UNIT 1 MEANING AND SCOPE OF STATISTICS

Structure

1.0 OBJECTIVES

After studying this unit, you should be able to:

- define the word 'statistics'
- distinguish between descriptive and inferential statistics
- describe the different functions of statistics
- explain the importance of statistical methods in different fields
- appreciate the limitations of statistical methods
- explain the reasons for distrust in statistics.

1.1 INTRODUCTION

Statistics is not a new discipline but is as old as the human activity itself. Its sphere of utility, however, has been increasing over the years. In the olden days, it was considered as the 'science of statecraft' and was regarded as a by-product of the administrative activity of the State thereby limiting its scope. The governments in those days used to keep records of population, birth, deaths, etc., for administrative purposes. In fact, the word 'statistics' seems to have been derived from the Latin word 'status' or Italian word 'statista' or the German word 'Statistik' each of which means a political state. Statistical methods are now widely used in various diversified fields such as agriculture, economics, sociology, business management, etc. In this unit you will study the meaning and definition of statistics, distinction between descriptive and inferential statistics, functions of statistics, importance and limitations of statistics, and distrust of statistics.

1.2 MEANING OF STATISTICS

The word 'statistics' has been used in a variety of ways. Sometimes it is used in the plural sense to refer to numerical statements of facts or data. On the other hand it is also used in the singular sense to refer to a subject of study like any other subject such as (mathematics, economics, etc. For instance, when we refer to a few 'statistics' relating 'to our country like — thereare 932 females per 1,000 males in India, the per capita national product at current prices has increased from Rs. 246 in 1950-51 to Rs. 2,596 in 1985-86 — we are using the word statistics in the plural sense (meaning data). To prepare these numerical statements, one must be familiar with those methods and techniques which are used in data collection, organisation, presentation, analysis and interpretations. A study of these methods and techniques is the science of statistics. The use of the word statistics here is in the singular sense. In this sense the word statistics means statistical methods or the science of statistics. Now let us study in detail about these two approaches.

1.2.1 Statistics Defined in Plural Sense

Statistics has been defined differently by different writers. According to Webster "statistics are the classified facts representing the conditions of the people in a state... specially those facts which can be stated in numbers or any tabular or classified arrangement." To Bowley statistics are "numerical statements of facts in any department of enquiry placed in relation to each other." According to Yule and Kendall statistics means "quantitative data affected to a marked extent by multiplicity of causes." These definitions are too narrow as they confine the scope of statistics to only such facts or figures which either relate to the conditions of the people in a state or specify some characteristics of the data.

A more comprehensive definition of statistics was given by Horace Secrist. According to him statistics means "aggregate of facts affected to marked extent by 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." This definition is quite comprehensive and points out the characteristics that numerical facts (data) must possess so that they may be called statistics. Let us discuss about these characteristics one by one.

a) They must be aggregate of facts: Individual and isolated figures cannot be called statistics. They should form a part of aggregate of facts relating to any particular field of enquiry. For example, Ram's monthly income is Rs. 2,000. This is not a statistical statement. However, when we say that monthly incomes of Ram, Mohan, and Sohan are Rs. 2,000, 2,500 and Rs. 3,000 respectively, they will be called statistics.

b) They are affected by multiplicity of factors: There are several factors that affect a phenomenon. For instance, the consumption of a household on any item would be affected by several factors as income, taste, education, etc. Similarly, production of wheat is affected by soil, seeds, rainfall, temperature, etc. The data relating to such phenomenon can be called statistics. But if we write the numbers one to ten along with their squares, then these figures though more thap one, cannot be called statistics. These figures are not affected by multiplicity of causes.

c) They must be numerically expressed: To call a statement as statistics, it must be expressed numerically. Therefore, qualitative characteristics such as beauty, colour of eyes, etc., cannot be measured directly and hence, in general, they do not fall under the purview of statistics. We have to quantify these characteristics before they become statistics. For example, in a college we may count the number of girls having black eyes or blue eyes or brown eyes.

d) They are enumerated or estimated according to a reasonable standard of accuracy: Statistics are either enumerated or estimated, but reasonable standards of accuracy must be maintained. The degree of accuracy will depend on the nature and the object of the study being undertaken. Suppose, as the Principal of a College you are interested in understanding the average level of performance of the students who take admission to B.Com. class. For this purpose you must collect the marks obtained by the students at the senior secondary level. It may be done in two ways. First you can have a complete enumeration of the marks of all the students and derive their average. Secondly if complete enumeration is not possible due to some reason, you may select a sample. On the basis of the result of the sample, you may then estimate the average level of performance of all students. Thus, statistics may be obtained by enumeration or estimation. Let us take another example to understand the point reasonable standard of accuracy. If you are estimating the total production of food crop in India the appropriate units of measurement (or the level of accuracy) may be lakhs of tons. But if you are reporting the total production of gold, the appropriate unit of measurement may be kilograms. Thus, degree of accuracy depends on nature and objective of the study.

e) They must be collected in a systematic manner for a predetermined purpose: The data should be collected in a systematic manner. Data collected in a haphazard mannek will not serve much purpose. The purpose for which data is collected, must be decided in advance. The purpose should be specific and well-defined. If the purpose of the enquiry is not specified, either we may collect too much or too little data.

f) They must be placed in relation to each other: The numerical facts should be comparable if they are to be called statistics. For instance, statistics on production and export of an item during a year are related. What they put together are statistics. But if you have three figures: 1) production of rice in India in 1986, 2) number of children born in USA in 1987, and 3) number of cars registered in UK in 1988. These figures may be facts alright, but taken together they cannot be called statistics as they have no relation among themselves.

It is, thus, clear that all statistics are numerical statements of facts but all **numerical statements** of facts are not statistics. They will be called statistics only if the above characteristics are present in them.

1.2.2 Statistics Defined in Singular Sense

Numerical information must be collected, organised, presented, analysed and interpreted if it has to be used for making wise decisions. We require methods that help us in this regard. Thus, statistics, when used in the singular sense, has been defined as a body of methods which provides tools for data collection, analysis and interpretation. Here too, different writers have interpreted statistics differently. Now let us also discuss about some of these definitions.

Bowley, for instance, has given a number of definitions. But none of them is comprehensive. They in fact point to the development of science of statistics over time. Some of these definitions are:

i) Statistics may be called the science of counting.

ii) Statistics may rightly be called the science of averages.

iii) Statistics is the science of measurement of social organism, regarded as a whole in all manifestations.

Croxtoa and Cowden have given a simple and precise definition of statistics. According to them "statistics may be defined as the collection, presentation, analysis and interpretation of numerical data."

The definition given by Selligrnan is equally simple but comprehensive. According to him "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."

The last two definitions are quite precise, comprehensive and point out the scope of statistical methods. The science of statistics teaches us the methods and techniques which are required for 1) collection of data, 2) classification and tabulation of data, 3) presentation of data, 4) analysis of data, and 5) interpretation of data.

Thus, from the above discussion, we can conclude that the word 'statistics' may be used either in plural sense to refer to data or in singular sense to refer to a body of methods for making wise decisions in the face of uncertainty.

1.3 DESCRIPTIVE AND INFERENTIAL STATISTICS

As you know, when used in singular sense, statistics is a study of the principles and methods used in the collection, presentation, analysis and interpretation of data in any sphere of enquiry. These methods and techniques are so diverse that statisticians generally categorise them into two: 1) descriptive statistics, and 2) inferential statistics.

Descriptive Statistics refer to various measures that are used to describe the characteristicfeatures of the data. Such measures include measures of central tendency, measures of dispersion, etc. Graphs, tables and charts that display data are also examples of descriptive statistics. Suppose the number of first year B.Com. students is 100 and you compute the average marks of these students. Here you are using descriptive statistics. Similarly, when you are computing the average marks of a sample of 25 students from the same class but without attempting any generalisation about the entire class, you are still using descriptive statistics.

Inferential Statistics on the other hand refer to statistical process of drawing valid inferences about the characteristics of population data on the basis of sample data. The

word population in statistics does n mean only human population. It stands for totality of items related to any field of study. If the teacher, in the above example, decides to estimate the average marks of the entire class on the basis of the sample average, we would say that he is using inferential statistics. It is noteworthy that most of the time we use sample data to understand the features of the population data. Inferences about population drawn from sample measures may involve some error or discrepancy. The magnitude of such errors can be estimated on the basis of probability theory.

Check Your Progress A

1 Are the following statements statistical data?

i) Weekly wages of 100 workers of a factory.

ii) Height of Ram is six feet.

iii) Mohan's weight is 70 Kgs, Sohan's height is 6.2 feet, and Ram's monthly income is Rs. 1,500.

iv) Sales of a company during the past 10 years.

2 Comment on the following statements in not more than one line.

i) Webster and Secrist defined descriptive statistics.

.....

ii) Definition of statistics given by Yule and Kendall is contained in the one by Secrist.

.....

iii) Qualitative data cannot be studied under statistics.

.....

iv) Methods of statistics relate to collection and analysis of the data only.

.....

v) The definition of science of statistics by Bowley covers the different stages of statistical methodology.

.....

vi) Inferential statistics is related to the study of samples.

.....

1.4 FUNCTIONS OF STATISTICS

You have studied the meaning and definitions of statistics. You have also learnt the difference between descriptive statistics and inferential statistics. Let us now discuss some of the important functions of statistics:

1 **To present facts in a proper form:** Statistical methods present general statements in a precise and definite form. For example, you may say that in India average yield of cotton per hectare is 180 Kg. This statement is more precise and convincing than saying that the average yield of cotton in India is very low.

2 **To simplify unwieldy and complex data:** Statistical methods simplify unwieldy and complex data to make them understandable easily. The raw data is often unintelligible. One cannot grasp their characteristics unless the data is classified according to some common characteristics, Suppose, you are given the weekly wages of 1,000 workers in a factory. You will not be in a position to draw any inference from the data unless they are condensed through classification such as the following:

Weekly Wages (Rs.)	No. of Workers
Below-600	100
600-700	200
700-800	400
800-900	200
Above 900	100
Total:	1000

3 **To provide techniques** for making **comparison:** The primary purpose ot statistics is to. facilitate a comparative study of different phenomena either over time or space. For instance, the estimation of national income is not done for its own sake. But it is done to compare the income over time to get an idea whether the standard of living of people is rising or not. Suppose, as compared to 1987, the per-capitaincome in India has increased by 10% in 1988. On the basis of this information, we shall be in a position to throw some light on the standard of living of an Indian in 1988.

4 **To formulate** policies in different fields: Statistical methods are very useful in formulating various policies in social, economic, and business fields. The government, for instance, utilises vital statistical data for formulating family planning programme. Similarly, the government utilises the information on consumer price indices for granting dearness allowance to its employees.

5 **To** study relationship between different phenomena: Statistical measures such as correlation and regression are used to study relationships between variables. Such relationships are important for making decisions. For instance, you may find a relationship between the demand of a product and its prices. In general, if the prices rise, the demand for the product is likely to decline.

6 **To** forecast future values: Some of the statistical techniques are used for forecasting future values of a variable. On the basis of sales figures of the last 10 years, a marketing manager can estimate the likely demand for his product during the next year.

