed base index even when there are more than one commodities. However, the actual values of these index numbers would differ from each other due to computations of average of link relatives. The magnitude of these differences would depend upon the suitability of an average in averaging link relatives. Often these differences are found to be very small.

3. In spite of the fact that a chained index number series is approximately equal to the fixed base index number series, the former is preferred to the latter because:

   (a) The difficulties regarding changes in relative importance of the items as well as that of changes in the composition of commodity bundle are avoided.

   (b) The chain base index numbers provide easy comparisons with its preceding year.

10.10.2 Conversion of Chain Base Index Number into Fixed Base Index Number and Vice-versa

We can write $P_t^{CB} = \frac{P_{0t}^{FB}}{P_{0t-1}^{FB}} \times 100$

$$\text{i.e., } \begin{pmatrix} \text{Chain Base Index} \\ \text{Number of Current} \\ \text{Year} \end{pmatrix} = \frac{\text{Fixed Base Index of Current Year}}{\text{Fixed Base Index of Previous Year}} \times 100 \quad \ldots (3)$$

$$\therefore \begin{pmatrix} \text{Fixed Base Index} \\ \text{of Current Year} \end{pmatrix} = \frac{\begin{pmatrix} \text{Chain Base Index} \\ \text{of Current Year} \end{pmatrix} \times \begin{pmatrix} \text{Fixed Base Index} \\ \text{of Previous Year} \end{pmatrix}}{100} \quad \ldots (4)$$

Example: From the following data on retail prices of wheat (in ₹/quintal), construct chain base index numbers. Also construct the chained index numbers and the fixed base index numbers with 1985 as base year.

Years	:	1985	1986	1987	1988	1989	1990
Prices	:	330	370	385	390	400	425

Solution:

Since it is a single commodity case, therefore, link relatives would be the chain base index numbers and the fixed base index numbers would be equal to the chained index numbers.

Year	Price of Wheat	Chain Base Index	Chained Index Number*
1985	330	100	100
1986	370	$\frac{370}{330} \times 100 = 112.1$	112.1
1987	385	$\frac{385}{370} \times 100 = 104.1$	$\frac{104.1 \times 112.1}{100} = 116.7$
1988	390	$\frac{390}{385} \times 100 = 101.3$	$\frac{101.3 \times 116.7}{100} = 118.7$
1989	400	$\frac{400}{390} \times 100 = 102.6$	$\frac{102.6 \times 118.2}{100} = 121.2$
1990	425	$\frac{425}{400} \times 100 = 106.3$	$\frac{106.3 \times 121.2}{100} = 128.8$

* Also equal to F.B.I

Example: From the following data, construct chain base index numbers:

Items	Years				
	1986	1987	1988	1989	1990
	Prices in ₹				
A	5	8	10	12	15
B	3	6	8	10	12
C	2	3	5	7	10.5

Solution:

Calculation of Chain Base Index Numbers

LR * → Items ↓	1986	1987	1988	1989	1990
A	100	$\frac{8}{5} \times 100 = 160$	$\frac{10}{8} \times 100 = 125$	$\frac{12}{10} \times 100 = 120$	$\frac{15}{12} \times 100 = 125$
B	100	$\frac{6}{3} \times 100 = 200$	$\frac{8}{6} \times 100 = 133.3$	$\frac{10}{8} \times 100 = 125$	$\frac{12}{10} \times 100 = 120$
C	100	$\frac{3}{2} \times 100 = 150$	$\frac{5}{3} \times 100 = 166.7$	$\frac{7}{5} \times 100 = 140$	$\frac{10.5}{7} \times 100 = 150$
Total	300	510	425.0	385	395
CBI	$\frac{300}{3} = 100$	$\frac{510}{3} = 170$	$\frac{425}{3} = 141.7$	$\frac{385}{3} = 128.3$	$\frac{395}{3} = 131.7$

*LR = Link Relatives

Example: From the following prices of three groups of commodities for the year 1983 to 1987, find chain base index numbers chained to 1983.

Groups	1983	1984	1985	1986
I	4	6	8	10
II	16	20	24	30
III	8	10	16	20

Solution:

Calculation Table

LR → Group ↓	1983	1984	1985	1986
I	100	$\frac{6}{4}\times 100 = 150$	$\frac{8}{6}\times 100 = 133.3$	$\frac{10}{8}\times 100 = 125$
II	100	$\frac{20}{16}\times 100 = 125$	$\frac{24}{20}\times 100 = 120$	$\frac{30}{24}\times 100 = 125$
III	100	$\frac{10}{8}\times 100 = 125$	$\frac{16}{10}\times 100 = 160$	$\frac{20}{16}\times 100 = 125$
Total	300	400	413.3	375
CBI	$\frac{300}{3} = 100$	$\frac{400}{3} = 133.3$	$\frac{413.3}{3} = 137.8$	$\frac{375}{3} = 125$
Chained Index	100	133.3	$\frac{137.8\times 133.3}{100} = 183.7$	$\frac{183.7\times 125}{100} = 229.6$

Example: Compute the chained base price index series with 1972-1973 = 100 for the following data:

	Commodity		1972-73	1973-74	1974-75	1975-76
(i)	Groundnut	Quantity (in'000 tonnes)	4092	5932	5111	6754
		Price (₹10000/1000 tonnes)	184	250	249	178
(ii)	Cotton	Quantity (in'000 bales)	5735	6309	7156	5950
		Price (₹10000/1000 bales)	123	182	166	142

Solution:

Calculation Table

Years → Comm. ↓	1972-73		1973-74		1974-75		1975-76	
	p_0	q_0	P_1	q_1	p_2	q_2	p_3	q_3
Groundnut	184	4092	250	5932	249	5111	178	6754
Cotton	123	5735	182	6309	166	7156	142	5950

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0}\times 100 = \frac{4092\times 250 + 5735\times 182}{4092\times 184 + 5735\times 123}\times 100 = 141.72$$

(using Laspeyres's formula)

$$P_{12} = \frac{5932\times 249 + 6309\times 166}{5932\times 250 + 6309\times 182}\times 100 = 95.94$$

$$P_{23} = \frac{5111\times 178 + 7156\times 142}{5111\times 249 + 7166\times 166}\times 100 = 78.27$$

$$P_{01}^{Ch} = 141.72,\ P_{02}^{Ch} = \frac{141.72\times 95.94}{100} = 135.97$$

and $P_{03}^{Ch} = \frac{135.97\times 78.27}{100} = 106.42$

10.11 BASE SHIFTING

Base shifting is needed either because the existing base has become too old and hence, useless for comparison or the given series is to be compared with another, having a different base year.

To shift the base year of a series, various index numbers are divided by the index number of the year selected for new base and the quotient, thus obtained, is multiplied by 100, *i.e.*,

$$\begin{pmatrix}\text{Index Number with}\\ \text{new base year}\end{pmatrix}=\frac{\text{Index Number with old base year}}{\text{Index Number of year selected as base year}}\times 100$$

The process of base shifting assumes that the index number formula satisfies, at least approximately, the circular test.

Example: Shift the base of the following series to 1987 and to 1989.

Years	:	1985	1986	1987	1988	1989	1990
Index No.	:	125	155	185	220	265	320

Solution:

Shifting of Base to 1987 and 1989

Years	Index No.	Index No. with 1987 = 100	Index No. with 1989 = 100
1985	125	$\frac{125}{185}\times 100=67.6$	$\frac{125}{265}\times 100=47.2$
1986	155	$\frac{155}{185}\times 100=83.8$	$\frac{155}{265}\times 100=58.5$
1987	185	$\frac{185}{185}\times 100=100$	$\frac{185}{265}\times 100=69.8$
1988	220	$\frac{220}{185}\times 100=118.9$	$\frac{220}{265}\times 100=83.0$
1989	265	$\frac{265}{185}\times 100=143.2$	$\frac{265}{265}\times 100=100$
1990	320	$\frac{320}{185}\times 100=173.0$	$\frac{320}{265}\times 100=120.8$

10.12 SPLICING

When the base year of an index number series becomes very old, it is usually discontinued and a new series with a recent past year as base is started. Sometimes, it may become necessary to combine these two series. The process of combining of two or more overlapping index number series is called splicing.

If there are two index number series, say 'A' or old series and 'B' or new series, then B can be spliced to A or vice-versa. When B is spliced to A the base of the spliced series would be same as that of series A. Similarly, when A is spliced to B, the base of the spliced series would be same as that of series B.

For splicing, there must be at least one year having index numbers of both the series. Normally, this year is the base year of the new series. Given this, we can find a correction factor for each of the two situations.

1. When B is spliced to A

   Correction factor =

$$\frac{\text{Index of Series A in the year corresponding to Base year of series B}}{100}$$

   All the index numbers of series B are multiplied by this correction factor to get the spliced series.

2. When A is spliced to B

   In this case, all the index numbers of series A are divided by the above correction factor to get the spliced series.

Example: Given below are the two index number series, one with 1981 as base and the other with 1989 as base.

Series A:	**Years** :	1981	1982	1983	1984	1985	1986	1987	1988	1989
	Index No. :	100	110	120	130	170	200	240	300	350

Series B:	**Years** :	1989	1990	1991	1992
	Index No. :	100	125	160	190

(a) Splice series B to series A (or series A forward)

(b) Splice series A to series B (or series B forward)

Solution:

The correction factor

Years	Series A	Series B	Series B spliced to series A	Series A spliced to series B
1981	100		100	$\frac{100}{3.5}=28.6$
1982	110		110	$\frac{110}{3.5}=31.4$
1983	120		120	$\frac{120}{3.5}=34.3$
1984	130		130	$\frac{130}{3.5}=37.1$
1985	170		170	$\frac{170}{3.5}=48.6$
1986	200		200	$\frac{200}{3.5}=57.1$
1987	240		240	$\frac{240}{3.5}=68.6$
1988	300		300	$\frac{300}{3.5}=85.7$
1989	350	100	350	100
1990		125	$125\times3.5=437.5$	125
1991		160	$160\times3.5=560$	160
1992		190	$190\times3.5=665$	190

10.13 USE OF PRICE INDEX NUMBERS IN DEFLATING

This is perhaps the most important application of price index numbers. Deflating implies making adjustments for price changes. A rise of price level implies a fall in the value of money. Therefore, in a situation of rising prices, the workers who are getting a fixed sum in the form of wages are in fact getting less real wages. Similarly,

in a situation of falling prices, the real wages of the workers are greater than their money wages. Thus, to determine the real wages, the money wages of the workers are to be adjusted for price changes by using relevant price index number.

The following formula is used for conversion of money wages into real wages:

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Consumer Price Index}} \times 100 \qquad \ldots(1)$$

Another application of the process of deflating to find the value of output at constant prices so as to facilitate the comparison of real changes in output. It may be pointed out here that the output of a given year is often valued at the current year prices. Since prices in various years are often different, the comparison of output at current year prices has no relevance.

The output at constant prices is obtained using the following formula:

$$\text{Output at Constant Prices} = \frac{\text{Output at Current Prices}}{\text{Price Index}} \times 100 \qquad \ldots(2)$$

Example: The following table gives the average monthly wages of a worker along with the respective consumer price index numbers for ten years.

Years	:	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Average monthly wages (₹)	:	500	525	560	600	630	635	700	740	800	900
Consumer Price Index	:	100	110	120	125	135	160	185	200	210	240

Compute his real average monthly wages in various years.