7 **To** measure uncertainty: With the help of probability theory, you can measure the chance of occurrence of uncertain event. Probability concepts are quite useful in decision-making. Suppose, if you are interested in estimating the chance of your passing the B.Com examination, you may get an idea about it by studying the pass percentages of students during the last 10 years.

8 To test a **hypothesis:** Statistical methods are extremely useful in formulating and testing hypotheses and for the development of new theories. For instance, a company is desirous of knowing the effectiveness of its new drug to control malaria. It could do so by using a statistical technique called Chi-square Test.

9 To draw valid inferences: Statistical methods are also useful in drawing inferences regarding the characteristics of the universe (population) on the basis of sample data.

1.5 IMPORTANCE OF STATISTICS

In the ancient times statistics was used as the science of statecraft only. Data on a wide range of activities such as population, births and deaths were collected by the State for administrative purposes. However, in recent years, the scope of statistics has widened considerably to bring to its fold social and economic phenomena. The developments in the statistical techniques over the years also widened its scope considerably. It is no longer considered to be a by-productof the administrative setup of the State but now it embraces practically all sciences, social, physical, and natural sciences. As a matter of fact, now statistics finds its applications in various diversified fields such as agriculture, business and industry, sociology; economics, biometry, etc. Thus, these days statistics finds its application in almost all spheres of human activity.

Statistics and State

In earlier times, the role of the State was confined to the maintenance of law and order. For that purpose, it used to collect data relating to manpower, crimes, income and wealth, etc., for formulating suitable military and fiscal policies. But the role of State has enlarged considerably with the inception of the concept of Welfare State. Thus, today statistical data relating to prices, production, consumption, income and expenditure, etc., are extensively used by the governments worldover for formulating their economic and other policies. To raise the standards of living of its population, developing countries such as India are following the policy of planned economic development. For that purpose the government must base its decisions on correct and sound analysis of statistical data. For instance, in formulating its five year plans, the government must have an idea about the availability of raw materials, capital goods,

financial resources, the distribution of population according to various characteristics such as age, sex, income, etc., to evolve various policies.

Statistics in Economics

Statistical analysis is immensely useful in the solution of a variety of economic problems such as production, consumption, distribution, etc. For example, an analysis of data on consumption may reveal the pattern of consumption of various commodities by different sections of the society. Data on prices, wages, consumption, savings and investment, etc., are vital in formulating various economic policies. Likewise, data on national income and wealth are useful in formulating policies for reducing disparities of income. Use of statistics in economics has led to the formulation of several economic laws such as Engel's Law of Consumption, Law of Income Distribution, etc. Statistical tools of index numbers, time series analysis, regression analysis, etc., are vital in economic planning. For instance, the consumer price index is used for grant of dearness allowance (DA) or bonus to workers. Demand forecasting could also be made by using time series analysis. For testing various economic hypotheses, statistical data is now being increasingly used.

Statistics in Business and Management

With the growing size and increasing competition, the activities of modern business enterprises are becoming more complex and demanding. The separation of ownership and management in the case of big enterprises has resulted in the emergence of professional management. The success of the managerial decision-making depends upon the timely availability of relevant information much of which comes from statistical data. Statistical data has, therefore, been increasingly used in business and industry in all operations like sales, purchases, production, marketing, finance, etc. Statistical methods are nbw widely applied in market and production research, investment policies, quality control of manufactured products, economic forecasting, auditing and many other fields. One element common to all problems faced by managers is the need to take decisions under uncertainty. And statistical methods provide techniques to deal with such situations. It is, therefore, not surprising when Wallis and Roberts say that "statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty."

Check Your Progress B

1 Enumerate the functions of statistics.

.....

.....

.....

.....

2 Write brief comments in one line on the following statements.

i) Statistics only perform the function of simplifying complexities.

.....

ii) Statistics help in testing the laws of other sciences.

.....

iii) Future course of events is uncertain, so statistics can hardly be of any help in their study.

.....

iv) Planning is not conceivable without statistics.

.....

v) A personnel officer of a big corporation can draw a workable personnel plan without the knowledge of statistics.

.....

1.6 LIMITATIONS OF STATISTICS

We have discussed the importance and functions of statistics. Now we shall discuss about the limitations of statistics. The following are some of the limitations of statistical methods which should be kept in mind while using them:

1 Statistics deals only with the quantitative characteristics: Statistics deals with facts which are expressed in numerical terms. Therefore, those phenomena that cannot be described in numerical terms do not fall under the scope of statistics. Beauty, colour of eyes, intelligence, etc., are qualitative characteristics and hence cannot be studied directly. These characteristics can be studied only indirectly, by expressing them numerically after assigning particular scores. For example, we can study the level of intelligence of a group of persons by using intelligence quotients (I.Q's).

2 Statistics does not deal with individuals: Since statistics deals with aggregate of facts, a single and isolated figure cannot be regarded as statistics. For example, the height of one individual is not of much relevance but the average height of a group of people is relevant from statistical point of view. In this context, you may recall the definition given by Secrist here.

3 Statistical laws are not exact: Unlike the laws of natural sciences, statistical laws are not exact. They are true under certain conditions and always some chance factor is associated with them for being true. Therefore, conclusions based on them are only approximate and not exact. They cannot be applied universally. Laws of pure sciences like Physics and Chemistry are universal in their application.

4 Statistical results are true only on an average: Statistical methods reveal only the average behaviour of a phenomenon. The average income of employees of a company will, therefore, not throw much light on the income of a specific individual. They are therefore, useful for studying a general appraisal of a phenomenon.

5 Statistics is only one of the methods of studying a problem: A problem can be studied by several methods. Statistical methods arc only one of them. Under all circumstances, statistical tools do not provide the best solution. Quite often it is necessary to consider a problem in the light of social considerations like culture, region, etc. Therefore, statistical conclusions need to be supplemented by other evidences.

6 Statistics can be misused: The various statistical methods have their own limitations. If used without caution they are subject to wrong conclusions. So one of the main limitations of statistics is that, if put into wrong hands, it can be misused. This misuse can be, at times, accidental or intentional. Many government agencies and research organisations are tempted to use statistics to misrepresent the facts to prove their own point of view. Suppose you are told that during a year the number of car accidents in a city by women drivers is 10 while those committed by men drivers is 40. On the basis of this information, you may conclude that women are safe drivers. If you conclude like that you are misinterpreting the information. You must know the total number of drivers of both types before you could arrive at a correct conclusion.

1.7 DISTRUST OF STATISTICS

Despite its importance and usefulness the science of statistics is looked upon with suspicion. Quite often it is discredited, by people who do not know its real purpose and limitations. We often hear statements such as:

"There are three types of lies: lies, damned lies, and statistics". "Statistics can prove anything". "Statistics cannot prove anything". "Statistics are lies of the first order". These are expressions of distrust in statistics. By distrust of statistics, we mean lack of confidence in statistical data, statistical methods and the conclusions drawn. You may ask, why distrust in statistics? Some of the important reasons for distrust in statistics are as follows:

1 Arguments based upon data are more convincing. But data can be manipulated according to wishes of an individual. To prove a particular point of view, sometimes arguments are supported by inaccurate data.

2 Even if correct figures are used, they may be incomplete and presented in such a manner that the reader is misled. Suppose, it has been found that the number of traffic accidents is lower in foggy weather than on clear weather days. It may be concluded that it is safer to drive in fog. The conclusion drawn is wrong. To arrive at a valid conclusion, we must take into account the difference between the rush of traffic under the two weather conditions.

3 Statistical data does not bear on their face the label of their quality. Sometimes even unintentionally inaccurate or incomplete data is used leading to faulty conclusions.

4 The statistical tools have their own limitations. The investigator must use them with precaution. But sometimes these tools or methods are handled by those who have little or no knowledge about them. As a result, by applying wrong methods to even correct and complete data, faulty conclusions may be obtained. This is not the fault of statistical methods, but of the persons who use them.

We may conclude by taking an illustration. Suppose a child cuts his finger with a knife. His parents started blaming the knife. Here the fault does not lie with the knife but with the child who misused the knife. It should be kept in mind that statistics neither proves anything nor disproves anything. It is only a tool (i.e. a method of approach) which should be used with caution and by those who are knowledgeable in the subject.

1.8 LET US SUM UP

The word statistics can be used either plural sense or in singular senseß. When used in plural sense, the word statistics refers to numerical statements of facts or data. To be called statistics, numerical data should possess the following characteristics : 1) they must be aggregate of facts, 2) they must be affected by multiplicity of factors, 3) they must be numerically expressed, 4) they must be enumerated or estimated according to areasonable standard of accuracy, 5) they must be collected in a systematic manner for a predetermined purpose, and 6) they must be placed in relation to each other. The word statistics, when used in singular sense, refers to a body of knowledge which provides methods and techniques required for, 1) collection of data, 2) classification and tabulation of data, 3) presentation of data, 4) analysis of data, and 5) interpretation of data.

Statistical methods can be divided into: 1) descriptive statistics, and 2) inferential statistics. Statistical methods are helpful in: 1) presenting facts in proper form, 2) simplifying unwieldy and complex data, 3) providing techniques for making comparison, 4) formulating policies in different fields, 5) studying relationships between different phenomena, 6) forecasting future values, 7) measuring uncertainty of events, 8) testing statistical hypotheses, and 9) drawing valid inferences. Statistical methods are useful in various fields such as state administration, economics, business management, etc. With the growing complexity of managing today's business, statistical tools are proving quite handy and useful in the decision-making process. However, there are limitations in using these tools. Statistics does not study qualitative phenomenon nor does it study individuals. Statistical laws are not exact and may be misused. A blind fold application of these tools, particularly by those who are no fully conversant with them, has resulted in lot of distrust. The science of statistics is a useful servant to those who understand its proper use.

1.9 KEY WORDS

Descriptive Statistics: Refers to methods and techniques of summarising and describing the characteristics of the data.

Inferential Statistics: Refers to those methods which are helpful in drawing inferences about the characteristics of the population on the basis of sample data.

Statistical Data: Information expressed in quantitative or numerical form is called statistical data. All statistical data is numerical statements of facts but all numerical statements of facts are nor statistics. Numerical statements must possess certain

characteristics in order that they may be called data.

Statistical Methods: A body of methods and principles that are helpful in the collection, summarisation, description, analysis and interpretation of numerical data.

Statistics: When used in plural sense, refer to numerical statements of facts or data. When used in singular sense, refers to a body of methods which provides tools for data collection, analysis and interpretation.

1.10 ANSWERS TO CHECK YOUR PROGRESS

A 1 i) Yes ii) No iii) No iv) Yes

2 i No. Their definitions related to data.
ii) Yes.
iii) Yes. Not directly, after quantifying them.
iv) No. Other aspects are also there.
v) Yes.
vi) No. They are methods to derive population values from sample results.

B 2 i) No. There are other functions also.
ii) Yes. By collecting relevant data.
iii) No. Probability theory and methods of forecasting helps.
iv) Yes. Lots of Statistics are required.
v) No. Statistical methods will be used.

1.11 TERMINAL QUESTIONS

1 "Statistics are numerical statements of facts but all facts numerically stated are not statistics." Comment.

2 Define statistics and discuss the various functions of statistics.

3 Discuss the usefulness of statistics and explain the limitations of statistics.

4 What do you understand by distrust of statistics? Is the science of statistics to be blamed for it?

Note: These questions will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the university. These are for your practice only.

UNIT 2 ORGANISING A STATISTICAL SURVEY

Structure

2.0 OBJECTIVES

After studying this unit, you should be able to:

- describe the steps involved in a statistical survey
- distinguish between primary data and secondary data, and identify their resources
- state the salient features of different types of enquiries
- explain the law of statistical regularity and the law of inertia of large numbers
- appreciate the significance of statistical unit and the degree of accuracy.

2.1 INTRODUCTION

In the previous unit you learnt the meaning and scope of statistics. We have also discussed the importance and limitations of statistics. In this unit you will study the various steps in organising a statistical survey, the sources of data, different types of enquiries, and some laws connected with them. You will also learn certain other aspects like statistical units and degree of accuracy which are to be kept in mind, while conducting a survey.

2.2 STEPS IN STATISTICAL SURVEY

When we conduct a statistical survey, there are certain steps which are to be followed in a sequential order, Unless we follow these steps systematically, we may not be able to achieve purposeful results from the survey. The important steps concerning a statistical survey are presented below in a sequential order:

1 Defining the problem
2 Determining the objective and scope
3 Preliminaries to the collection of data
   i) Source of data

ii) Type of enquiry
iii) Statistical unit
iv) Degree of accuracy

4 Collection of data
5 Editing of data
6 Classification and tabulation of data
7 Analysis of data
8 Interpretation of data
9 Writing the report

Now let us discuss briefly about all these steps.

1 Defining the Problem

In any statistical survey, first of all, we have to state very clearly the problem to be investigated. Clear definition of the problem is of utmost importance as it is helpful to identify the relevant data. As you know, statistics is concerned with the aggregate of facts which are numerically expressed. Therefore, while defining the problem we should ensure the possibility of quantitative measurement.

2 Determining the Objective and Scope

After defining the problem, the next step is to determine the objective and scope of the survey. If the objective of the survey is clearly stated it serves as a guide in the collection of required information. If objective is stated precisely, you can also adopt a uniform approach to different problems which arise during the course of survey.

Scope of survey refers to the area to be covered, the period of study, the population or items to be covered, the type of information to be collected, etc. All these depend on the problem to be investigated and the objective of the study. The accuracy of the final result depends on correct assessment of all the items mentioned above. So you must determine the scope of the survey precisely.

3 Preliminaries to the Collection of Data

Before you proceed to collect the data, you should accomplish the following preliminaries:

i) Source of Data: You should decide about the sources from which the data is to be collected. For the collection of data, there are two approaches: (1) you may collect the data yourself, or (2) you may take the data from published sources. The data collected for the first time by the investigator is known as primary data. On the other hand if you use the data already collected by someone else, such data is referred to as secondary data. You will study in detail about these two types of data later on in this unit.

ii) Type of Enquiry: You should determine the type of enquiry to be conducted. There are different types of statistical enquiries such as census or sample, initial or repetitive, direct or indirect, regular or ad-hoc, confidential or non-confidential, official or non-official, etc. A decision about the most suitable type of enquiry for the proposed study should be taken keeping in view the objective and scope of enquiry, the party (client) interested in the enquiry, the source of data etc. You will study about this in detail later in this unit.

iii) Defining the Statistical Unit: You should define the statistical unit or units in which the data is to be collected. The unit should be appropriate and be free from ambiguity. If the statistical unit is defined clearly, we can avoid the possibility of collecting erroneous data. Once the statistical unit is defined, the same unit should be adopted throughout the investigation. You will study in detail about the statistical unit later in this unit.

iv) Degree of Accuracy: You should also decide the degree of accuracy to be achieved in the collection of data. Absolute accuracy, even if it can be achieved, is seldom desired in statistical investigations. This is because it is expensive and time consuming without much addition to the required standard of accuracy. But, you should attempt to achieve a reasonable level of accuracy, depending on the type of data that are being used, and the purpose of the investigation. We will discuss in detail about the degree of accuracy later in this unit.

4 Collection of Data

After completion of these preliminaries, the next step is the actual collection of data. There are many methods of collecting data and any one of them can be employed. A

suitable method of data collection should be decided after considering various factors such as the nature of the study, objective and scope of enquiry, availability of financial resources, availability of time, etc. Methods of data collection will be discussed in detail later in this unit.

5 Editing the Data

Once the data is collected, the next step involved is the scrutiny of the collected information. This is known as editing of data. It is necessary because in most cases the collected data contains various mistakes and errors. But at the time of editing one should not attempt to tamper with the data.

6 Classification and Tabulation of Data

The mass of collected and edited data is to be organised in the form of tables or charts or graphs or in a compact form called frequency distribution. This would enable us to find out the salient features of the data. Once the data is classified and tabulated, it facilitates easy comparison.

7 Analysis of Data

The next step is the analysis of the data through various statistical measures such as averages, percentages, coefficients, etc. It is not possible to compare a large number of raw figures but comparison is possible when it is presented in the form of a figure which gives overall idea of the data. There are different statistical measures which describe different characteristics of the data in a summary form. You will learn in detail about some of these methods later in this unit. Out of a long list of statistical methods for analysing, you should select only those measures which are suited to the purpose of the survey.

8 Interpretation of Data

After analysing the data, we have to accomplish the task of drawing inferences. This has to be done very carefully. Otherwise there is the danger of drawing misleading conclusions. It is through interpretation we can give broader meaning to survey findings. Relations and processes that underlie survey findings can be focused well by proper interpretation.

9 Writing the Report

The last step of a statistical survey is to write the report. The survey remains incomplete till the written report is presented. The purpose of survey is not well served if the findings are not effectively communicated to people at large. Survey results must invariably enter the general store of knowledge. All this explains the significance of writing the survey report.

Check Your Progress **A**

1 What is the purpose of conducting a statistical survey?

.....

.....

.....

.....

2 List the major steps to be followed in a statistical survey.

.....

.....

.....

.....

.....

3 Can the following steps relating to statistical survey be accomplished by way of preliminaries to the collection of data? State Yes or No.

i) Editing of data .....

ii) Type of enquiry .....

iii) Degree of accuracy .....

iv) Analysis of data .....

v) Tabulation .....

2.3 SOURCES OF STATISTICAL DATA

As you know, after defining the problem, and determining the objective and scope of the enquiry, the next step is to decide the sources from which data is to be collected. You also know that, based on the source, the data may be classified into two categories: (1) primary data and (2) secondary data. Let us now discuss about these two categories of data in detail.

2.3.1 Primary Data and Secondary Data

The data which is collected for the first time for your own use is known as **primary data.** The source happens to be primary if the data is collected for the first time by you as original data. On the other hand, if you are using data which has been collected, classifiedand analysed by someone else, then such data is known as **secondary data.** The sources of secondary data are called **secondary sources.** For instance, national income data collected by the Government in a country is primary data for that Government. But the same data becomes secondary for those research workers who use it later. We may, thus, state that primary data is in the shape of raw materials to which statistical methods are applied for analysis. At the same time secondary data is in the shape of finished products since it has already been treated in some form or the other by statistical methods.

In case you have decided to collect primary data for your survey, you have to identify the sources from which you can collect that data. Big enquiries like population census involve very large number of persons to be surveyed but in case of small enquiries like cost of living of industrial workers in a city, the persons to be surveyed may be few. If you have decided to use secondary data, it is necessary for you to edit and scrutinise such data. Otherwise it may not have the desired level of accuracy or it may not be suitable or adequate for your purpose. If you do not edit and scrutinise the secondary data before you use it in your survey, the results of your investigation may not be fully correct. Therefore, secondary data should always be used with great caution. Bowley writes: **It is never safe to take published statistics at their face value without knowing their meaning and limitations.**

2.3.2 Methods of Collecting Primary Data

There are several methods which one can use for the collection of primary data. The important methods are: (i) observation, (ii) interview, (iii) quertionnaire, and (iv) schedulb. Let us briefly study these methods.

i) **Observation:** In this case you have to collect the information through personal observation and intensive study of the phenomenon when it actually occurs.

ii) **Interview:** The desired information is obtained by interviewing those persons who are supposed to have knowledge about the problem under investigation.

iii) **Questionnaire:** In this method, the information is collected from varioussources by mailing the questionnaire containing a list of questions relating to the problem under investigation. The questionnaire is mailed to the persons concerned and the respondents are requested to answer the questions and return the questionnaire.

iv) **Schedule:** In the case of schedule method, the questionnaires are sent through enumerators. These enumerators help the informants in filling the answers.

To collect the primary data any of these four methods can be used depending on the circumstances, and the availability of persons, funds and time.

2.3.3 Sources of Secondary Data

Secondary data can be collected from two sources: (1)published sources, and (2) unpublished sources. The sources of published data are usually the official publications of the Government, governments of foreign countries, international bodies (e.g. United Nations Organisation, World Bank, etc.), trade associations,

chambers of commerce, banks, stock exchanges, technical and trade journals, books, newspapers and magazines etc. The sources of unpublished data are varied and such material may be found with scholars, research workers, labour bureaus, trade associations, etc.

2.4 TYPES OF ENQUIRIES

While organising a statistical survey, after deciding about the source of data, you have to take a decision about the type of enquiry. There are various types of enquiries sach as census or sample, original or repetitive, direct or indirect, and open or confidential. Before we discuss about these types, let us first explain the factors which affect the decision relating to type of enquiry.

2.4.1 Factors Affecting the Type of Enquiry

The decision regarding the type of enquiry is influenced by a number of factors. They are explained as follows:

1 **Objective** and **Scope of the Survey:** This is one of the factors which determines the type of enquiry. For instance, the objective of your enquiry is to find out the total area under rice cultivation in West Bengal. In this case, the type of enquiry best suited would be one in which there is complete enumeration. If the objective is to find out the yield per hectare in West Bengal, you can take some sample plots in different locations and estimate the yield per hectare. In such case there is no need for complete enumeration. A sample survey may give fairly accurate results. Similarly, if the scope of the enquiry is wide (i.e., information is to be collected from large number of items), you go for one type of enquiry and you go for another type of enquiry if the scope is narrow.

2 **Who Conducts the Survey:** Another factor to be considered while determining the type of enquiry is who conducts the statistical enquiry. The facilities for collection of data differ depending upon whether the survey is conducted by the State or by some organisation or by some individual. The State can spend more money and also can use compulsion to extract information. If the investigation is being conducted by an institution or organisation other than the State, they can use moral pressure and persuade people to give the necessary information. The type of enquiry in such cases is bound to be of a different type. And if the survey is conducted by individuals on their own behalf, the enquiry would be of a still different type because the resources at the disposal of individuals are limited.