Solution:

Computation of Real Wages

Years	Average Monthly wage	Consumer Price Index	Real average monthly wage
1980	500	100	$\frac{500}{100} \times 100 = 500.00$
1981	525	110	$\frac{525}{110} \times 100 = 477.27$
1982	560	120	$\frac{560}{120} \times 100 = 466.67$
1983	600	125	$\frac{600}{125} \times 100 = 480.00$
1984	630	135	$\frac{630}{135} \times 100 = 466.67$
1985	635	160	$\frac{635}{160} \times 100 = 396.88$
1986	700	185	$\frac{700}{185} \times 100 = 378.38$
1987	740	200	$\frac{740}{200} \times 100 = 370.00$
1988	800	210	$\frac{800}{210} \times 100 = 380.95$
1989	900	240	$\frac{900}{240} \times 100 = 375.00$

10.13.1 Purchasing Power of Money

The concept of deflating can also be used to determine the purchasing power or real value of a rupee. When prices in general are rising, the real value of a rupee is declining. If, e.g., the price index in 1992 with base 1990 is 120, the real value of a rupee in 1992 as compared with its value in 1990 $= \frac{1}{120} \times 100 = 0.83$. This implies that a rupee in 1992 is worth only 83 paise of 1990.

From the above, we note that the purchasing power of a rupee in current year is equal to the reciprocal of the price index multiplied by 100. Thus, we can write

Purchasing Power of a Rupee or

$$\text{Constant Rupee} = \frac{\text{Current Rupee} \times 100}{\text{Price Index}} = \frac{100}{\text{Price Index}}$$

Note that the Current Rupee is always equal to unity.

Example: Given the following information on the Gross Domestic Product (in ₹ crores) at the constant (1980-81) prices and at current prices for five years. Calculate the series of price index numbers and of quantity index numbers for each of the five years with 1980-81 as base year.

	GDP at constant (1980-81) Prices	**GDP at current Prices**
1980-81	200	200
1981-82	150	240
1982-83	125	350
1983-84	120	360
1984-85	160	400

Solution:

Calculation of Price and Quantity Index Numbers

Years	**GDP at Constant Prices**	**GDP at Current Prices**	**Quantity Index Number Series**	**Price Index * Number Series**
1980-81	200	200	100	$\frac{200}{200} \times 100 = 100$
1981-82	150	240	$\frac{150}{200} \times 100 = 75$	$\frac{240}{150} \times 100 = 160$
1982-83	125	350	$\frac{125}{200} \times 100 = 62.5$	$\frac{350}{125} \times 100 = 280$
1983-84	120	360	$\frac{120}{200} \times 100 = 60$	$\frac{360}{120} \times 100 = 300$
1984-85	160	400	$\frac{160}{200} \times 100 = 80$	$\frac{400}{160} \times 100 = 250$

* Price Index $= \frac{\text{Output at Current Prices}}{\text{Output at Constant Prices}} \times 100$

Example: For the following data of a firm, construct the index of average wage and salaries at constant prices (Base year = 1980).

Years	:	1980	1981	1982	1983	1984
Average wages and salaries Paid (₹)	:	5000	5670	5865	6240	6820
Consumer Price Index	:	100	108	102	104	110

Solution:

Calculation Table

Years	Average Wages and Salaries Paid (₹)	Consumer Price Inbox	Wages and Salaries at Const. Prices	Index of Wages and Salaries (1980 = 100)
1980	5000	100	$\frac{5000}{100} \times 100 = 5000$	$\frac{5000}{5000} \times 100 = 100$
1981	5670	108	$\frac{5670}{108} \times 100 = 5250$	$\frac{5250}{5000} \times 100 = 105$
1982	5865	102	$\frac{5865}{102} \times 100 = 5750$	$\frac{5750}{5000} \times 100 = 115$
1983	6240	104	$\frac{6240}{104} \times 100 = 6000$	$\frac{6000}{5000} \times 100 = 120$
1984	6820	110	$\frac{6820}{110} \times 100 = 6200$	$\frac{6200}{5000} \times 100 = 124$

10.14 CONSUMER PRICE INDEX NUMBER

The consumer price or the retail price is the price at which the ultimate consumer purchases his goods and services from the retailer. According to the Labour Bureau, *"with the help of Consumer Price Index Number, it is intended to show over time the average change in prices paid by the consumers belonging to the population group proposed to be covered by the index for a fixed list of goods and services consumed by them".*

Formerly, this index was also known as the cost of living index. However, since this index measures changes in cost of living due to changes in retail prices only and not due to changes in living standards, etc., the name was changed to consumer price index or retail price index.

10.15 CONSTRUCTION OF CONSUMER PRICE INDEX

The following steps are involved in the construction of a consumer price index:

1. ***Scope and Coverage:*** The scope of consumer price index, proposed to be constructed, must be very clearly defined. This implies the identification of the class of people for whom the index will be constructed such as industrial workers, agricultural workers, urban wage earners, etc. Further, it is also necessary to define the coverage of the class of people, i.e., the definition of geographical location of their stay such as a city or two or more villages, etc. The selected class of people should form a homogeneous group so that weights of various commodities are same for all the people.

2. ***Selection of Base Period:*** A normal period having comparative economic stability should be selected as a base period in order that the consumption pattern used in the construction of the index remains practically stable over a fairly long period.

3. ***Conducting Family Budget Enquiry:*** A family budget gives the details of expenditure incurred by the family on various items in a given period. In order to estimate the consumption pattern, a sample survey of family budgets of the group of people, for whom the index is to be constructed, is conducted and from this an average family budget is prepared. The goods and services that are to be included in the construction of the index are selected from this average family budget. Efforts should be made to include as many commodities as possible. Generally the commodities are divided into five broad groups: (i) Food, (ii) Clothing, (iii) Fuel and Lighting, (iv) House Rent and (v) Miscellaneous.

   If necessary, these groups may further be divided into sub-groups. Percentage expenditure of a group is taken as its weight.

4. ***Obtaining Price Quotations:*** The next step in the construction of consumer price index is to obtain the retail price quotations of various items that are selected. The price quotations should be obtained from those markets from which the group of people, for whom the index number is being constructed, normally make purchases. The quality of various goods and services used by the group of people should also be kept in mind while obtaining price quotations.

5. ***Computation of the Index Number:*** After the collection of necessary data, the consumer price index can be computed by using either of the following formulae:

   (a) *Aggregate Expenditure Method:* Base year quantities are taken as weights in the aggregate expenditure method. The formula for the consumer price index is given by:

   $$p_{01}^{CP} = \frac{\sum p_1 p_0}{\sum p_0 p_0} \times 100$$ which is the Laspeyres's formula.

   (b) *Family Budget Method:* This method is also known as weighted average of price relatives method and accordingly values are taken as weights. The formula for the consumer price index is given by $P_{01}^{CP} = \frac{\sum Pw}{\sum w}$, where

   $$p = \frac{p_1}{p_0} \times 100 .$$

Example: From the information given below, construct the consumer price index number of 1985 by (i) Aggregate Expenditure Method and (ii) Family Budget Method.

Commodities	Quantities (q_0)	Price in 1980 (p_0)	Price in 1985 (p_1)
A	2	75	125
B	25	12	16
C	10	12	16
D	5	10	15
E	25	4.5	5
F	40	10	12
G	1	25	40

Solution:

Calculation of Consumer Price Index

Com.	p_0q_0	p_1q_0	$P = \frac{p_1}{p_0} \times 100$	$w = p_0q_0$	Pw
A	150	250	166.67	150	25000.5
B	300	400	133.33	300	39999.0
C	120	160	133.33	120	15999.6
D	50	75	150.00	50	7500.0
E	112.5	125	111.11	112.5	12499.9
F	400	480	120.00	400	48000.0
G	25	40	160.00	25	4000.0
Total	**1157.5**	**1530**		**1157.5**	**152999.0**

1. Index by agg. exp. Method $= \frac{1530}{1157.5} \times 100 = 132.18$

2. Index by F.B. method $= \frac{152999}{1157.5} = 132.18$

Example: Compute the Cost of Living Index Numbers for 1982 and 1983 with base year 1981, from the following data:

Items	Unit	Price in ₹		
		1981 (p_0)	1982 (p_1)	1983 (p_2)
Food	40 kg.	32	36	40
Clothing	Per Metre	4	3.60	4.40
Fuel	40 kg	8	10	11
Electricity	Per unit	0.40	0.50	0.50
House Rent	1 Block	20	24	30
Miscellaneous	Per unit	1	1.20	1.50

Assign weights to the above items as 3, 2, 1, 1, 2 and 1 respectively.

Solution:

Calculation of the Cost of Living Index Number

Items	$P_1 = \frac{p_1}{p_0} \times 100$	$P_2 = \frac{p_2}{p_0} \times 100$	Weights (w)	P_1w	P_2w
Food	112.5	125.0	3	337.5	375.0
Clothing	90.0	110.0	2	180.0	220.0
Fuel	125.0	137.5	1	125.0	137.5
Electricity	125.0	125.0	1	125.0	125.0
House rent	120.0	150.0	2	240.0	300.0
Miscellaneous	120.0	150.0	1	120.0	150.0
Total			**10**	**1127.5**	**1307.5**

$\therefore$ Cost of living index of 1982 $= \frac{1127.5}{10} = 112.75$

and Cost of living index of 1983 $= \frac{1307.5}{10} = 130.75$

Example: Construct the Bombay working class weighted cost of living index number from the following indices with given weights:

Group	:	Food	Fuel & light	Clothing	Rent	Miscellaneous
Weights	:	47	7	8	13	14
Indices	:	247	293	289	100	236

Solution:

Calculation of Cost of Living Index

Group	Weights (w)	Indices (I)	Iw
Food	47	247	11609
Fuel & light	7	293	2051
Clothing	8	289	2312
Rent	13	100	1300
Miscellaneous	14	236	3304
Total	**89**		**20576**

$\therefore$ Cost of living index number $= \frac{\sum Iw}{\sum w} = \frac{20576}{89} = 231.19$

10.15.1 Uses of Consumer Price Index

1. A consumer price index is used to determine the real wages from money wages and the purchasing power of money.
2. It is also used to determine the dearness allowance to compensate the workers for the rise in prices.
3. It can be used in the formulation of various economic policies of the government.
4. It may be useful in the analysis of markets of certain goods or services.

Example: A particular series of consumer price index covers five groups of items. Between 1975 and 1980 the index rose from 180 to 225. Over the same period, the price index numbers of various groups changed as follows:

Food from 198 to 252; clothing from 185 to 205; fuel and lighting from 175 to 195; miscellaneous from 138 to 212; house rent remained unchanged at 150.

Given that the weights of clothing, house rent and fuel and lighting are equal, determine the weights for individual groups of items.

Solution:

Let w_1% be the weight of food, w_2% be the weight of miscellaneous group and w% be the weight of each of the remaining three groups. Therefore, we can write $w_1 + w_2 + 3w = 100$ or $w_2 = 100 - w_1 - 3w$.

The given data can be written in the form of table as given below:

Groups	Weights	Index in 1975 (I_1)	Index in 1980 (I_2)
Food	w_1	198	252
Clothing	w	185	205
Fuel & Lighting	w	175	195
House Rent	w	150	150
Miscellaneous	$100 - w_1 - 3w$	138	212
Total	**100**		

On the basis of above, the consumer price index of 1975 is

$$\frac{1.98w_1 + (185+175+150)w + 138(100 - w_1 - 3w)}{100} = 180 \text{ (given)}$$

or $60w_1 + 96w = 4200$ ... (1)

Further, the consumer price index of 1980 is

$$= \frac{252w_1 + (205+195+150)w + 212(100 - w_1 - 3w)}{100} = 225 \text{ (given)}$$

or $40w_1 - 86w = 1300$ ... (2)

Solving equations (1) and (2) simultaneously, we get $w = 10$ and $w_1 = 54$

Substituting these values in expression for w_2, we get

$$w_2 = 100 - w_1 - 3w = 100 - 54 - 30 = 16$$

Example: In the consumer price index of a working class of a particular town, the weights corresponding to different groups of items were as follows:

Food = 55, Fuel = 15, Clothing = 10, Rent = 8 and Miscellaneous = 12

In October 1972, the dearness allowance equal to 182% of the workers' wages was given by a mill of the town which only compensated them for the rise in price of food and rent. Another mill of the same town gave a dearness allowance equal to 46.5% of their wages which compensated the workers of that mill for the rise in price of fuel and miscellaneous groups. It is known that price index of food is double the price index of fuel and price index of miscellaneous is double the price index of rent. Find the rise in price of food, fuel, rent and miscellaneous groups if there is no rise in the price of clothing.