3 **Financial Implications:** The decision about the type of enquiry is also affected by its financial implications. As you know money is required to conduct statistical survey. A survey on a large scale requires more money than a survey on a small scale. We all know that the financial resources of different institutions or persons conducting surveys differ. A State can spend much more than a private institution, and a private institution can spend much more than an individual. Therefore, while deciding about the type of enquiry, one has to think about the financial resources involved in it.

4 Sources of Data: One more factor which influences the type of enquiry is the source from which statistical information is obtained. If primary data has to be collected (i.e., the data are to be collected originally), the type of enquiry would differ from the type which would be ideal if secondary data is to be gathered. This is so because in case of primary data we have to define various terms, units, etc., in the light of the objects of the enquiry. But such decisions are not needed while using secondary data.

2.4.2 Different Types of Enquiries

As discussed earlier, there are different types of enquiries. Let us now discuss briefly about each of those methods:

1 Census or Sample Enquiry

You must be knowing that all the items in any field of inquiry constitute a universe or population. In statistics 'population' does not mean only human population. It means sum total of all the items which relate to a certain study. In **census enquiry the whole group is to be surveyed while in a sample enquiry only a part of the group is studied.**

As explained earlier, a complete enumeration of all the items in the population is known as census enquiry. In this enquiry it can be presumed that, when all the items are covered, no element of chance is left and the highest accuracy is obtained. But in reality. this may not be entirely true. There is an error called 'bias' in this type of enquiry which will become larger and larger as the number of observations increase. Moreover, to check this bias there is no other way except through aresurvey or use of sample checks. — You will learn more about bias in Unit 3. Besides, census enquiry involves a great deal of time, money and energy. Therefore, organising census enquiry on large scale becomes difficult because of the resources involved. At times, this type of enquiry is practically beyond the reach of individuals. Perhaps, government alone can get the complete enumeration carried out. Even the government adopts this type of enquiry in very rare cases. For instance, Government of India conducts population census once in a decade. Further, many a time it may not be possible to examine every item in the population. Sometimes it is possible to obtain reasonably accurate results by studying only a part of the total population. In this case, there is no utility of census surveys.

As you know, in case of sample enquiry only a part of the population is studied. When field studies are undertaken, as discussed earlier, considerations of time, cost, convenience, etc., lead to selection of sample survey. The basic assumption in the sample survey is that the sample items selected truly represent the total population. The sample items, therefore, would enable the investigator to estimate the characteristics of the population without any bias and would produce valid and reliable results. The advantages of sample enquiry'are:

i) A sample study is relatively less expensive as compared to a census study and produces results at a relatively faster speed.

ii) It enables more accurate measurements, as it is generally conducted by trained and experienced investigators.

iii) When the population is very large, sample survey is the most suitable method of data collection.

iv) Sample survey method is very suitable, when a test involves the destruction of the item under study. For instance, in physical sciences, you take fresh samples of chemicals every time.

v) It also enables us to estimate errors due to sampling.

In spite of these advantages of sample enquiry, we should remember that if the universe happens to be small, resorting to a sample survey is not useful. In fact, the decision about the type of enquiry (i.e. sample enquiry or census enquiry) depends upon a variety of factors like objective, scope, nature of enquiry, availability of resources, etc.

2 Original or Repetitive Enquiry

An original enquiry is one which is carried out for the first time whereas a repetitive survey is one which is conducted in continuation of previous surveys. In case of a original survey (also known as initial survey), there is freedom for adopting any method of data collection but in case of repetitive enquiry the old method is usually continued. It can only be modified to suit the new situation. However, in repetitive enquiry the definition of the various terms should not be altered, as this would make comparisons inaccurate.

3 Confidential or Open Enquiry

A confidential survey is that where the results of the survey are kept secret and are not made known to the general public, But in case of open enquiry the results are open to the general public. The modes of treatment in open and confidential enquiries will be different. Mast of the enquiries conducted by the State, private institutions and even by individuals are of the non-confidential type. But sometimes private bodies like manufacturers' associations, trade unions, etc., collect information, the details of which are confined only to their members and not to anybody else.

4 Direct or Indirect Enquiry

Direct enquiry is one where data is capable of direct quantitative measurcment. For instance, factors such as height, weight, income, etc., can be measured in quantitative terms. Indirect enquiry is one whdre direct quantitative measurement is not possible. For example, factors such as intelligence, efficiency, honesty, etc, cannot be measured quantitatively. In case of indirect enquiry we have to consider all the factors which have a bearing on the problem under study even though they cannot be quantitatively measured. But, those factors which cannot be quantified directly, should be measured

(quantitatively) indirectly. Forinstapce, to study intelligenceof students, we may study the marks obtained by the concerned group of students.

5 Regular or Ad-hoc Enquiry
A regular enquiry is one in which data is collected at regular intervals over a period of time whereasin an ad-hoc enquiry, data is collectedas and when necessary without any regularity.

6 Official or Semi-officialor Non-official Enquiry
When survey is conducted on behalf of a government, it is an **official enquiry.** When the survey is being done by bodies enjoying government patronage, it is termed as **semi-official** enquiry. The enquiry conducted by private bodies or individuals is known as **non-official or private enquiry.** The facilities available will differ in these three enquiries. In case of official enquiry people may be compelled to supply information. In a semi-officialenquiry people may be requested and the information can be acquired with relative ease. But in a private enquiry the investigator may have to face a lot of difficulty, in spite of his best efforts, in collecting data.

2.5 SAMPLING METHODS

As mentioned earlier, there are two types of surveys: (1) census survey where the whole group is to be surveyed, and (2) sample survey where a selected representative items of the group are studied. In the sample survey, the representative items so selected are referred as **sample.** The technique of selecting items for the sample is usually referred as **sampling method.** There are several sampling methods. They are generally categorised as: (1) probability sampling methods, and (2) non-probability sampling methods.

2.5.1 Probability Sampling Methods

In the case of probability sampling method, each and every item in the population has a probability or chance of being included in the sample. Thus, in this method every member of the population has an equal chance of selection into the sample. Under this probability sampling, there are various methods such as:

1 Simple random sampling
2 Systematic sampling
3 Stratified sampling
4 Cluster sampling
5 Area sampling
6 Multi-stage sampling

1 **Simple Random Sampling:** This method is also known as chance or lottery sampling method. In this case each and every item in the population has an equal chance of inclusion in the sample and each one of the possible samples has the same probability of being selected. This is the most common method used when the population is a homogeneous group. To identify the sample unit, normally, random numbers are used.

2 **Systematic** Sampling: Under this method, population is arranged in alphabetical, serial order etc. Then the sample units appearing at fixed intervals are selected. Thus, you may select every 14th name on a list, every 10th house on the side of a street and so on. Element of randomness is introduced into this method of sampling by using random numbers to pick up the first unit with which to start. Thus, in this method, the selection process starts by picking some random point in the list of population, and the units are to be selected until the desired number is secured.

3 **Stratified** Sampling: This method is generally used when population is not a homogeneous group. Under this method, population is divided into a number of homogeneous sub-populations or strata. While doing this, care should be taken to avoid overlapping. After stratification, the sample items are randomly selected from each stratum either on proportionate or equal basis. To understand this method clearly let us take an example. Suppose we want to survey the economic conditions of the employees of a university and its various affiliated and constituent collegest -

There are different categories of employees: (i) principals, professors, (ii) readers, (iii) lecturers, (iv) administrative staff, and (v) class IV staff. Each of these groups is more or less a homogeneous group. These five groups will, therefore, be called 'strata'. From each of these five groups, you can randomly select a suitable size of sample. This method of selection is called stratified sampling.

4 **Cluster Sampling:** This method involves grouping the population into heterogeneous groups called 'clusters' and then selecting a few of such groups (or the clusters) by simple random sampling method. All the items in the selected clusters are studied for accomplishing the survey work. Let us consider the same example discussed under stratified sampling method. Each of the affiliated and constituent colleges and the different departments of the University have all the five categories of employees: (i) principals, professors, (ii) readers, (iii) lecturers, (iv) administrative staff, and (v) class IV staff. So from the point of economic conditions, employees of an institution form a heterogeneous group. Each institution will therefore be called a 'cluster'. You select a few institutions by a simple random sampling method and then survey all the employees of the selected institutions. This method is called cluster sampling.

5 **Area Sampling:** This method is very close to cluster sampling. It is generally followed when the total geographical area to be covered under the survey is spread very widely. In this sampling method, the geographical area is first divided into a number of smaller areas and then a suitable number of these smaller areas are randomly selected. All units of these selected small areas are then studied and examined for accomplishing the survey work.

6 **Multi-stage Sampling:** This method is suitable for big surveys extending to a considerably large geographical area or the population is heterogeneous. For instance, in a survey you want to select some families from all over the country. Under this multi-stage sampling method, the first stage may be to randomly select a few states. At the next stage, from each sample state you can randomly select a few districts. Then at the third stage you can select a few towns from each of the selected districts. Finally, certain families may be randomly selected within the selected towns. Thus, in this method stratification is done at four stages to constitute a final sample. It may be noted that in this multi-stage sampling, each and every item of the population has a chance of being selected but this chance need not be same for all items.

2.5.2 Non-probability Sampling Methods

This method involves purposive or deliberate selection of particular item(s) of the universe for constituting a sample. This means that if the investigator thinks that certain units are 'not representative'. such units may not get equal chance of being included in the sample. Hence the method is called non-probability sampling. The following methods come under this category:

1 **Convenience Sampling:** When you select the sample items from the population based. on the ease of access, the method is called **convenience sampling.** For example, we want to collect data from the consumers of petrol. We may select a few petrol pump stations within our reach and then may interview the persons who buy petrol at these stations. This would be an example of convenience sample of petrol buyers.

2 **Judgment Sampling:** When investigator's judgment is used for selecting sample items for constituting a representative sample, we call it **judgment sampling.** Judgment sampling is generally used in case of qualitative research surveys where the purpose is to develop hypotheses rather than to generalise larger populations.

3 **Quota Sampling:** This is another variety of non-probability sampling. Under it the population is first divided into homogeneous groups and the interviewers are simply allotted quota to be filled from each group. The actual selection of sample items is left to the interviewers' judgment. The size of the quota for each group is usually proportionate to the size of that group in the population.

As discussed above you find that there are several sampling methods. You can adopt any of these methods whichever is suitable for your enquiry. However, if you resort to random sampling, errors due to personal judgment entering into selection of items can generally be eliminated. In this case sampling error can also be estimated. There are

methods for estimating sampling errors which are outside the scope of this course. Purposive sampling is desirable when the universe is small and a known characteristic of it is to be studied intensively. Sample designs other than random sampling may be used only for reasons like convenience and low costs. Therefore, sampling methods to be used must be decided by taking into consideration the nature and scope of enquiry and other related factors like the time, money, staff, convenience, etc.

2.6 LAW OF STATISTICAL REGULARITY

The law of statistical regularity tells us that the random selection of items from the universe is very likely to give a representative sample. This law, thus, states that **"on an average the sample chosen at random from the universe, will have the same composition and characteristic as the universe."** For instance, if there are 700 boys and 300 girls in a school, a random selection of 100 students would yield about 70 boys and 30 girls. Conversely, it can as well be stated that if a random selection of 100 students from a school reveals 70 boys and 30 girls, it is not unreasonable to conclude that, out of 1,000 students in the school, there will be about 700 boys and 300 girls. In this case, the results obtained from the study of 100 items are applied to 1,000 items and this is precisely the purpose of sampling.