Solution:

Let X denote the price index of fuel and Y denote the price index of rent. Therefore, the price index of food = 2X and the price index of miscellaneous = 2Y. The dearness allowance of 182% to the workers of first mill compensated them only for food and rent. Thus, 100 + 182 = 282, can be taken as the consumer price index of the workers, where price of food and rent have gone up while prices of other groups have remained same, i.e., the price index of each of these groups is 100.

From the above, we can write the following equation:

$$\frac{55 \times 2X + 15 \times 100 + 10 \times 100 + 8 \times Y + 12 \times 100}{55+15+10+8+12} = 282$$

or $110X + 8Y = 24500$ ... (1)

Similarly, for the workers of second mill, we can write

$$\frac{55 \times 100 + 15 \times X + 10 \times 100 + 8 \times 100 + 12 \times 2Y}{100} = 146.5$$

or $15X + 24Y = 7350$ ... (2)

On solving these equations simultaneously, we get X = 210 and Y = 175. Since X = 210, therefore, the price of fuel has gone up by 110%. Similarly the index of food is 420, therefore, its price has gone up by 320%. Proceeding in a similar way, the price of rent has gone up by 75% and the price of miscellaneous group has gone up by 250%.

10.16 CALCULATION OF INFLATION

The rate of change of price index is termed as the rate of inflation. If P_t is price index in time period t and P_{t-1} is the price index of the previous time period, the rate of inflation for time period t, denoted by I_t is given by

$$I_t = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100$$

Many countries of the world use consumer price index to measure inflation. However, in India, it is measured by the use of whole sale price index. The rate of inflation is often computed from year to year basis where as in India, it is also computed on weekly basis.

Example: Compute the rate of inflation for the following data:

Years	2000	2001	2002	2003	2004	2005
Price Index	108.2	114.6	120.5	123.3	128.2	130.1

Solution:

Computation of Inflation

Years	Price Index	Inflation Rate (%)
2000	108.2	—
2001	114.6	5.91
2002	120.5	5.15
2003	123.3	2.32
2004	128.2	3.97
2005	130.1	1.48

The rate of inflation for 2003, for example is $\frac{(123.3 - 120.5)}{120.5} \times 100 = 2.32\%$

10.17 STOCK MARKET INDEX

A stock market index measures the change in prices of a set of stocks, which are included in the index. This is given by the weighted average of the prices of the equities, included in the index, relative to the weighted average of prices in base year. This index is calculated after every 15 seconds on each trading day.

The stock market index indicates the overall market sentiments through a set of stocks that are representative of the markets. This index serves:

1. As a barometer of market behaviour,
2. As an indicator of day to day fluctuations in stock prices.

10.17.1 Methods of Index Number Construction

The method of construction of an index number of a stock market depends upon the types of weights assigned to different stocks included in the index. The weights in an index can be of following three types, namely, market capitalisation weights, price weights and equal weights.

Based on these weights, we have the following methods of index number construction:

1. ***Market Capitalisation Weighted Index:*** In this index, the weights can be assigned on the basis of any one of the following methods:

   (a) *Full Market Capitalisation Method:* In this method, the equity price is assigned a weight on the basis of full market capitalisation. Market capitalisation is determined as the product of equity price and the number of shares issued. Thus a company with higher market capitalisation would be assigned a greater weight.

   (b) *Free-float Market Capitalisation Method:* Free-float implies the percentage of shares that is freely available for sale and purchase in the market. It excludes strategic holdings in the company which would not, in the normal course, come in the open market for trading. Stock held by government, controlling shareholders, the company's management, restricted shares due to IPO regulations, and shares locked under the employee stock ownership plan are generally excluded from the definition of free-float. Free-float market-capitalisation reflects the investible market capitalisation which may be much less than the total market capitalisation. A company with higher free-float market capitalisation would be assigned a greater weight and thus be more prominent in the index. Thus the companies which provide high value to share holders but have less free-float would be marginalised.

   The above methodology has become very popular the world over. All major index providers like MSCI, FTSE, STOXX, S&P and Dow Jones use this method.

   (c) *Modified Capitalisation Method:* This method seeks to limit the influence of the largest stock in the index which otherwise would dominate in the index. This method sets a limit on the percentage weight of the largest stock or a group of stocks.

2. ***Price Weighted Index:*** In a price weighted index, each stock is given a weight proportional to its stock price. This index is given by the formula.

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

   where ΣP_1 is the aggregate of stock prices in current year and ΣP_0 is the aggregate of stock prices in base year. Note that simple aggregation of prices also involves implicit weighing because a firm with higher price is automatically assigned a greater weight and vice-versa.

3. ***Equally Weighted Index:*** In this index, all the stocks have same weightage irrespective of their price or their market-capitalisation. The formula for the index is given below.

$$P_{01} = \Sigma\left(\frac{p_1}{p_0} \times 100\right) \div n$$

   Note that this index is obtained by taking simple average of price relatives.

10.17.2 SENSEX Calculation Methodology

SENSEX is calculated using the "Free-float Market Capitalisation" methodology with effect from September 1, 2003. Before this date, the index used to be calculated by "Full Market Capitalisation" methodology. First complied in 1986, SENSEX is a basket of 30 constituent stocks representing a sample of large, liquid and

representative companies. The base year of SENSEX is 1978-79 and the base value is 100 index points.

The calculation of SENSEX involves dividing the free-float market-capitalisation of 30 companies, in the index, by a number called the Index Divisor. This divisor is the only link to the original base period value of the SENSEX. It keeps the index comparable over time and is the adjustment point for all index adjustments arising out of corporate actions, replacement of scrips, etc. During market hours prices of the index scrips, at which latest trades are executed, are used to calculate SENSEX every 15 seconds.

10.17.3 Dollex-30

BSE also calculates a dollar-linked version of SENSEX which is known as Dollex-30.

10.17.4 Index Closure Algorithm

The closing SENSEX on any trading day is computed by taking the weighted average of all the trades on SENSEX constituents in the last 30 minutes of the trading session. If a SENSEX constituent has not traded in the last 30 minutes, the last trade price is taken for computation of index closure. If a SENSEX constituent has not traded at all in the day, then its last day's price is taken for computation of index closure. The use of Index Closure Algorithm prevents any intentional manipulation of the closing index value.

10.17.5 Maintenance of SENSEX

One of the important aspects of maintaining continuity with the past is to update the base year average. The base year value adjustment ensures that replacement of stocks in the Index, additional issue of capital and other corporate announcements like 'right issue', etc., do not destroy the historical value of the index. The beauty of maintenance lies in the fact that adjustments for corporate actions in the index should not per-se affect the index values.

10.17.6 SENSEX–Scrip Selection Criteria

The general guidelines for selection of constituents in the SENSEX are as follows:

1. ***Listed History:*** The scrip should have a listing history of at least 3 months at BSE. Exception may be considered if full market capitalisation of a newly listed company ranks among top 10 in the list of BSE universe. In case a company is listed on account of merger/demerger/amalgamation, minimum listing history would not be required.
2. ***Trading Frequency:*** The scrip should have been traded on each and every trading day in the last three months. Exceptions can be made for extreme reasons like scrip suspension, etc.
3. ***Final Rank:*** The scrip should figure in the top 100 companies listed by final rank. The final rank is arrived at by assigning 75% weightage to the rank on the basis of three-month average full market capitalisation and 25% weightage to the liquidity rank based on three-month average daily turn over and three-month average impact cost.
4. ***Market Capitalisation Weightage:*** The weightage of each scrip in SENSEX based on three-month average free-float market capitalisation should be at least 0.5% of the index.

5. ***Industry Representative:*** Scrip selection would generally take into account a balanced representation of the listed companies in the universe of BSE.
6. ***Track Record:*** In the opinion of the committee, the company should have an acceptable track record.

10.17.7 Index Review Frequency

The index committee meets every quarter to discuss the index related issues. In case of a revision in the index constituents, the announcement of the incoming and outgoing scrips is made six weeks in advance of the actual implementation of the revision of the index.

10.17.8 S&P CNX Nifty

This is an index of National Stock Exchange (NSE). The NSE – 50, Nifty, was launched in April 1996. It was rechristened as S&P CNX Nifty in July 1998. This index is widely used as it reflects the state of market sentiments through 50 blue chip, large cap, highly liquid and highly traded stocks covering 23 sectors.

10.17.9 Method of Computation

S&P CNX Nifty is computed by using market capitalisation weighted method, wherein the level of the index reflects the total market value of all the stocks in the index relative to base period. This method also takes into account constituent changes in the index and importantly corporate actions such as stock splits, rights, etc. without affecting the index value.

10.17.10 Base Date and Value

The base period selected for S&P CNX Nifty index is the close of prices on November 3, 1995, which marks the completion of one year of operations of NSE's Capital Market Segment. The base value of the index has been set at 1000 and a base capital of ₹ 2.06 trillion.

10.17.11 Criteria for Selection of Constituent Stocks

1. ***Liquidity (Impact Cost):*** For inclusion in the index, the security should have traded at an average impact cost of 0.75% or less during the last six months for 90% of the observations for a basket size of ₹ 5 million.

   Impact cost is the cost of executing a transaction in a security in proportion to the weightage of its market capitalisation as against the index market capitalisation at any point of time. This is the percentage markup suffered while buying/selling the desired quantity of a security compared to its ideal price (best buy + best sell)/2.
2. ***Market Capitalisation:*** Companies eligible for inclusion in the Nifty must have a six monthly average market capitalisation of ₹ 500 crores or more during the last six months.
3. ***Floating Stock:*** Companies eligible for inclusion in Nifty should have at least 10% floating stock. For this purpose, floating stock shall mean stocks which are not held by promoters and associated entities of such companies.
4. ***Others:*** A company which comes out with a IPO will be eligible for inclusion in the index, if it fulfills the normal eligibility criteria for the index like impact cost, market capitalisation and floating stock, for a three month period instead of a six month period.

10.17.12 Replacement of Stock from the Index

A stock may be replaced from the index for the following reasons:

1. Compulsory changes like corporate actions, delisting, etc. in such a scenario, the stock having the largest market capitalisation and satisfying other requirements related to liquidity, turn over and free float will be considered for inclusion.
2. When a better candidate is available in the replacement pool, which can replace the index stock i.e. the stock with the highest market capitalisation in the replacement pool has at least twice the market capitalisation of the index stock with the lowest market capitalisation.

10.17.13 CNX Nifty Junior

The next rung of liquid securities after S&P CNX Nifty is the CNX Nifty Junior. This index is built out of next 50 highly liquid stocks.

As with S&P CNX Nifty, the stocks in the CNX Nifty Junior are filtered for liquidity, the most liquid of the stocks excluded from S&P CNX Nifty. The maintenance of the two indices are synchronised so that the two sets of selected stocks are disjoint. Hence, it is always meaningful to pool the two indices into a composite 100 stocks index. CNX Nifty Junior was introduced on January 1, 1997 with base date November 3, 1996 and base value 1000 with base capital of ₹ 0.43 trillion.

10.17.14 S&P CNX Defty

This index was introduced on November 26, 1997. It shows returns on S&P CNX Nifty index in dollar terms. This index serves as a performance indicator to Foreign Institutional Investors (FIIs), off-shore funds and others.

10.17.15 Total Returns Index

Nifty is a price index and hence reflects the returns one would earn if investment is made in the index portfolio. However, a share price index does not consider the returns arising from dividend receipts. Only capital gains arising due to price movements of constituent stocks are indicated in a price index. Therefore, to get a true picture of returns, the dividends received from the constituent stocks also need to be reflected in the index values. The Total Return Index is an index which reflects the return from gain/loss from changes in prices plus dividend payments by constituent stocks.

10.18 PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

The following are some general problems that are faced in the construction of any index number:

1. ***Definition of the purpose:*** Since it is possible to construct index numbers for a number of purposes and one cannot have an all-purpose index, therefore, it is very essential to define the specific purpose of its construction. For example, if we are interested in the construction of a price index number, we must have knowledge about the purpose to be served by it, i.e., what is to be measured by it; like the cost of living of workers or the change in wholesale prices, etc. In the absence of this information, it may be difficult to carry out various steps in the construction of an index number. The questions like what are items to be included, from which of the markets the price quotations are to be obtained, what will be the weights of different items, etc., cannot be answered unless the purpose of the index number

construction is known. Further, an index number can be of sensitive or general nature. In case of sensitive index, only those items are included whose variables (like prices in case of price index) fluctuate very often; while efforts are made to include as many items as possible when the index is of general nature. It may be pointed out that the index numbers are specialised tools and as such are more useful and efficient when properly used. The first step in this direction is a specific definition of the purpose of its construction.

2. ***Selection of the base period:*** Every index number is constructed with reference to a base period. There are two important points that must be kept in mind while selecting the base period of an index number.

   (a) The base period should correspond to a period of relative economic and political stability, i.e., it should be a normal or representative period in some way. In certain situations, where identification of such a period is not possible, the average of certain periods can also be taken as base.

   (b) The base period should not be too distant from the current period. The comparison of current period with a remote base doesn't have much relevance. In the words of Morris Hamburg, "It is desirable that the base period be not too far away in time from the present. The further away we move from the base period the dimmer are our recollections of economic conditions prevailing at that time. Consequently, comparisons with these remote base periods tend to lose significance and become rather tenuous in meaning".

   Another problem with a remote base period can be that certain items that were in use in the base period are no longer in use while certain new items are in use in current period. In such a situation, the two item bundles are no longer homogeneous and comparable. This problem is less likely to occur when fairly recent period is chosen as base.

3. ***Selection of number and type of items:*** An index number of a particular group of items is in fact based on a sample of items taken from it. It is neither possible nor necessary to include all the items of the group in the construction of an index number. The number of items to be included depends largely upon the purpose of the index number.

   There are no hard and fast rules that can be laid down with regard to the selection of the number of items, however, it must be remembered that more is the number of items the more representative will be the index number and more cumbersome will be the task of computations. Therefore, it is necessary to have some sort of balance between having a representative index and the work of computation involved in its construction.

   The following points should be kept in mind in selecting the type of items:

   (a) The items should be representative of the tastes, habits and customs of the people for whom the index is to be constructed.

   (b) The selected items should be of stable quality. The standardised items should be given preference.

   (c) As far as possible, the non-tangible items like personal services, goodwill, etc., should be excluded because it is difficult to ascertain their value.

4. ***Collection of data:*** The next important step in the construction of an index number is the collection of data. For example, for the construction of price index, price quotations are to be obtained. Since the prices of commodities may vary from one market to another and in certain cases from one shop to another, it is necessary to select those markets which are representative in the sense that the

group under consideration generally makes purchases from these markets. The next logical step is to select an agency through which price quotations are to be obtained. The selected agency should be highly reliable and if necessary the accuracy of price quotations reported by it may also be checked by appointing some other agency or agencies. Furthermore, care should always be taken to obtain price quotations for the same quality of items.

Similar type of considerations is necessary for the collection of data for the construction of index numbers such as quantity index, value index, unemployment index, etc.

5. ***Selection of a suitable average:*** Since the index numbers are also averages, any of the five averages, viz. arithmetic mean, median, mode, geometric mean and harmonic mean can be used in its construction. However, since in most of the situations we have to average ratios of the values in current period to that in base period, geometric mean is the most suitable average in the construction of index numbers. The main difficulty of using the geometric mean is the complexities of its computations and hence, the use of arithmetic mean is more popular inspite of its being less suitable.

6. ***Selection of suitable weights:*** According to John I. Griffin, *"Weighing is designed to give component series an importance in proper relation to their real significance."* The basic purpose of weighing is to enable each item to have an influence, on the index number, in proportion to its importance in the group. It is, therefore, necessary to design a system of weighing such that true importance of the items is reflected by it. The system of weighing may be either arbitrary or rational. Arbitrary or chance weighing implies that the statistician is free to assign weights to different items as he thinks fit or reasonable. Rational or logical weighing, on the other hand, implies that some criterion has been fixed for assigning weights. Two types of weights are commonly used in the construction of a price index number: (i) physical quantities and (ii) money values. These weights can be quantities (or values) produced or consumed or sold in base or current or in any other period.

Another problem, to be tackled, with regard to system of weights is whether weights should be fixed or fluctuating. When relative importance of various items change in different periods, it is desirable to have fluctuating system of weights to get better results.

10.19 COMPARISON OF LASPEYRES'S AND PAASCHE'S INDEX NUMBERS

Out of various formulae discussed so far, the Laspeyres' and Paasche's formulae are generally preferred for the construction of index numbers. The main reason for this is that the values of these index numbers have a simple interpretation. For example, in case of Laspeyres' index, the base year quantities are used as weights and $\Sigma p_1 q_0$ gives the cost of base year bundle of goods valued at current year prices. Similarly, $\Sigma p_0 q_0$ gives the cost of base year bundle valued at base year prices. Therefore, the ratio

$$\frac{\sum p_1 q_0}{\sum p_0 q_0}$$

gives the change in cost of purchasing the bundle q_0.

In a similar manner, the Paasche's price index can be interpreted as the change in cost of purchasing the bundle q_1. Out of these two, the Laspeyres' index is preferred because weights do not change over different periods and hence the index numbers of

various periods remain comparable. Furthermore, Laspeyres' index requires less calculation work than the one with changing weights in every period. The main disadvantage of Laspeyres' formula is that with passage of time the relative importance of various items may change and the base year weights may become outdated. Paasche's index, on the other hand, uses current year weights which truly reflect the relative importance of the items. The main difficulty, in this case, is that index numbers of various periods are not comparable because of changing weights. Moreover, it may be too expensive and difficult to obtain these weights.

When both the index number formulae are applied to the same data, they will in general give different values. However, "if prices of all the commodities change in the same ratio, the Laspeyres' index is equal to Paasche's index, for then the two weighing systems become irrelevant; or, if quantities of all the commodities change in same ratio, the two index numbers will again be equal, for then the two weighing systems are same relatively." In order to show this, let p_{1i} be the price of i^{th} commodity in current year and p_{0i} be its price in base year. If prices of all the commodities increase by 5%, then we can write $p_{1i}/p_{0i} = 105/100$ or $p_{1i} = 1.05 \times p_{0i}$, for all values of i. To generalise, we assume that $p_{1i} = a.p_{0i}$ (or $p_1 = a.p_0$, on dropping the subscript i), where a is constant.

We can write the Laspeyres' index as:

$$P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Substituting $p_1 = ap_0$, we have

$$P_{01}^{La} = \frac{\sum \alpha p_0 q_0}{\sum p_0 q_0} \times 100 = \alpha \frac{\sum p_0 q_0}{\sum p_0 q_0} \times 100 = 100\alpha \qquad \ldots(1)$$

Similarly, the Paasche's index is given by

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{\sum \alpha p_0 q_1}{\sum p_0 q_1} \times 100 = \alpha \frac{\sum p_0 q_1}{\sum p_0 q_1} = 100\alpha \qquad \ldots(2)$$

Hence, $P_{01}^{La} = P_{01}^{Pa}$

Further, when quantities of all the commodities change in same proportion, we can write, $q_1 = b.q_0$, for all commodities. Here b is a constant.

Thus, we can write the Paasche's index as

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{\beta \sum p_1 q_0}{\beta \sum p_0 q_0} \times 100 = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Hence, $P_{01}^{Pa} = P_{01}^{La}$

10.20 RELATION BETWEEN WEIGHTED AGGREGATIVE AND WEIGHTED ARITHMETIC AVERAGE OF PRICE RELATIVES INDEX NUMBERS

It will be shown here that, basically, the two types of index numbers, weighted aggregative and weighted arithmetic average of price relatives, are same and that one type of index number can be obtained from the other by suitable selection of weights. Since the weighted aggregative index numbers are easy to calculate and have simple

interpretation, they are preferred to weighted arithmetic average of price relatives indices.

Consider the Laspeyre's index

$$P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Rewriting $\sum p_1 q_0$ as $\sum \frac{p_1}{p_0} . p_0 q_0$

We have $P_{01}^{La} = \frac{\sum \frac{p_1}{p_0} . p_0 q_0}{\sum p_0 q_0} \times 100 = \frac{\sum Pw}{\sum w}$ ...(1)

Here $P = \frac{p_1}{p_0} \times 100$ and $w = p_0 q_0$

In a similar way, the other aggregative type of index numbers can also be converted into average type index numbers.

Further, it can be shown that an arithmetic average type of index number can be converted into an aggregative type by a suitable selection of weights.

Consider $P_{01} = \frac{\sum \frac{p_1}{p_0} \times w}{\sum w} \times 100$

Let $w = p_0 q_1$, then the above equation can be written as

$$P_{01} = \frac{\sum \frac{p_1}{p_0} \times p_0 q_1}{\sum p_0 q_1} \times 100 = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = P_{01}^{Pa}$$

From the practical point of view, the weighted aggregative methods are preferred to weighted average of price relatives because the former are easy to compute and have simple interpretation.

However, the weighted average of price relatives may be useful when we are interested in knowing the extent of homogeneity in price movements of a certain sub-group or the whole group of commodities. Further, to determine the relative importance of an item, it is necessary to write the index number formula in weighted average of price relative form.

From equation (1), we can write

$$P_{01}^{La} = \sum \left(\frac{p_1}{p_0} . \frac{p_0 q_0}{\sum p_0 q_0} \right) \times 100$$

Let $w = \frac{p_0 q_0}{\sum p_0 q_0}$ the Sw = 1. Thus, $w = \frac{p_0 q_0}{\sum p_0 q_0}$ for an item gives its relative importance (or weight) in the group.

10.20.1 Change in the Cost of Living due to Change in Price of an Item

Let $q_1, q_2, \ldots\ldots q_n$ be the fixed quantities of the n commodities consumed by a group of consumers irrespective of price changes; and $p_{01}, p_{02}, \ldots\ldots p_{0n}$ and $p_{11}, p_{12}, \ldots\ldots p_{1n}$ be their prices in base and current years respectively. Their cost of living index, measured by the change in expenditure of purchasing a given bundle of commodities, can be written as

$$P_{01} = \frac{\sum p_{1i}q_i}{\sum p_{0i}q_i} \times 100$$

Let the price of i^{th} commodity changes by 100a%. Thus, the new price, denoted as p'_{01}, can be written as $(1 + a)p_{1i}$ and the changed index number would be

$$p'_{01} = \frac{\left(\sum p_{1i}q_i + \alpha p_{1i}q_i\right)}{\sum p_{0i}q_i} \times 100 = p_{01} + \frac{\alpha p_{1i}q_i}{\sum p_{0i}q_i} \times 100$$

Hence, the absolute change in the cost of living is given by

$$P'_{01} - P_{01} = \frac{\alpha p_{1i}q_i}{\sum p_{0i}q_i} \times 100$$

and the proportionate change is given by

$$\frac{P'_{01} - P_{01}}{P_{01}} = \frac{\alpha p_{1i}q_i}{\sum p_{0i}q_i} \times 100 \times \frac{\sum p_{0i}q_i}{\sum p_{1i}q_i} \times \frac{1}{100} = \frac{\alpha p_{1i}q_i}{\sum p_{1i}q_i}$$

We note that

$$\frac{p_{1i}q_i}{\sum p_{1i}q_i}$$

is the proportion of expenditure on the i^{th} commodity before the change of price.