This law of statistical regularity will operate when the following two conditions are fulfilled:

i) The selection of items for the sample should be random. It means that every item in the population/universe should have an equal chance of being included in the sample.
ii) The number of items to be included in the sample should be reasonably large enough so that sample is sufficiently representative.

Thus, if from the universe a moderately large sized sample is chosen at random, it is almost certain that on an average the sample so taken will show the same characteristics as that of the universe.

2.7 LAW OF INERTIA OF LARGE NUMBERS

Law of Inertia of Large Numbers is a corollary to the law of statistical regularity, which we have discussed earlier. There is a relationship between the size of a sample and its accuracy. The reason for this lies in the fact that in large numbers the chances of compensatory errors are greater. In other words, data collected from large samples has a higher degree of stability than the data collected from small samples. For instance, if a coin is tossed 40 times, heads are expected 20 times. But in actual tossing, head may appear 25 times and tail only 15 times. If the coin is tossed further, a reverse situation may arise. If the coin is tossed 1,000 times, it is quite possible there are 500 heads and 500 tails. This is so because when the number of tosses become larger and larger, sometimes errors (difference between actual and expected) move in the opposite direction thus cancelling out each other. In the above example, the larger the number of such tosses, the greater are the chances of one irregularity compensating the other. It is on this basis that we say that large numbers have 'inertia'. In simple words, this means that large numbers are more constant. The production of rice in a given district might show great variations year after year. But the production in the whole state would not vary much, because if in some districts the crop is above normal, it is just possible that in other districts it might be below normal. Thus, the production at the state level would be stable. Similarly, the rice production figures for the whole country would show only a small variation from one year to another. This very phenomenon is referred as the 'Inertia of Large Numbers'.

However, from this discussion we should not infer that the law of inertia of large numbers does not allow any change in figures with the passage of time. All that it means is that there are no violent or significant fluctuations in large numbers. The fluctuations in large numbers are slow and gradual. As the number of items becomes larger and larger, the proportionate deviation from the expected value becomes smaller and smaller.

Check Your Progress B

1 Name a few sources for obtaining secondary data.

.....

.....

.....

.....

.....

2 Mention the methods which are used for collecting primary data.

.....

.....

.....

.....

.....

3 What is simple random sampling? What are its important advantages?

.....

.....

.....

.....

.....

4 State whether the following sampling methods are examples of probability samples or not. State Yes or No.

i) Quota sampling .....

ii) Stratified sampling .....

iii) Area sampling .....

iv) Judgment sampling .....

5 State whether the following statements are True or False.

i) The law of Statistical regularity presumes convenience sampling method of, choosing a sample.

ii) The law of inertia of large numbers states that large numbers are relatively stable.

iii) The law of inertia of large numbers does not allow any change in figures with the passage of time.

iv) Stratified sampling is used when population does not constitute a homogeneous group of items.

2.8 STATISTICAL UNIT

As you know, while planning the statistical survey, it is essential that the unit in which the data is to be collected should be properly defined. Statistical unit may be defined as the unit in terms of which the investigator measures the variable (or counts attributes) selected for enumeration, analysis and interpretation. Proper definition of statistical unit is essential in order to collect relevant data. In the absence of a well defined unit, it is just possible that the data which should have been collected may be omitted and the data which should have been omitted may be collected. The task of defining a unit is not so easy as it may appear to be in the first instance.

2.8.1 Features of a Good Statistical Unit

While deciding the statistical unit for the enquiry, we must pay attention to the following requirements:

i) The unit must be appropriate: The statistical unit must suit the purpose of the enquiry. For instance, you know there are different types of prices such as retail price, wholesale price, cost price etc. When you select the price unit for your enquiry, you should select the price suitable for the enquiry. If the retail price is suitable and you select the wholesale price, you get misleading results.

ii) The unit should be specific and unambiguous: The unit should be defined very specifically and the meaning should not be ambiguous. Otherwise, the data collected may not be fully correct and become inaccurate.

iii) The unit must be stable: If there are fluctuations in its value, the data collected at different times or at different places may not be comparable. At times, the results may mislead.

iv) The unit must be homogeneous: Once the statistical unit is defined, it must be uniform throughout the enquiry so that valid comparisons can be made on the basis of collected data.

v) The unit must be simple: The statistical unit must be simple to understand and complete in itself.

2.8.2 Types of Units

You have learnt the meaning of a statistical unit, and the characteristics of a good statistical unit. Now let us study about the types of statistical units.

1 The statistical unit may be either a physical unit or an arbitrary unit. Units of measurements like ton, kilogram, metre, inch, pound etc., are examples of physical units. Such units are prevalent in common usage and do not need any explanation. In many studies these physical units are not suitable. For instance, you are conducting an enquiry on workers' wages in an industry. In this case the statistical unit to be defined is wage. There are different types of wages such as money wage, real wage, piece wage, monthly wage, and so on. In such a situation, you have to arbitrarily decide which wage you have to collect and give it a proper definition.

2 The statistical units also can be categorised as (i) units of estimation or enumeration, and (ii) units of analysis and interpretation. ,

i) Units of enumeration are those in terms of which the data is collected. Units of enumeration may be either simple units or composite units. A simple unit is one which represents a single condition without qualifications. Examples of such units are worker, house, ton, meter, hour, etc. A composite unit is formed by adding a qualifying word to a simple unit with the result that its scope becomes restricted and its definition becomes relatively difficult. For example, take the two units, 'worker' and 'skilled worker'. Here the first unit is a simple unit and the second unit a composite unit. In the second case we should know not only the meaning of worker but also that of the term 'skilled worker'. Other examples of composite units are machine-hour, passenger-mile, kilowatt-hour, and so on.

ii) Units of analysis and interpretation are those units which are used for comparison and interpretation of statistical data. They include ratios, rates, percentages, coefficients, etc. You will learn more about ratios, rates, percentages, etc. in Unit 4.

2.9 DEGREE OF ACCURACY

As discussed earlier in this unit, while conducting an enquiry we have to decide the degree of accuracy to be achieved in the collection of data. While determining the degree of accuracy we should bear in mind two aspects: (i) the accuracy which is normally possible, and (ii) the degree of accuracy that is considered necessary in that particular investigation. It is very difficult to achieve absolute accuracy i.e., to describe a phenomenon exactly as it is. We may not be able to describe the phenomenon with perfect accuracy either because of the imperfection of the investigator and/or because of the imperfection of the measuring instruments. Hence, it is futile to expect complete accuracy in statistical investigations. Even in physical sciences, where controlled

experiments are performed, absolute accuracy cannot be achieved. Then it is of no use to talk about it in social sciences.

2.9.1 Significance of Reasonable Accuracy

As stated above, there is no need of absolute accuracy in statistical investigations. When reasonably accurate estimates are available, there is no difficulty in understanding or analysing a phenomenon. For instance, when we weight foodgrains in quintals, we do not correct the weight to a gram. It is enough if the weight is corrected to a kilogram. Similarly, the distance between two cities is expressed in kilometers, and a few meters have no significance. Even in counting also absolute accuracy is a rare happening. The population census requires the greatest possible degree of accuracy to count the actual number of people. But even in such a case, it is possible that some persons are left out while carrying out the enumeration. Similarly, accuracy in respect of ages in ordinary use, need not be as great as in the case of population census. It is sufficient for all general purposes if ages are given in completed years only. Thus, there is no need of absolute accuracy, only reasonable accuracy can serve the purpose.

Now the question arises, what is reasonable accuracy? We cannot say anything categorically about this. The reasonable accuracy depends upon the nature and objective of the enquiry and the type of data required. In many cases there are conventional standards of accuracy. In measuring the distance between two cities a few metres can be left out but in the measurement of cloth even a few centimetres cannot be ignored. If we are weighing coal, we may ignore few grams, but we cannot do so while weighing gold. In statistical investigations, we may follow these conventions while deciding the reasonable degree of accuracy.

The investigator should adopt those methods and units which will give him the requisite degree of accuracy. The accuracy of measurement depends upon two factors; (i) the fineness of the measuring instruments, and (ii) the care with which it is being employed by the investigator. For instance, if a ruler is marked up to only centimetres, it is unreasonable to measure lengths correct to millimetres. In the same way, when the ages of the persons are stated in years and months in an enquiry, information down to actual days cannot be obtained therefrom.

2.9.2 Concept of Spurious Accuracy

You have learnt about reasonable accuracy. Now you should also study about spurious accuracy. You can understand the meaning of spurious accuracy by an example. Let the ages of five Xth class students be 16 years 7 months, 17 years 2 months, 16 years 8 months, 15 years 9 months, and 15 years 10 months respectively. From these figures it would be obviously misleading to say that the average age of the students is $(16+17+16+15+15)/5 = 15.8$ years. The highest degree of accuracy that can be attained in this case is to express the average age in years, i.e., as 15 completed years. The degree of accuracy imputed by the figures 15.8 years is called 'spurious accuracy'. In expressing numerical facts it is necessary to guard against such spurious accuracy.

Check Your Progress C

1 Distinguish between complete accuracy and reasonable accuracy.

.....

.....

.....

.....

2 Distinguish between reasonable accuracy and spurious accuracy.

.....

.....

.....

.....

.....

.....

3 **Name** the characteristics of a good statistical unit.

.....

.....

.....

.....

4 Why proper definition of statistical unit is essential?

.....

.....

.....

.....

5 Differentiate between units of enumeration and units of analysis.

.....

.....

.....

.....

6 State whether the following statements are True or False.

i) Rates and percentages are categorised as units of analysis and interpretation.

ii) 'Passenger-mile' is an example of simple unit of enumeration.

iii) Spurious accuracy and reasonable accuracy are inter-changeable terms.

iv) Degree of accuracy to be attained in a survey is always decided before starting the work of data collection.

2.10 LET US SUM UP

Statistical surveys, which are fact finding enquiries, concerning phenomena of interest, are to be properly planned and executed so that their results may depict realities. In organising a statistical survey you have to follow several steps: (1) defining the problem, (2) determining the objective and scope of the survey, (3) accomplishing the preliminaries like deciding the sources of data, type of enquiry, statistical unit and the degree of accuracy desired, (4) data collection, (5) editing the data, (6) classification and tabulation of data, (7) analysis of data, (8) interpretation of data, and (9) writing the report.

Sources of statistical data may either be primary or secondary. If data is collected for the first time by the investigator as original data, such data is called primary data. The sources from which primary data is collected are called primary sources. When already collected data is used, it is secondary for the investigator. Sources of such data are called secondary sources. There are several methods of collecting primary data such as personal observation, questionnaire, interview, schedule, etc. You must decide which method to use depending upon the nature, object and scope of the enquiry along with time and money constraints. There are several sources like books, reports, journals, newspapers, and other published sources from where secondary data can be obtained. They may even be obtained from unpublished sources.