Alternatively, the above equation can be written as:

$$\begin{pmatrix} \textit{Proportionate change in} \\ \textit{price of the commodity} \end{pmatrix} \times \begin{pmatrix} \textit{Proportion of expenditure} \\ \textit{on the commodity} \end{pmatrix} = \begin{pmatrix} \textit{Proportionate change in} \\ \textit{the cost of living} \end{pmatrix}$$

It may be pointed out here that the above result assumes that the consumption of the commodity remains unchanged as a result of change in its price.

10.21 LIMITATIONS OF INDEX NUMBERS

Despite the fact that index numbers are very useful for the measurement of relative changes, these suffer from the following limitations:

1. The computation of an index number is based on the data obtained from a sample, which may not be a true representative of the universe.
2. The composition of the bundle of commodities may be for different years. This cannot be taken into account by the fixed base method. Although this difficulty can be overcome by the use of chain base index numbers, but their calculations are quite cumbersome.
3. An index number doesn't take into account the quality of the items. Since a superior item generally has a higher price and the increase in index may be due to an improvement in the quality of the items and not due to rise of prices.

4. Index numbers are specialised averages and as such these also suffer from all the limitations of an average.
5. An index number can be computed by using a number of formulae and different formulae will give different results. Unless a proper method is used, the results are likely to be inaccurate and misleading.
6. By the choice of a wrong base period or weighing system, the results of the index number can be manipulated and, thus, are likely to be misused.

Check Your Progress

Fill in the blanks:

1. Every index number is constructed with reference to a _____ period.
2. Index numbers are also called _____ of economic activity.
3. Ideal index formula _____ both, the time reversal and factor reversal tests.
4. The rate of change of _____ is termed as the rate of inflation.
5. A _____ is used to determine the real wages from money wages and the purchasing power of money.
6. The basic purpose of _____ is to enable each item to have an influence, on the index number, in proportion to its importance in the group.

10.22 LET US SUM UP

- An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.
- In real situations, neither the prices of all the items change in the same ratio nor in the same direction, i.e., the prices of some commodities may change to a greater extent as compared to prices of other commodities.
- The index numbers are very useful device for measuring the average change in prices or any other characteristics like quantity, value, etc., for the group as a whole.
- Index numbers are specialized type of averages that are used to measure the changes in characteristics which is not capable of being directly measured.
- The changes in magnitude of a group are expressed in terms of percentages which are independent of the units of measurement. This facilitates the comparison of two or more index numbers in different situations.
- Index numbers are indispensable tools for the management of any government or non-government organizations.
- There is inverse relation between the purchasing power of money and the general price level measured in terms of a price index number.
- The reciprocal of the relevant price index can be taken as a measure of the purchasing power of money.
- The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

- While taking weighted average of price relatives, the values are often taken as weights. These weights can be the values of base year quantities valued at base year prices.
- In case of weighted aggregative price index numbers, quantities are often taken as weights. These quantities can be the quantities purchased in base year or in current year or an average of base year and current year quantities or any other quantities.
- A quantity index number measures the change in quantities in current year as compared with a base year.
- Index numbers where comparisons of various periods were done with reference to a particular period, termed as base period. Such type of index number series is known as fixed base series.
- Simple Average of Price Relatives Index

$$P_{01} = \frac{\sum \frac{p_1}{p_0}}{n} \text{ (using A.M.)}$$

$$P_{01} = \text{Antilog}\left[\frac{\sum \log \frac{p_1}{p_0} \times 100}{n}\right] \text{ (using G.M.)}$$

- Simple Aggregative Index $P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$
- Weighted Average of Price relatives Index

$$P_{01} = \frac{\sum Pw}{\sum w} \text{ (Using weighted A.M.)}$$

$$P_{01} = \text{Antilog}\left[\frac{\sum w \log P}{\sum w}\right] \text{ (Using weighted G.M.)}$$

Here $P = \frac{p_1}{p_0} \times 100$ and w denotes values (weights)

- Weighted Aggregative Index Numbers

(a) Laspeyres's Index $P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

(b) Paasche's Index $P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$

(c) Fisher's Ideal Index $P_{01}^{DB} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$

(d) Dorbish and Bowley's Index $P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100$

(e) Marshall and Edgeworth Index $P_{01}^{ME} = \dfrac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_0} \times 100$

(f) Walsh's Index $P_{01}^{Wa} = \dfrac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$

(g) Kelly's Index $P_{01}^{Ke} = \dfrac{\sum p_1 q}{\sum p_0 q} \times 100$

- Real Wage $= \dfrac{\text{Money Wage}}{\text{CPI}} \times 100$
- Output at Constant Prices $= \dfrac{\text{Output at Current Prices}}{\text{Price Index}} \times 100$
- Purchasing Power of Money $= \dfrac{1}{\text{Price Index}} \times 100$

10.23 UNIT END ACTIVITY

Taking 1983 as base year, calculate an index number of prices for 1990, for the following data given in appropriate units, using:

1. Weighted arithmetic mean of price relatives by taking weights as the values of current year quantities at base year prices, and
2. Weighted geometric mean of price relatives by taking weights as the values of base year quantities at base year prices.

Commodity	1983		1990	
	Price	Quantity	Price	Quantity
A	82	63	160	56
B	80	75	182	53
C	105	92	185	64
D	102	25	177	13
E	102	63	175	54
F	190	61	140	60

10.24 KEYWORDS

Barometers of Economic Activity: Sometimes index numbers are termed as barometers of economic activity.

Base Year: The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Current Year: The year under consideration for which the comparisons are to be computed is called the current year. It is commonly denoted by writing '1' as a subscript of the variable.

Cyclical Variations: The oscillatory movements are termed as Cyclical Variations if their period of oscillation is greater than one year.

Dorbish and Bowley's Index: This index number is constructed by taking the arithmetic mean of the Laspeyres' and Paasche's indices.

Fisher's Index: Fisher suggested that an ideal index should be the geometric mean of Laspeyres' and Paasche's indices.

Index Number: An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.

Kelly's Fixed Weights Aggregative Index: The weights, in this index number. are quantities which may not necessarily relate to base or current year. The weights, once decided, remain fixed.

Laspeyres' Index: Laspeyres' price index number uses base year quantities as weights.

Link Relatives Method: This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern.

Marshall and Edgeworth's Index: This index number uses arithmetic mean of base and current year quantities.

Paasche's Index: This index number uses current year quantities as weights.

Periodic Variations: These variations, also known as oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation.

Quantity Index Number: A quantity index number measures the change in quantities in current year as compared with a base year.

10.25 QUESTIONS FOR DISCUSSION

1. What are index numbers? Discuss their uses.
2. Examine the various steps in the construction of an index number.
3. Distinguish between average type and aggregative type of index numbers. Discuss the nature of weights used in each case.
4. Distinguish between simple and weighted index numbers. Explain 'weighted aggregative' and 'weighted average of relatives' methods for the construction of index numbers.
5. Construct Laspeyres's, Paasche's and Fisher's indices from the following data:

	1986		1987	
Item	**Price (₹)**	**Expenditure (₹)**	**Price (₹)**	**Expenditure (₹)**
1	10	60	15	75
2	12	120	15	150
3	18	90	27	81
4	8	40	12	48

6. From the following data, prove that Fisher's Ideal Index satisfies both the time reversal and the factor reversal tests.

	Base Year		Current Year	
Commodity	**Price**	**Quantity**	**Price**	**Quantity**
A	6	50	10	60
B	2	100	2	120
C	4	60	6	60

7. Examine the problems involved in the construction of an index number.

8. Given the following data:

Year	**Average Weekly take-home wages** (₹)	**Consumer Price Index** (₹)
1968	109.50	112.8
1969	112.20	118.2
1970	116.40	127.4
1971	125.08	138.2
1972	135.40	143.5
1973	138.10	149.8

(a) What was the real weekly wage for each year?

(b) In which year did the employees had the greatest buying power?

(c) What percentage increase in the average weekly wages for the year 1973 is required to provide the same buying power that the employees enjoyed in the year in which they had the highest real wages?

9. Construct Consumer Price Index for the year 1981 with 1971 as the base year.

Items	:	Food	Rent	Clothes	Fuel	Others
Percentage Expenses	:	35%	15%	20%	10%	20%
Value Index (1971)	:	150	50	100	20	60
Value Index (1981)	:	174	60	125	25	90

10. Compute consumer price index number from the following data by aggregate expenditure.

Commodity	**Quantities consumed in base year**	**Units which prices are quoted**	**Prices base year**	**Prices in current year**
Wheat	400 kgs	/quintal	350	400
Rice	2 quintals	/quintal	580	700
Gram	100 kgs	/quintal	740	950
Pulses	2 quintals	/quintal	980	1200
Ghee	50 kgs	/kg	70	85
Sugar	50 kgs	/kg	8	11
Fire wood	5 quintals	/quintal	50	60
House Rent	1 house	/house	1600	1800

11. A textile worker in the city of Ahmedabad earns ₹ 750 per month. The cost of living index for January 1986 is given as 160. Using the following data, find out the amounts he spends on (i) Food and (ii) Rent.

12. "In the construction of index numbers the advantages of geometric mean are greater than those of arithmetic mean". Discuss.

13. Show that the Laspeyres's index has an upward bias and the Paasche's index has a downward bias. Under what conditions the two index numbers will be equal?

14. Compute price index number of 1992 with 1985 as base for the following data by (i) simple arithmetic and (ii) simple geometric average of price relatives.

Commodity	Price in 1985	Price in 1992
A	45	54
B	57	47
C	50	58
D	48	52
E	51	58
F	55	60

15. Calculate price index for 1985 and 1986 using 1980 as base from the following data by using the method of simple arithmetic average of price relatives.

Commodity	Prices (₹ per unit)		
	1980	1985	1986
1	8	9	7
2	10	13	10
3	11	15	9
4	50	47	36
5	150	200	180

16. Calculate the index number for 1992 with 1990 as base from the following data using (i) weighted arithmetic mean and (ii) weighted geometric mean of price relatives.

Commodity	Weights	Prices	
		1990	1992
A	22	250	620
B	48	330	440
C	17	625	1275
D	13	65	90

17. The price quotations of four different commodities for 1986 and 1987 are given below. Calculate the index number of 1987 with 1986 as base by using:

    (a) The simple average of price relatives and

    (b) The weighted average of price relatives.

Comm.	Unit	Weight	Price in ₹	
		(₹ '000)	1986	1987
1	Kg.	5	4.00	9.00
2	Quintal	7	5.00	6.40
3	Dozen	6	6.00	9.00
4	Kg.	2	2.00	3.60

18. An enquiry into the budgets of middle class families of a city gave the following information:

	Food	Clothing	Fuel	Rent	Miscellaneous
Percentage of Expenses on	30	25	15	20	10
Price in 1990 (₹)	180	150	125	90	65
Price in 1992 (₹)	200	160	150	105	65

Compute the price index number using:

(a) Weighted A.M. of price relatives.

(b) Weighted G.M. of price relatives.

Check Your Progress: Model Answer

1. Base
2. Barometer
3. Fisher
4. Price index
5. Consumer price index
6. Weighing

10.26 REFERENCE & SUGGESTED READINGS

- Levin, R. I., & Rubin, D. S. (2018). **Statistics for Management** (8th ed.). Pearson India. ISBN: 9788131772840
- Camm, J. D., Cochran, J. J., Fry, M. J., Ohlmann, J. W., & Anderson, D. R. (2019). **Essentials of Business Analytics** (2nd ed.). Cengage Learning. ISBN: 9781337406420
- Sharpe, N. R., De Veaux, R. D., & Velleman, P. F. (2019). **Business Statistics** (3rd ed.). Pearson. ISBN: 9780134684773
- Groebner, D. F., Shannon, P. W., & Fry, P. C. (2020). **Business Statistics: A Decision-Making Approach** (10th ed.). Pearson. ISBN: 9780134496499

UNIT - XI

DIAGRAMMATIC AND GRAPHIC PRESENTATION OF DATA

CONTENTS

11.0 AIMS AND OBJECTIVES

After studying this lesson, you should be able to:

- Define diagram and graph
- Describe the differences between diagram and graph
- Explain the types of diagram

11.1 INTRODUCTION

Besides textual and tabular presentations of statistical data, the most attractive and commonly used popular modem device to exhibit any data in a systematic manner is to represent them with suitable and appropriate diagrams and pictures. The usual and effective means in this context are: graphs, charts, pictures, etc. and they are really and

surely capable of depicting some important features of the data which they individually are not able to exhibit. Selection of the appropriate diagram actually depends on the nature of the raw data available and the purpose or the area in which it will be applied. However, only certain limited information can be supplied through a particular diagram and as such each diagram has certain specific limitations of its own.

Tabulation and grouping does make data simple to understand and analyze. However, just the numerical data is not attractive enough to present it to higher management, stakeholders and those not very familiar with the particular functional area. Moreover, pictorial or graphical representation is catchy to appreciate, remember, and grasp quickly and easy to explain. It allows us to obtain the underlying information in one glance. "One picture is equal to a thousand words" as the proverb goes. Hence, diagrams, graphs and charts have assumed importance for decision-making to the managers. To communicate the information effectively to the higher management, you must present the data in pictorial format whenever feasible, and support it with the numerical data as a reference. Remember, higher management may not have adequate time to analyze the numerical data. Similarly, always present the information to junior employees as diagrams, graphs and charts, because they may not have adequate knowledge and grasp of numerical analysis.

11.2 DIAGRAMS AND GRAPHS

One of the most effective and interesting alternative way in which a statistical data may be presented is through diagrams and graphs. There are several ways in which statistical data may be displayed pictorially such as different types of graphs and diagrams.

Through both diagrams and graphs are handy tools in the hands of a statistician for representation of statistical data, there are much differences between the two. A brief distinction between a diagram and a graph is given in Table 11.1.

Table 11.1: Differences between Diagram and Graph

Diagram	Graph
1. Can be drawn on an ordinary paper.	1. Can be drawn on a graph paper.
2. Easy to grasp.	2. Needs some effort to grasp.
3. Not capable of analytical treatment.	3. Capable of analytical treatment.
4. Can be used only for comparisons.	4. Can be used to represent a mathematical relation.
5. Data are represented by bars, rectangles, pictures, etc.	5. Data are represented by lines and Curves.

A graphic presentation is used to represent two types of statistical data: (i) Time Series Data and (ii) Frequency Distribution.

11.3 TYPES OF DIAGRAMS

There are a large number of diagrams which can be used for presentation of data. The selection of a particular diagram depends upon the nature of data, objective of presentation and the ability and experience of the person doing this task. For convenience, various diagrams can be grouped under the following categories:

11.3.1 One-dimensional Diagrams

One-dimensional diagrams are also known as bar diagrams. In case of one-dimensional diagrams, the magnitude of the characteristics is shown by the length or

height of the bar. The width of a bar is chosen arbitrarily so that the constructed diagram looks more elegant and attractive. It also depends upon the number of bars to be accommodated in the diagrams. If large numbers of items are to be included in the diagram, lines may also be used instead of bars. Different types of bar diagrams are line diagram, bar diagram and column diagram.

11.3.2 Two-dimensional Diagrams

In case of a two-dimensional diagram, the value of an item is represented by an area. Such diagrams are also known as 'surface' or 'area diagrams'. Popular forms of two-dimensional diagrams are:

- Rectangular Diagrams
- Square Diagrams
- Circular or Pie Diagrams

11.3.3 Three-dimensional Diagrams

With the help of three dimensional diagrams, the values of various items are represented by the volume of cube, sphere, cylinder, etc. These diagrams are normally used when the variations in the magnitudes of observations are very large.

11.3.4 Pictograms and Cartograms

These are like frequency plots. The data points are plotted on the graph in the same manner. Then instead of joining the data points, pictures or objects of the height of the data points are used to depict the data. In that case, heights of the pictures or objects represent the frequency. These include histograms and frequency polygon.

11.4 BAR DIAGRAM

Bar diagrams and column diagrams are very common in representing business data. These are used to depict the frequencies of different categories of variables. In case of bar diagrams, the bars are horizontal with their lengths proportional to the frequencies. On the other hand, in column diagrams the frequencies are depicted by vertical columns having their length proportional to the frequencies. We can also have multiple bars or columns representing different categories of variables. Further, data related to sub-categories in a category can be shown on same bar or column by overlapping the bars or column on top.

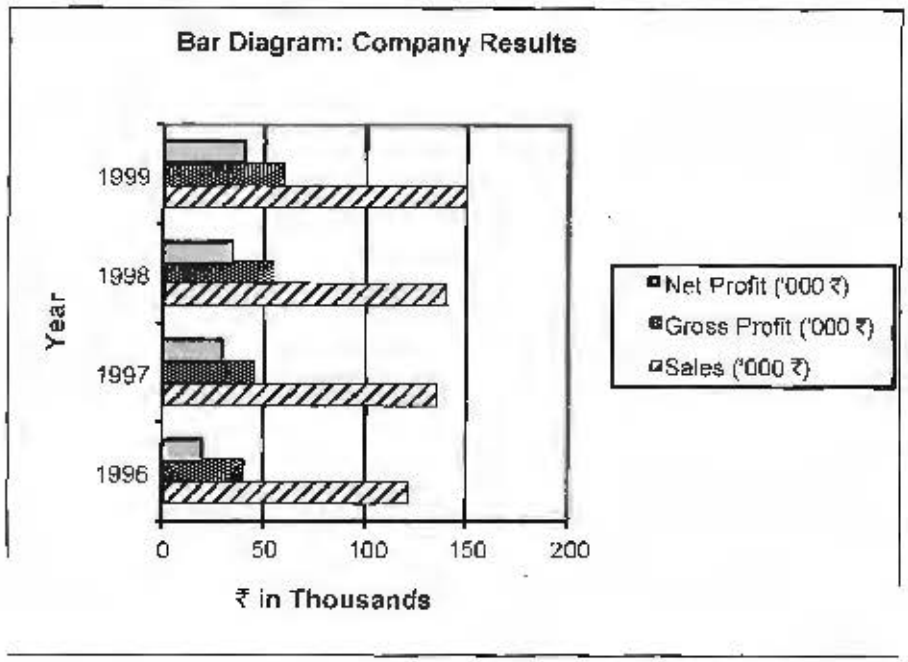
Example: Draw a multiple bar diagram to present following data. Also draw a multiple column diagram.

Years	Sales ('000 ₹)	Gross Profit ('000 ₹)	Net Profit ('000 ₹)
1996	120	40	20
1997	135	45	30
1998	140	55	35
1999	150	60	40

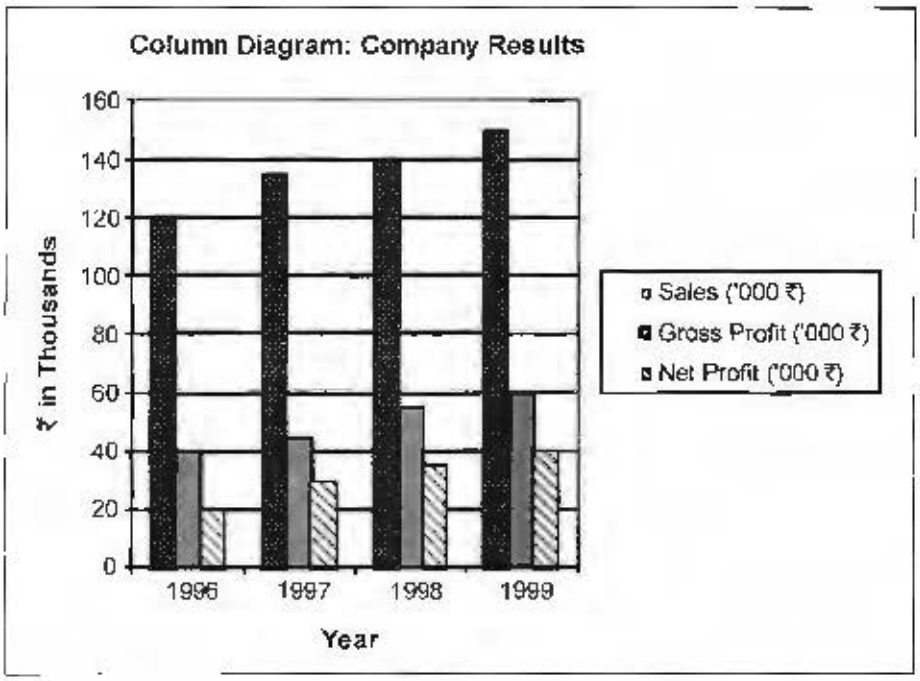
Solution:

Bar Diagram: We take year on the Y axis and rupees in thousands on the X axis. Then we draw horizontal bars with lengths proportional to the values of variables 'Sales', 'Gross Profits' and 'Net Profits'.

The bar diagram for the above data is as follows:



Column Diagram: We take year on the X axis and rupees in thousands on the Y axis. Then we draw vertical columns with lengths proportional to the values of variables 'Sales', 'Gross Profits' and 'Net Profits'. The column diagram for the above data is as follows:



11.5 LINE DIAGRAM

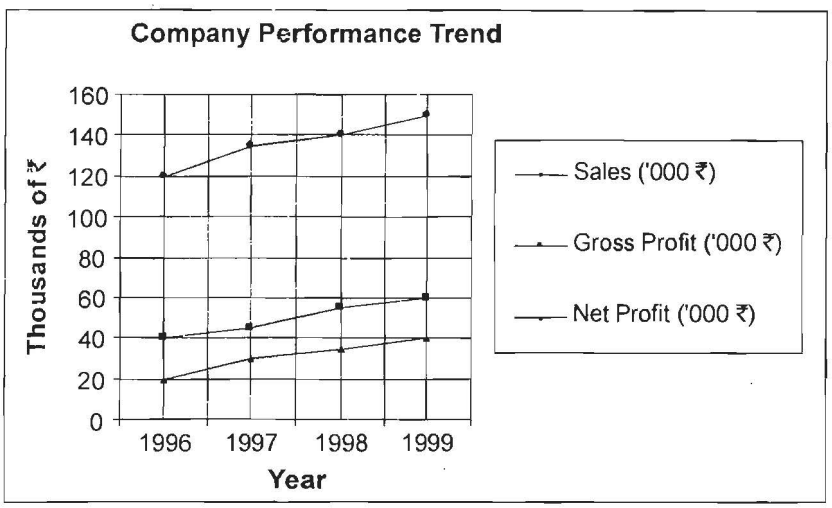
It is similar to the frequency polygon, where we plot one or more variables against one variable. One variable against which other variables are plotted is taken along the X axis. It is commonly used to depict the trends in anytime series data. We can show one or more variables like economic, market trends, financial results, etc. together so that these can be compared.

Example: Draw line diagram to present following data:

Year	Sales ('000 ₹)	Gross Profit ('000 ₹)	Net Profit ('000 ₹)
1996	120	40	20
1997	135	45	30
1998	140	55	35
1999	150	60	40

Solution:

We take year on the X axis and rupees in thousands on the Y axis. Then we plot the data points for the variables 'Sales', 'Gross Profits' and 'Net Profits'. These data points are then joined by straight lines to draw the line diagram. The line diagram for the above data is as follows:



11.6 HISTOGRAM

Besides the frequency polygon, histogram is one of the most popular and widely used graphical representations. It uses vertical bars whose height represents the frequency. In histogram, the vertical bars touch the neighbouring bars sharing one edge. Hence, if the data is of inclusive classes, it needs to be converter to exclusive classes so that the class boundaries overlap. Sometimes, we also use histograms superimposed with frequency polygons. This helps interpolation of data, at the same time retaining the attractive representation of histogram.

Example: In a city, the income tax department had the data as follows for the number of tax payers along with the range of income tax they paid for a particular year. Represent the data graphically with the help of a histogram.