The survey can be of several types. It may either be census survey or sample survey, In the former case the entire group is surveyed, but in the later case only a part of the group is studied. In practice, sample surveys are very popular because of several advantages. Other type of enquiries can be direct or indirect, original or repetitive, open or confidential, regular or ad-hoc, and so on. While deciding about the type of enquiry to be undertaken, you have to keep several factors in mind.

In the case of sample survey, there are various methods for the selection of the sample. Those methods can be broadly categorised as: (1) probability sampling methods, and (2) non-probability sampling methods. Relating to sampling, two laws are important:

(1) law of statistical regularity, and (2) law of inertia of large numbers. The law of statistical regularity states that, if a moderately large sized sample is taken at random from the universe it will, on an average, possess the same characteristic as the universe. The law of inertia of large numbers is a corollary to the law of statistical regularity. It states that large numbers are relatively more stable than small numbers. Fluctuations in large numbers are only slow and gradual.

Statistical unit is one in terms of which you measure the variables or count attributes selected for enumeration, analysis and interpretation. The statistical unit selected should be specific, simple, unambiguous, stable, complete and appropriate. In statistical surveys, generally it is very difficult to attain absolute accuracy. Our purpose is well served by reasonable accuracy which to a large extent depends upon the nature and object of enquiry.

2.11 KEY WORDS

Census Enquiry: A complete enumeration of all items in the population.

Law of Inertia of Large Numbers: Statistical law which states that large groups of data have a higher degree of stability than that possessed by small ones. It simply implies that there are no violent fluctuations in large numbers and they are relatively more stable.

Law of Statistical Regularity: Statistical law which states that a moderately large sized sample chosen at random will show on an average the same characteristics as the universe.

Population: Sum total of all items related to a study.

Primary Data: The original data collected for the first time by the investigator. They are in the shape of raw material, to which statistical methods are applied for analysis.

Random Sampling Method: A sampling technique which gives equal chance to each and every items of the population for being included in the sample.

Reasonable Accuracy: That level of accuracy which is considered necessary depending upon the circumstances of a particular investigation.

Sample Enquiry: The study of only a few representative items selected from the population.

Sampling Methods: The techniques of selecting sample items from the population.

Secondary Data. Those data which were collected by someone else earlier but are now being used by the investigator. They may be in published form or in unpublished form. They are in the shape of finished products since they have already been treated in one form or the other.

Statistical Survey: A fact finding enquiry concerning a phenomenon, spread over a time period in a given area. Quantitative information is collected through the survey on various aspects of the phenomenon under consideration.

Statistical Unit: A unit in terms of which the investigator measures the variables or counts attributes for enumeration, analysis and interpretation.

2.12 ANSWERS TO CHECK YOUR PROGRESS

A.3 i) No ii) Yes iii) Yes iv) No v) No

B.4 i) No ii) Yes iii) Yes iv) No

5 i) False ii) True iii) False iv) True

C.6 i) True ii) False iii) False iv) True.

2.13 TERMINAL QUESTIONS

1 What is a statistical survey? Describe the steps to be followed while organising a statistical survey.

2 What preliminariesshould be accomplished before the data collection work starts? Explain.

3 Differentiate between the primary and secondary data. Explain different methods of collecting primary data and the sources of secondary data.

4 Why sample survey is preferred compared to census survey? Explain.

5 What is sampling? Explain various methods of sampling?

6 Write short note'son the following:

i) Characteristics of a good statistical unit.
ii) Significance of reasonable degree of accuracy in a statistical survey.
iii) Law of Statistical Regularity.
iv) Law of Inertia of Large Numbers.
v) Differentiate between Reasonable Accuracy, Absolute Accuracy and Spurious Accuracy.

Note: These questions will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the university. These are for your practice only.

UNIT 3 ACCURACY, APPROXIMATION AND ERRORS

tructure

.0 OBJECTIVES

er studying this unit, you should be able to:
appreciate the need for accuracy, and distinguish between absolute **accuracy** and spurious accuracy
explain the meaning **and** methods of approximation
describe different kinds of errors in statistics
compute errors by different methods.

.1 INTRODUCTION

ou have already learnt that statistical **data must** have reasonable standard of accuracy nd whenever we conduct a statistical survey, we have to define very clearly the degree f accuracy. So the question is how to measure accuracy and how to make pproximation so that desired level of accuracy can' be achieved? Other aspects to be ept in mind while **conducting statistical** surveys are the errors which creep in at **various** ages. These errors may occur due to inaccurate measurements, **inappropriate** ethods, approximations of figures, the chance factor associated with selection of mple units, etc. In this unit **you will** study more about **the** concept of accuracy. You ill also learn **the concept of approximation,** the methods of approximation, **the** errors . rising from approximation and sampling, and the different methods of measuring rors.

.2 ACCURACY

s you know, statistical data may be obtained by counting or by measuring or by aking estimates. The data on cars **produced** may be obtained by counting the cars. he data on milk powder **produced may** be **recorded** by weighing the milk powder. But hen **government** requires data on production of wheat, before the crop is harvested, tatistician can only estimate the total production. Counting, if done properly, **results** exact numbers. But measurements and estimates on the other hand are not exact. or example, when a truck load of **milk** powder is weighed on a weigh-bridge, a ilogram more or less does not make a difference, Here, the weight is accurate **upto** a ilogram. **But if** a pinch of powder is weighed on chemical balance in the laboratory, ven a **milligram will** tilt the balance. In this case, the weight is accurate **upto** a illigram, Thus, when articles are measured, there is a limit to the **accuracy** depending n the measuring instrument used.

2.13 TERMINAL QUESTIONS

1 What is a statistical survey? Describe the steps to be followed while organising a statistical survey.

2 What preliminaries should be accomplished before the data collection work starts? Explain.

3 Differentiate between the primary and secondary data. Explain different methods of collecting primary data and the sources of secondary data.

4 Why sample survey is preferred compared to census survey? Explain.

5 What is sampling? Explain various methods of sampling?

6 Write short notes on the following:

i) Characteristics of a good statistical unit.
ii) Significance of reasonable degree of accuracy in a statistical survey.
iii) Law of Statistical Regularity.
iv) Law of Inertia of Large Numbers.
v) Differentiate between Reasonable Accuracy, Absolute Accuracy and Spurious Accuracy.

Note: These questions will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the university. These are for y o u practice only.

UNIT 3 ACCURACY, APPROXIMATION AND ERRORS

tructure

.0 OBJECTIVES

fter studying this unit, you should be able to:

- appreciate the need for accuracy, and distinguish between absolute accuracy and spurious accuracy
- explain the meaning **and** methods of **approximation**
- describe different kinds of errors in statistics
- compute errors by different methods.

.1 INTRODUCTION

ou have already learnt that statistical data must have reasonable standard of accuracy nd whenever we conduct a statistical survey, we have to define very clearly the degree f accuracy. So the question is how to measure accuracy and how to make pproximation so that desired level of accuracy can be achieved? Other aspects to be ept in mind while conducting statistical surveys are the errors which creep in at various ages. These errors may occur due to inaccurate measurements, inappropriate ethods, approximations of figures, the chance factor associated with selection of **mple units, etc. In this unit you will study more about the concept** of accuracy. You **ill also learn the concept of approximation, the methods of approximation, the errors rising from approximation and sampling, and the different methods of measuring rors.**

.2 ACCURACY

s you know, statistical data may be obtained by **counting** or by measuring or by taking estimates. The data on cars **produced** may be obtained by counting the cars. he data on **milk** powder **produced may** be **recorded** by weighing the milk powder. But hen government requires data on production of wheat, before the crop is harvested, **atistician** can only estimate the total production. Counting, if done properly, results exact numbers. But measurements and estimates on the **other** hand are not exact. **or** example, when a truck load of milk powder is weighed on a weigh-bridge, a **ilogram** more or less does not make a difference. Here, the weight is accurate **upto** a **ilogram.** But **if** a pinch of powder is weighed on chemical **balance** in the laboratory, **ven a** milligram will tilt the balance. In this case, the weight is accurate **upto** a **illigram. Thus,** when articles are measured, there is a **limit** to the accuracy depending **n the measuring** instrument used,

There are various other factors which also effect the accuracy and lead to errors. You will read about such factors later in this unit when we discuss about the sources of errors. As we have discussed in the previous unit, it is very difficult to attain perfect accuracy. Even in physical sciences like Physics, Chemistry, etc., it is very difficult to achieve complete accuracy. In statistical measurement, as discussed in previous unit, we are content with a reasonable degree of accuracy which is decided by keeping in view its practical value, its use, cost of attaining it, nature and purpose of the survey, etc. In many cases a high degree of accuracy is not even necessary. For examples, it may be more meaningful to say that the population of a country is one million instead of saying more accurately as 1,004,601. One should not, therefore, be particular about absolute accuracy when it is not desirable.

Absolute accuracy may not give the desired clarity. For instance, you are comparing the annual sale of polyester cloth and cotton cloth from a retail shop. It is better to say that they are in the ratio of 3:2 than to state the actual sale figure of Rs. 60,340 in the case of polyester and Rs. 40,105 in the case of cotton cloths.

The extent of accuracy required will also depend upon the situation. A blacksmith weighing iron may not worry about grams, but a goldsmith weighing gold will try to be very accurate to a milligram.

The accuracy is a relative term. Measurements which are accurate for one purpose may be inaccurate for another purpose. For example, in stating the population one may not give the exact number. But when election results are stated, the exact number of votes polled is very important as sometimes the victory margin may be even one vote. Thus, the desired level of accuracy is guided by the purpose of enquiry.

Spurious Accuracy: When the level of accuracy is greater than its real or desired level, it is called spurious accuracy. In statistics claim should not be made for such an accuracy which does not exist. Note that spurious accuracy may be misleading also. In statistical treatment of data, spurious accuracy often results in greater accuracy than is warranted by data. Sometimes the find answer is represented with greater accuracy while the intermediate calculations are done with less accuracy. This also leads to spurious accuracy. A false appearance of extreme accuracy should be avoided.

3.3 APPROXIMATION

You must be aware that in several cases only approximate figures could be arrived at. This is especially so in the case of measurement, as it is difficult to have accurate physical measurement, although one can achieve the reasonable degree of precision. If numerous digits are included in a figure, it may confuse. Through approximation we can exclude the unnecessary digits and avoid any such confusion. Approximation enables clear grasping and facilitates calculations and comparisons. The extent to which approximation should be done depends upon the degree of accuracy desired in the data. The approximation is done by rounding off the digits.