Tax paid (in ₹ '000)	20-24	25-29	30-34	35-39	40-44	45-49	Total
Number of Tax Payers	45	130	200	65	45	15	500

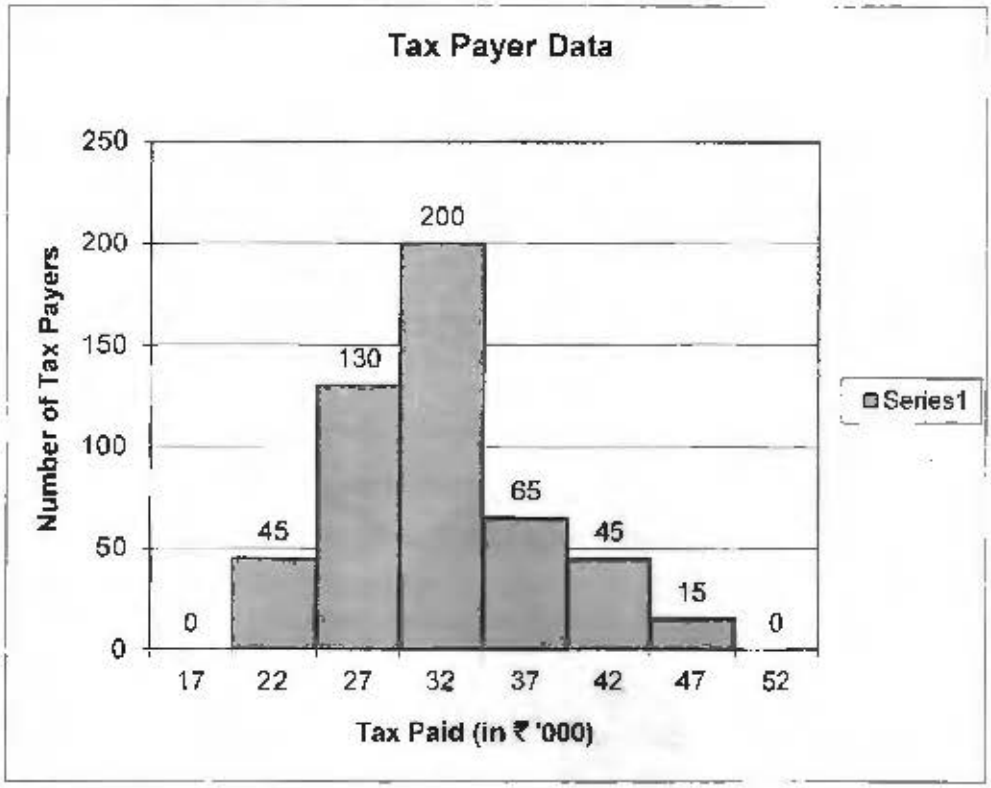
Solution:

For plotting the data, we will first convert the data as exclusive classes. This is done by increasing the upper limits and decreasing the lower limits by an amount equal to half of the difference between upper limit of any class and lower limit of the subsequent class. This makes the class boundary to join. Then class boundaries of tax paid classes are plotted on the X axis and number of tax payers on the Y axis.

Then vertical bars are drawn of widths equal to classes and heights equal to the frequencies of corresponding classes. This is depicted as follows:

Tax paid (in ₹ '000)	19.5-24.5	24.5-29.5	29.5-34.5	34.5-39.5	39.5-44.5	44.5-49.5
Number of Tax Payers	45	130	200	65	45	15

The histogram is shown below:



11.7 PIE DIAGRAM

Pie diagram is very popular visual representation in business reports, when manager wants to show the share of various categories in total. Total is represented as a circle. Each category is depicted as a sector with its central angle proportional to its share. The share percent in total of each category is converted to a sector angle using formula:

$$\text{Sector Angle in degrees} = \frac{\text{Share Percentage}}{100} \times 360$$

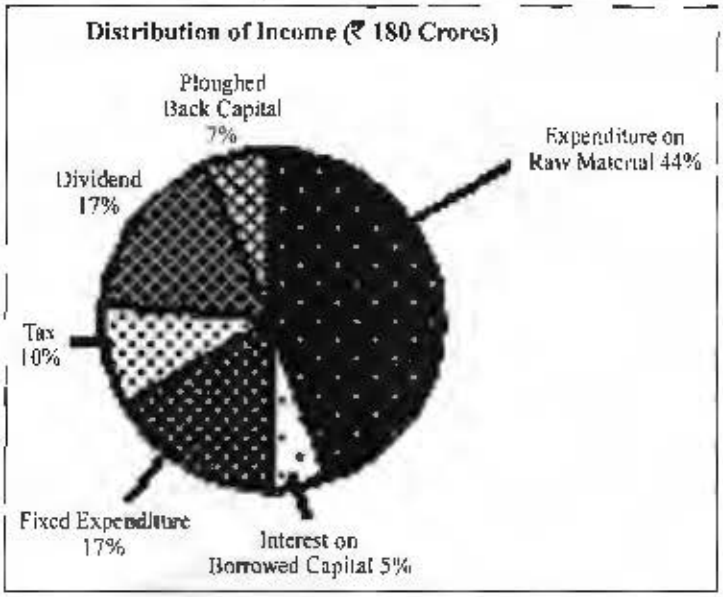
Other variations of pie diagrams are doughnut diagrams and exploded pie diagram. These are shown below.

Example: ABC Company has a total income of ₹180 crores. Out of this it has paid ₹10 crores as interest on borrowed capital. It has spent ₹80 crores for raw materials and other running expenditure. Its fixed costs (overheads) are ₹30 crores. On the net profit it has to pay the tax at the rate of 30% on net profit. Further, the board of directors decides to pay the dividend at the rate of 50% on the paid up capital of ₹60 crores. The remaining amount is retained as profit ploughed back. Depict the data as a pie diagram, doughnut diagram and exploded pie diagram.

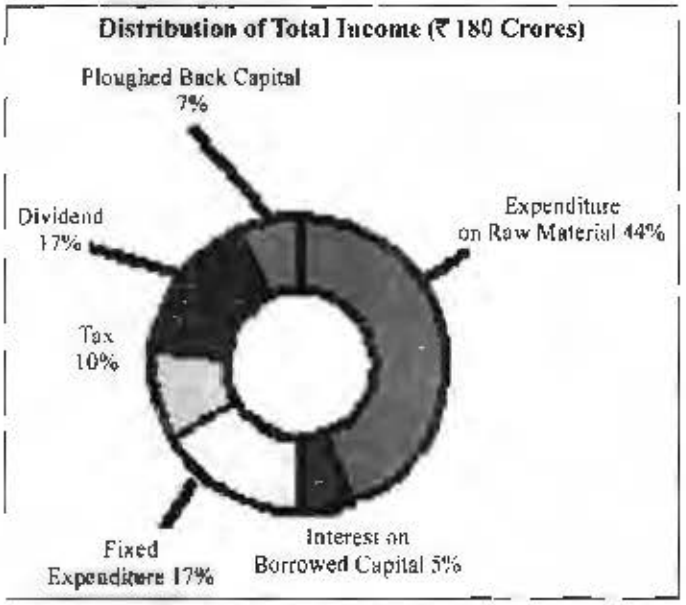
Solution:

We need to calculate the proportion of each category of the income distribution. Then we convert it as degrees, with total is 360 degrees. The calculations are shown as follows:

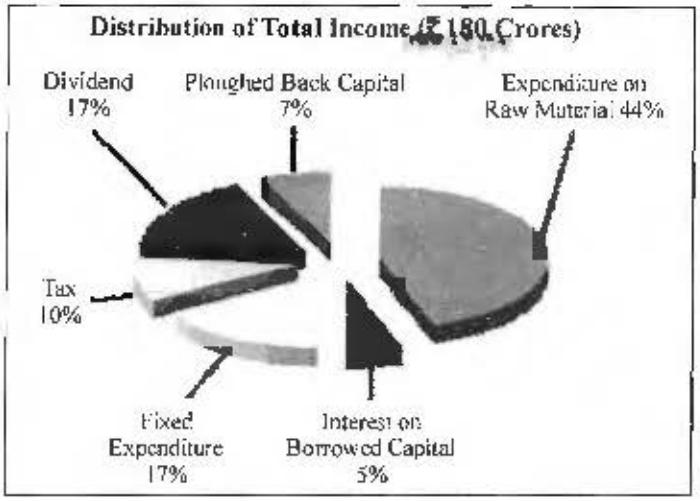
		Amount in (₹ in Crore)	Proportion to Total Income	Equivalent Angle
Total Income	(a)	180	1	360
Expenditure on Raw Material	(b)	80	0.44	160
Interest on Borrowed Capital	(c)	10	0.056	20
Fixed Expenditure	(d)	30	0.167	60
Net Profit [a – b – c - d]	(e)	60		
Tax $\left[\frac{30}{100} \times e\right]$	(f)	18	0.1	36
Dividend	(g)	30	0.167	60
Ploughed Back Capital	(h)	12	0.067	24



Pie Chart



Doughnut Diagram



Exploded Pie Diagram

11.8 FREQUENCY POLYGON

Frequency polygon is used for presenting the frequency distribution in graphical form. This can be used for discrete distribution with grouped as well as ungrouped data. This can also be used for continuous data by converting it to approximate discrete data through grouping. In all these cases, values of variables are represented on the X axis and their frequency (number of occurrences) on the Y axis. In case of probability distributions, we use the probability as frequency by choosing a suitable scale on the Y axis. For plotting the frequency polygon, we need to choose appropriate scale and origin so that the main data features occupy the reasonable area on the paper. This helps the readability. Although usually the scale chosen is linear, however, depending on the data type we could use logarithmic or other types of scale. Examples of these are audio noise plots, earthquake intensity plots, etc. Once the scale and origin is chosen, we need to draw grid lines (or use graph paper with grid lines) to facilitate accurate plotting. Then we take each data point and mark it on the graph. In case of a grouped data, we use class marks (mid points of the class intervals) as variable values on the X axis. These data points are joined by straight lines or a smooth curve to get frequency polygon or the frequency distribution in graphical form. To plot frequency distribution we can also join the data points by smooth lines.

Example: In a city, the income tax department had the data as follows for the number of tax payers along with the range of income tax they paid for a particular year. Represent the data graphically with the help of a frequency polygon and frequency distribution chart.

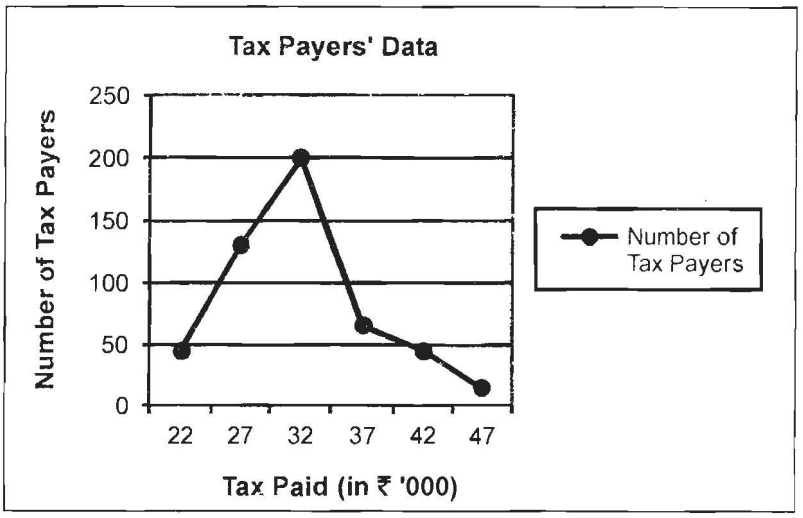
Tax paid (in ₹ '000)	20-24	25-29	30-34	35-39	40-44	45-49	Total
Number of Taxpayers	45	130	200	65	45	15	500

Solution:

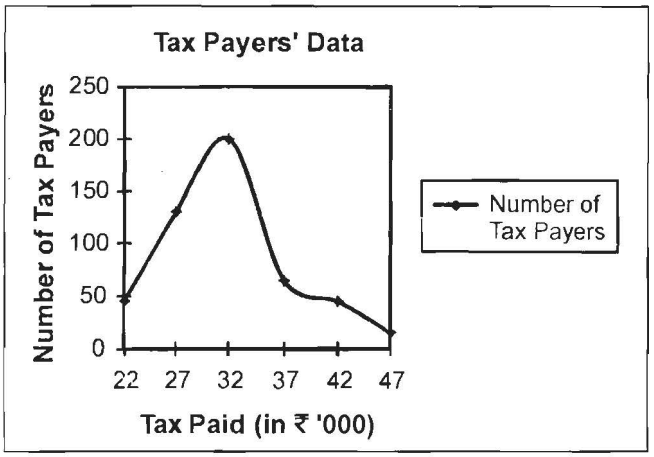
For plotting the data, we will use class marks of tax paid classes on the X axis and number of tax payers on Y axis. Thus, the points for plotting are as follows. Then we join these points by straight lines.