3.3.1 Methods of Approximation

As stated above, approximation is done by rounding off the digits. Now the question is, what is rounding off? The practice of expressing large figures in a simplified form by dropping the last few digits is described as rounding. There are several methods of rounding. Let us discuss about such methods.

i) Rounding Up: If the figures are raised by raising them to the next full unit, it is called rounding up method. For example, the weight of a postal parcel is 8.9 gms. If the weight is to be rounded up, the parcel charges would be levied for 9 gms.

ii) Rounding Down: If the figures are reduced by reducing them to the next lower full unit, it is called rounding down. A common example is that of stating the age as of last birthday. If you are 19 years 10 months old, you would still state your age as of last birthday as 19 years. This is called rounding down.

iii) Stating the Value of the Nearest Unit: The rules for rounding to the nearest full digit are as follows:

a) When the first of the digits to be dropped is less than 5, the preceding digit remains unchanged. For example, the figure 2,23,490 when approximated to the nearest thousand makes it 2,23,000. In this case 490 is dropped as first of the digit to be dropped is 4 which is less than 5.

b) When the first of the digits to be dropped is greater than five, increase the preceding digit by one. For example, the figure 1,42,896 when approximated to the nearest thousand makes it 1,43,000. Here 896 is dropped and as the first of the digit to be dropped (i.e., 8) is more than 5, the previous digit 2 is increased to 3. Similarly if 1,83,503 is rounded to the nearest thousands, it will be 1,84,000.

c) If the first digit to be dropped is an exact 5 (also called neutral), with only zeros to its right, leave the preceding digit unchanged if it is an even number; increase by 1, if it is a odd number. That is to say "round it so that the rounded digit is an even figure". For example, 2,23,500 when approximated to the nearest thousand, it would be 2,24,000. Similarly, a figure of 2,24,500 when approximated to the nearest thousand would also be 2,24,000. In the first case when 500 is dropped the previous digit (i.e., 3) is an odd number. So it is increased to 4. But in the second case,, previous digit is already an even number (i.e., 4), hence on dropping 500 it is unchanged. However, zero is considered as an even digit for this purpose.

iv) **Round to "So Many Significant Figures":** In a simple number or in the process of a computation, the digits that show the extent to which the figure is accurate are called significant digits. A non-zero digit is significant as the term is usually defined. Zeros may or may not be significant. Zeros are significant if there is a significant digit at some place on the right and also a significant digit on the left. In the figure 14,005, the zeros have significance because non-zero digits are there at the left as well as at the right. Zeros at the extreme left of a number are not significant. For example, in a figure like 0,00,500, there is no value for the three zeros at the extreme left. The numbers 501, 0.0501 and 0.000501 each have three significant digits, namely 5,0 & 1. Zeros at the extreme right of a number are significant only occasionally. For example, if 17,000 is correct to unit's place, then all the five digits are significant. If 17,000 is accurate only to the nearest thousand, then there are only two significant digits 1 and 7, and zeros are insignificant.

Zeros on the extreme right i.e., at the right of the decimal point without any non-zero digit after them, indicate the number of places to which the given number is correct. The value 123.00 indicates that it is correct up to two decimal places. Therefore, the two zeros are significant. **Expressing the value to "so many significant places" means presenting figures upto which the given number is accurate.** For example, expressing the number 3.4752 to two significant figures means rounding off the number to first two digits, i.e., write it by dropping 752. The adjusted number, therefore, will be 3.5. Similarly, the number 2,23,490 rounded to four significant figures will be 2,23,500.

Significant figures are the digits that carry real information and are free from inaccuracies. Study the Illustration 1 carefully. It will further clarify the rounding process as well as the concept of significant digits.

Illustration 1

Original Number	Rounded Number	Significant Digits in Rounded Number
5,99,502	600 thousands	600
5,99,500	6 0 0 "	600
5,99,498	599 "	599
5,98,500	598 "	598
999.051	999.1 to one decimal	9991
999.049	999.0 to one decimal	9990
999.150	999.2 to one decimal	9992
999.950	1,000.0 to one decimal	10000
0.00723	0.007 to three decimals	7

Check Your Progress A

1 What is spurious accuracy?

.....

.....

.....

.....

2 List the methods of approximation.

.....

.....

.....

.....

3 What do you understand by significant digits?

.....

.....

.....

.....

4 State whether the following statements are True or False.

i) It is not possible to achieve absolute accuracy in statistical data.

ii) Absolute accuracy carries the same meaning as spurious accuracy.

iii) Rounding up yields a higher result than actual value.

iv) '0' is never a significant digit.

5 Express the following figures after rounding to thousand:

i) 262500 .....

ii) 87634 .....

iii) 96215 .....

iv) 4399510 .....

v) 321501 .....

6 Express the following numbers to two significant digits.

i) 2358 .....

ii) 76435 .....

iii) 8.901 .....

iv) 0.00635 .....

v) 2.031 .....

3.4 ERRORS IN STATISTICS

The word 'error' has a specific meaning in statistics. It means the difference between the true value and the estimated or approximated value of an item. Errors are bound to be there, as the estimates are often based on the sample observations, and themethods involved also include approximation through rounding. Suppose you are interested to estimate the percentage (by weight) of nitrogen in fertiliser. The samples may be taken from different parts of the fertiliser mixture on different days. These samples are sent to different laboratories and analysed by different analysts using different methods. There will be slight variations in the composition of the mixture owing to day-to-day changes in temperature, humidity and other factors. Some of the variations are also due to differentsamples and results in sampling errors. Differences among the analysts may also give rise to some errors. Moreover, there may also be errors of measurement. These will be termed as errors of observations. Thus, in experiments and surveys there are several kinds of errors, namely, (i) sampling errors, (ii) analytical errors, and

(iii) errors of observations and measurements. One must hear in mind that errors are not mistakes arising from the compilation of data. Inaccuracy due to arithmetical miscalculationis not an error but simply a 'mistake'. Since statistics deals with estimated and/or approximated values, the errors are inevitable. These errors cannot be eliminated completely but there are ways to minimise them. However, mistakes can be eliminated completely.

Sources of Errors

,There are three main sources of errors in statistics. They are discussed below:

1 Errors of origin: It is not possible to attain precision while measuring variables such as height, weight, distance, etc. This is due to the limitations of the measuring instruments. Therefore, there is always scope for difference between the measurement and the actual state. Several times the selection of unsuitable statistical units gives incorrect measurement. Another possibility is the informants giving incorrect answers. Biased collection of data i.e., bias of the person collecting the data, also causes errors. The errors so generated are called errors of origin. This type of error tends to increase with the number of observations.

2 Errors of inadequacy: In any enquiry the sample should be representative of the population. If the size of sample is very small and the sample is not correctly representing the population, errors creep in. Such errors are called errors of inadequacy.

3 Errors of manipulation: There may be several errors unconsciously committed by the investigator in measuring, counting and classifying the objects. Such errors together with the errors due to approximation are termed as errors of manipulation. These errors increase with an increase in the number of observations.

In any statistical investigation all these three types of errors prevail.

3.4.1 Errors of Approximation

In statistical reports, figures are usually rounded off for conveniencg. When the figures are rounded, the degree of accuracy (or conversely errors) can be stated in one of the following ways:

1 Expressing data to the nearest thousand or hundred or whole number, For example, 4,672.4 is approximated to the nearest thousand i.e., 5,000 and to the nearest whole number i.e. 4,672.

2 Using the signs + and – to indicate approximation in absolute terms, i.e., 5,000 ± 500. This indicates that the actual value can be 500 more or less than 5,000.

3 Using the signs + and – to indicate proportion of error i.e., 5,000 ± 0.1. This indicates that the actual value can be 0.1 of 5,000 i.e., 500 more or less than 5,000.

4 Using a percentage does much the same as the third case above. For example 5,000 ± 10% means that the error is 10% of 5,000.

5 Expressing the level of approximation of accuracy to the significant figures. For example, 4,672.4 is correct to five significant figures.

The use of symbol ± is a useful way of giving the degree of approximation (or error). The plus and minus signs are used to denote the limits within which the error lies. The limits between which the actual error lies are known as possible errors.

Consider 5,000 ± 500. The maximum possible error is 500, If a certain quantity is rounded off to the nearest thousand, the upper limit of the error will be ± 500 and lower limit will be – 500. So the error will be written as ± 500. If a certain quantity is rounded to the nearest hundred and ten, the possible errors are ± 50 and ± 5 respectively. The maximum possible error can be estimated by looking at the method of rounding.

3.4.2 Measurement of Errors of Approximation

We have discussed the meaning and causes of errors in approximation. Now let us study various methods of measuring these errors.

1. Absolute Error: The difference between a true value and its approximate value (estimated or observed) is called absolute error.

Absolute Error (AE) = $x - x^1$
where x is the true value, and
x^1 is the approximated value.

Absolute error may be positive or negative. For instance, in the case of 5,000 ± 500, the maximum absolute error in either of the directions is 500. If the true value is greater than estimated value, the error is positive and if it is less than the estimated value, the error is negative.

Suppose the population of a state is 2,71,70,314 and of its capital is 26,39,766. When these figures are approximated to the nearest lakh, the population of the state would be 272 lakhs and that of its capital would be 26 lakhs. In the first case approximated value is more than the true value, and in the latter case the approximated value is less than the true value. Now we can calculate the absolute errors (AE) in these cases as follows:

For the state population
A.E. = True Value – Approximated Value
= 2,71,70,314 – 2,72,00,000
= –29,686

For state capital
A.E. = True Value – Approximated Value
= 26,39,766 – 26,00,000
= 39,766.

In these two illustrations, AE is negative in the first case and positive in the second case.

2 Relative Error: The magnitude of the absolute errors of the state population (i.e., –29,686) and its capital (i.e., 39,766) are not very much different, though the state population is ten times more than its capital's. If we want to know which error is more significant, the absolute error is not useful. That is to say, whether an error of 29,686 in 272 lakhs is more significant than an error of 39,766 in 26 lakhs. To find out this, the error should be expressed as a fraction of true value or the approximated value. For this purpose, relative error is more useful. Therefore, the relative error (RE) is defined as the ratio of absolute error to the approximated or estimated value. This can be expressed as follows:

$$\text{Relative Error (RE)} = \frac{\text{Absolute Error (AE)}}{\text{Corresponding Approximated Value } x^1}$$

Now let us take the illustration discussed under the absolute error, and estimate the Relative Error in approximating the state population and capital.

RE in approximating population = – 2,91,686 ÷ 2,72,00,000 = –0.0011
RE in approximating capital = 39,766 ÷ 26,00,000 = 0.0153

The error in approximating the population of the capital is nearly ten times higher. This is because the capital is nearly ten times lower compared to population.

Note that the population of the state can be written as 272 lakhs – 0.0011 and population of the capital as 26 lakhs + 0.0153.

3 **Percentage Error:** When the relative errors (REs) are expressed in percentages, they are called percentage errors. Conversion of a relative error to a percentage error is easy for comprehension.
Percentage Error (PE) = RE × 100
For example, the percentage error (PE) in approximating the state population is -0.0011 × 100 = –0.11% and similarly PE of the capital is 0.0153 × 100 = 1.53%. The percentage error of the capital is about ten times more than that of the state population. Thus, relative error and the percentage error take into account the base of the error. For comparison purposes, therefore, relative error and percentage error are more meaningful than absolute error.

Illustration 2
Find the relative error and percentage error when 2,234.752 is rounded to the
1 nearest two digits after decimal
2 nearest whole number

3 nearest hundred
4 nearest thousand.