Value on X axis	22	27	32	37	42	47
Value on Y axis	45	130	200	65	45	15

The plot is shown below:



To draw the plot as frequency distribution, we follow the same procedure for plotting the data points. Then we join the data points with a smooth curve as shown below. This gives better interpolation results. It also helps in comparing it with standard distributions.



11.9 OGIVES

Ogives are used to present cumulative frequency of a distribution in graphical format. There are two kinds of ogives. 'Less than' ogive represents cumulative frequency just below the variable value plotted on X axis. On the other hand, 'More than' ogive plots the sum of the frequencies corresponding to above the variable value. For this, we first calculate 'Less than' and 'More than' cumulative frequencies for the entire variable values (corresponding to classes). Then we plot these as points on the graph with class marks along the X axis and cumulative frequencies ('Less than' or 'More than') along the Y axis. These points are then joined by a smooth curve like frequency distribution. The value of the variable (on the X axis) at an ordinate from the point where two ogives intersect is 'Median' i.e. mid-value of the data (more about Median is in next chapter). The following example demonstrates drawing of ogives.

Example: Before constructing a dam on a river the central water research institute performed a series of tests to measure the water flow, past the proposed location of the dam during the period of 246 days, when there was a sufficient flow of water.

The results of testing were used to construct the following frequency distribution.

River Flow (thousand cubic meters per min)	1001-1050	1051-1100	1100-1150	1151-1200	1201-1250	1251-1300	1301-1350	1351-1400
Number of Days (frequency)	7	21	32	49	58	41	27	11

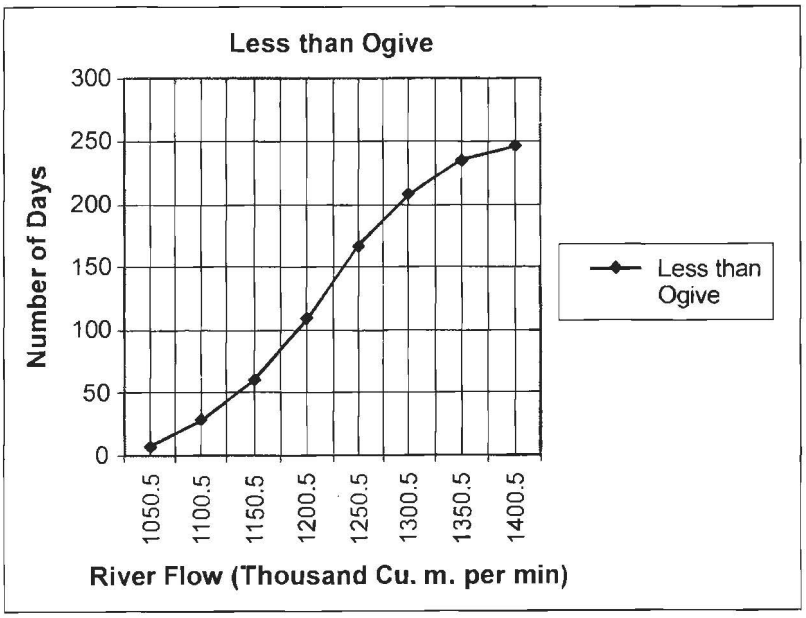
1. Draw ogive curves for the above data.
2. From the ogive curve estimate the proportion of the days on which flow occurs at less than 1300 thousands of cubic metres per minute.

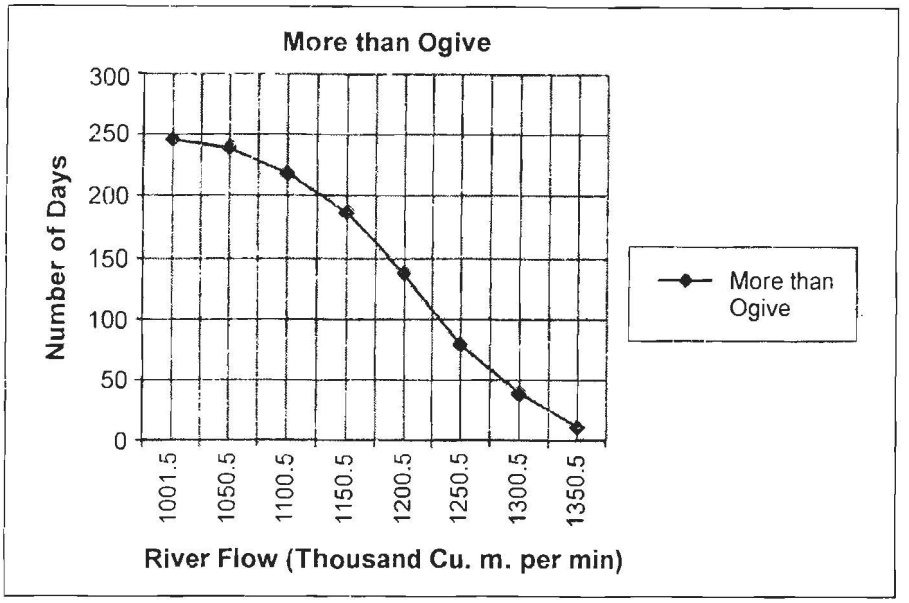
Solution:

First we calculate and prepare 'less than' and 'more than' frequency table as follows.

River Flow 1000 cu. m per min	No of Days	Upper Class Limit	'Less than' Frequency	Lower Class Limit	'More than' Frequency
1001-1050	7	1050.5	7	1001.5	246
1051-1100	21	1100.5	28	1050.5	239
1101-1150	32	1150.5	60	1100.5	218
1151-1200	49	1200.5	109	1150.5	186
1201-1250	58	1250.5	167	1200.5	137
1251-1300	41	1300.5	208	1250.5	79
1301-1350	27	1350.5	235	1300.5	38
1351-1400	11	1400.5	246	1350.5	11
Total	**246**				

1. Now we plot the ogives with class limits on the X axis and frequencies (less than or more than) on the Y axis, and joining the points with smooth curves. We also plot both ogives superimposed. The ogives are shown below.





2. From the 'less than' give we can read that the number of days on which flow occurs at less than 1,300 thousand of cubic metres per minute is 208.

   Thus, the proportion of days on which flow occurs at less than 1,300 thousand of cubic meters per minute is 0.846 or 84.6%.

Check Your Progress

Fill in the blanks:

1. A graphic presentation is used to represent _____ types of statistical data.
2. _____ diagrams are also known as bar diagrams.
3. _____ diagrams are normally used when the variations in the magnitudes of observations are very large.
4. In _____, the vertical bars touch the neighbouring bars sharing one edge.
5. _____ ogive represents cumulative frequency just below the variable value plotted on X axis.
6. _____ representation allows us to obtain the underlying information in one glance.

11.10 LET US SUM UP

- The charts help in grasping the data and analyze it qualitatively. This also helps managers to effectively present the data as a part of reports.
- Various types of chart are bar diagram, multiple bar diagrams, component bar diagram, deviation bar diagram, sliding bar diagram, Histogram and Pie charts.
- A graphic presentation is another way of representing the statistical data in a simple and intelligible form.
- There are two types of graphs which we have discussed, line graphs and ogives.

11.11 UNIT END ACTIVITY

Show the following data of expenditure of an average working class family by a suitable diagram.

Item of Expenditure		Percent of Total Expenditure
(i)	Food	65
(ii)	Clothing	10
(iii)	Housing	12
(iv)	Fuel and Lighting	5
(v)	Miscellaneous	8

11.12 KEYWORDS

Bar Graph: A graphical device for depicting data that have been summarized in a frequency distribution, relative frequency distribution or percent frequency distribution.

Histogram: A graphical presentation of a frequency distribution, relative frequency distribution or percent frequency distribution of quantitative data constructed by placing the class intervals in the horizontal axis and the frequencies on the vertical axis.

Relative Frequency Distribution: A tabular summary of data showing the fraction or proportion (relative frequency) of observations in the data set in each of several non-overlapping classes.

Line Graph: We plot one or more variables against one variable. One variable against which other variables are plotted is taken along the X axis. It is commonly used to depict the trends in anytime series data.

Ogives: Ogives are used to present cumulative frequency of a distribution in graphical format.

11.13 QUESTIONS FOR DISCUSSION

1. What is the difference between diagram and graphs?
2. Explain number of diagrams which can be used for presentation of data.
3. Discuss the importance of using frequency polygon.
4. What is ogives? Explain its two kinds.
5. The income of 12 workers on a particular day was recorded as given below. Represent the data by a line diagram.

   S. No. of Workers : 1 2 3 4 5 6 7 8 9 10 11 12

   Income (in ₹) : 25 35 30 45 50 55 40 50 60 55 40 35

6. Represent the following data by a suitable diagram.

   Years : 1987 1988 1989 1990 1991

   C.F.A Enrolments : 7300 9400 12100 14600 16700

7. Represent the following data, on revenue and costs, of a company during July 1991 to December 1991 by a net balance chart.

Months (1991) :	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Revenue in (₹'000) :	20	25	18	20	23	22
Cost in (₹'000) :	18	22	20	21	19	20

8. Draw 'less than' and 'more than' ogives for the following distribution of monthly salary of 250 families of a certain locality.

Income Intervals :	0-500	500-1000	1000-1500	1500-2000
No. of Families :	50	80	40	25
Income Intervals :	2000-2500	2500-3000	3000-3500	3500-4000
No. of Families :	25	15	10	5

Check Your Progress: Model Answer

1. Two
2. One-dimensional
3. Three-dimensional
4. Histogram
5. Less than
6. Pictorial or Graphical

11.14 REFERENCE & SUGGESTED READINGS

- Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2020). **Statistics for Business and Economics** (14th ed.). Cengage Learning. ISBN: 9780357114474
- Jaggia, S., Kelly, A., & Lertwachara, K. (2021). **Essentials of Business Statistics** (2nd ed.). McGraw-Hill Education. ISBN: 9781260205799
- Bowerman, B. L., O'Connell, R. T., Murphree, E. S., & Orris, J. B. (2018). **Essentials of Business Statistics** (6th ed.). McGraw-Hill Education. ISBN: 9781259549939
- Doane, D. P., & Seward, L. E. (2019). **Applied Statistics in Business and Economics** (6th ed.). McGraw-Hill Education. ISBN: 9781260224035
- Gupta, S. C., & Kapoor, V. K. (2018). **Fundamentals of Applied Statistics** (4th ed.). Sultan Chand & Sons. ISBN: 9788180547967
- Keller, G. (2022). **Business Analytics: A Data-Driven Decision Making Approach** (1st ed.). Cengage Learning. ISBN: 9780357717828
- Evans, J. R. (2020). **Business Analytics: Methods, Models, and Decisions** (2nd ed.). Pearson. ISBN: 9780135231679