Solution:

Method of Rounding	Rounded Value	Maximum Absolute Possible Error	Relative Error (Columns 3-2)	Percentage Error (Col 4 × 100)
(1)	(2)	(3)	(4)	(5)
Nearest two digits after decimal	2,234.75	±0.005	k0.000002 Negligible	50.0002% Negligible
Nearest whole number	2.235	±0.5	±0.0002	50.02%
Nearest hundred	2,200	±50	k0.0227	52.27%
Nearest thousand	2,000	±500	k0.25	±25%

If you study the above illustration carefully, you can notice the following aspects:

1 The maximum absolute error increases as the order of rounding increases, i.e., increasing number of digits are left out.

2 The relative error also increases as the order of rounding increases.

Thus, precision is reduced due to the higher order of rounding.

3.4.3 Computation with Rounded Numbers

While adding and subtracting with rounded figures, it is important to note that the answer cannot be more accurate than the least accurate figures. For example, add the three figures: 1) 357, 2) 574, and 3) 600 where the figure 600 is the least accurate figure as it is rounded to the nearest hundred. The result of adding these three figures is 1,531. But here the answer should be stated as only 1,500 i,e. rounded to nearest hundred. Any attempt to give a higher exactness leads to spurious accuracy.

Similarly, while multiplying or dividing with rounded figures ensure that the answer does not contain more significant figures than the minimum in the rounded figure used in the calculation. For example, mutiplying 2.92 by 2.6 (both rounded) gives 7.592, Here the answer must be in only two significant figures as 2.6 has only two significant figures: Therefore, the answer should be given as 7.6 only. **Thus, while calculations are made with rounded numbers, the final result has only limited accuracy.** Let us study this point in more detail.

3.4.4 Effect of Mathematical Operations on Errors

Errors associated with the approximated figures are affected by operations of addition, subtraction, multiplication and division. Let us study these points one after the other.

Effect of Addition

The absolute error of a sum is equal to the sum of absolute errors of its components. For example, add 500 (to the nearest 10) and 400 (to the nearest 100). This statement can be presented as below:

$(500 \pm 5) + (400 \pm 50) = 900 \pm 55.$

This can be explained in more detail as follows:

Figures	Error	Absolute Error	Maximum Value	Minimum Value
500	Nearest to 10	5	505	495
400	Nearest to 100	50	450	350
Total		55	955 (= 900+55)	845 (=!No−55)

Note that the absolute error is ± 55 (i.e., 5 + 50), the relative error is ±0.061 (±55/900), and the percentage error is ±6.1% (±0.061 × 100.)

Effect of Subtraction

The absolute error of difference equals the sum of errors of its components. For example, from 500 (to the nearest 10) subtract 400 (to the nearest 100). The difference 500 – 400 = 100 will have the error as 5 + 50 = 55. This can be expressed as follows.

$(500 \pm 5) - (400 \pm 50) = 100 \pm 55.$

Let us explain it in detail. Maximum error will occur when the greater figure is at the greatest and lower figure is at the lowest or vice versa. Thus, the absolute error of difference can be calculated as follows:

Figures	Error	Absolute Error	Subtraction will have Maximum Value	Subtraction will have Minimum Value
500	Nearest to 10	5	505 (Max)	495 (Min)
400	Nearest to 100	50	350 (Min)	450 (Max)
100	Total	55	155 (= 100+55)	45 (=loo–55)

The absolute error is ±55 (i.e., 5+50), the relative error is kO.55 (±55/100), and the percentage error is ±55% (k0.55 × 100). If you compare the errors in addition and subtraction, you will notice that the relative error in subtraction is much higher than in addition (because the base is smaller) while the absolute errors are equal. You may note that in both addition and subtraction, absolute error is the sum total of absolute errors of the components.

Effect of Multiplication

The relative error of a product is approximately equal to the sum of the relative errors of its components. Multiply 500 (to the nearest 10) by 40 (to the nearest unit). Now absolute error in 500 is ±5 and the relative error is ±1%. Absolute error in 40 is ±0.5, and the relative errors is ±1.25%. The multiplication of 500 and 40 is 2,000. This can be presented as follows:

$(500 \pm 1\%) \times (40 \pm 1.25\%) = 2{,}000 \pm 2.25\%$

Here relative error ±2.25% is the sum of +1% and ±1.25%. Let us explain it further. The maximum value of the product will be:

$(500 + 5) \times (40 + 0.5) = (500 \times 40) + (500 \times 0.5) + (5 \times 40) + (5 \times 0.5)$ ..... (a)

Similarly, the minimum value of the product will be:

$(500 - 5) \times (40 - 0.5) = (500 \times 40) - (5 \times 40) - (0.5 \times 500) + (5 \times 0.5)$ ..... (b)

Normally, when the errors are small the product of the two errors i.e., the term (5 × 0.5) in (a) and (b), is ignored as it is small. So the absolute error in 500 × 40 i.e. 2,000 is (5 × 40) + (0.5 × 500) which is equal to 450. This means relative error is 450/2,000 = 0.0225 and the percentage error is 2.25%.

Effect of Division

The relative error of a quotient is approximately equal to the sum of the relative errors of its components. To explain this, let us take the example discussed under multiplication and divide it. We have 500 / 40 = 12.5. Taking into account the relative errors, we will have as per statement:

$(500 \pm 1\%) \div (40 \pm 1.25\%) = 12.5 \pm 2.25\%$

This means relative error in the quotient 2.25% is the sum of two relative errors 1% and 1.25%. To check it, we find out the smallest value and the largest value of the division and see how much it is less or more than the division of 500 by 40 i.e. 12.5%.

The smallest value of the division will be obtained when the smallest value of the numerator
(i.e. 500 – 1%) is divided by the largest value of the denominator (i.e., 40 + 1.25%).

Now 500 – 1% = 500 – 5 = 495, and 40 + 1.25% = 40 + 0.5 = 40.5

So we have the smallest value of the division as 495 ÷ 40.5 = 12.22. The difference between 12.50 and 12.22 is 0.28, which is the absolute error. The relative error is 0.28 ÷ 12.5 × 100 = 2.24% or approximately 1% + 1.25% which is the sum of the relative errors in two numbers.

Similarly, we can find out the largest value of the division and see how much it is more than the value 500 ÷ 40 **i.e.**, 12.5. This will be (500 + 5) ÷ (40 − 0.5) − 12.5 = 505 ÷ 39.5 −39.5 − 12.5 = 0.28. This is also the same as the earlier difference. So the **relative** error of a quotient is approximately equal to the sum of the relative errors of its components.

Check Your Progress B

1 What are the sources of statistical errors?

.....

.....

.....

.....

2 What are the methods of expressing errors in approximation?

.....

.....

.....

.....

3 **State** whether the following statements are True or False.

i) Statistical errors means errors in calculation.

ii) Rounding down yields positive errors.

iii) There is no difference in absolute and relative error.

iv) The total absolute error of a product is equal to the product of the absolute **errors** of its components.

v) The relative errors of a quotient equals the sum of the relative errors of its components.

4 A = 25 **thousands,** B = 5 hundreds. Calculate the following and find out the extent of error in the result:

i) A + B .....

ii) A − B .....

iii) AX B .....

iv) A ◄ B .....

3.4.5 Biased and Unbiased Errors

As you know we cannot avoid errors in statistics. Although we accept the inevitability of errors, it is important to know if such errors are biased or unbiased. Let us now discuss about these two types of errors.

Biased **Errors**

When the errors are in one direction, they are called biased **errors. In** the case of these biased errors; the sum of the estimated figures will be either too large or too small than the sum of actual figures.

Suppose in an exercise you have rounded **down all** figures. In this case a biased error results, because **all** the figures after rounding down would be below their true values. For example, 14 is rounded as 10, the figure 132 as 100, and the figure 5,396 as 5,000. Here the errors take place only in one direction. They are **+4,** +32 and **+396. So** the total error in the sum **14+132+5,396 (i.e.,** 5,542) when rounded by the sum of **10+100+5,000 (i.e.,** 5,110) will be the sum of the errors **4+32+396=432,** which is true as 5,5425,110. Such errors are cumulative in nature and, therefore, also known as **cumulative** errors. These biased errors may also creep in because of the bias of

persons or instruments involved in collection of data. For example, the respondents may overstate or understate facts due to personal bias. The meter rod for measuring cloth may be slightly smaller than actual length. Both the cases will give rise to cumulative errors or biased errors. The method of rounding up or rounding down give rise to biased errors. The rounding off to nearest digit does not give rise to biased errors, as in a large number of observations about half the figures may be raised up and the rest may be decreased. Hence the errors in the total tend to cancel out each other.

Unbiased Errors

When the errors tend to cancel each other, they are called **unbiased or compensating errors.** For example, let us find out the total error in rounding of six numbers, 21, 22, 24, 26, 27 and 28 to nearest tens. The first three figures i.e. 21, 22, and 24 would be approximated to 20 each with a total error of +7. The remaining three figures i.e. 26, 27 and 28, would be approximated to 30 each with a total error of –9. The total error in the sum of all these six figures is 7 – 9 = –2 only. Thus, when errors are unbiased, in some cases the approximated value is less than the true value and in other cases the approximated value is more than the true value. Therefore, errors are positive as well as negative and as a result they nullify each other. The net error is negligible. The larger the number of items under review, the smaller will be the unbiased errors. As the number of observations increase, the unbiased errors have a tendency to decrease. Thus, to minimise the unbiased error one of the methods used is to increase the number of observations. Study Illustration 3 carefully.

Illustration 3

Actual Figures	Case (i)		Case (ii)			Case (iii)
	Unbiased Rounding	Unbiased Absolute Error	Lower (000's)	Biased Absolute Error	Upper (000's)	Biased Absolute Error
17,118	17,000	+118	17,000	+118	18,000	–882
8,362	8,000	+362	8,000	+362	9,000	–638
10,509	11,000	–491	10,000	+509	11,000	–491
15,443	15,000	+443	15,000	+443	16,000	–557
Actual absolute error		+432		+1,432		–2,568
Relative error		+0.847%		+2.864%		-4.756%

From lllustration 3 presented above, we can conclude that:

i) The absolute error in the case of unbiased error is lower compared to biased error.
ii) In the case of unbiased error, the relative error is also small. It also decreases with an increase in the number of items.
iii) In the case of biased errors, both the absolute and the relative errors are high. In fact they will increase as the number of items increases.

3.4.6 Estimation of Biased and Unbiased Errors

When figures are approximated, it is necessary to estimate the amount of error involved in that approximation. Sometimes, the exact figures are not known and only approximated figures are given. In such cases the actual amount of error i.e., difference between the actual and the approximate figure, cannot be found out. It can only be estimated. Take Illustration 3, and assume that the actual figures in the first column are not known. The questionis how to estimate the relative error in the total. The estimation procedure would depend on whether the errors are unbiased or biased. Let us now study the methods under both the situations separately.

Estimation when the Error is Unbiased

The absolute unbiased error in an item is in between 0 and 500 when figures are rounded to the nearest thousand. In the Illustration 3, one figure i.e. 17,118 has an error as low as 118 and another figure i.e. 10,509 has as high as 491. In any number rounded to thousand, the possible lowest error can be '0' and possible highest error can be 500. So the **average absolute error (AAE)** in any figure can be taken as (0+500) ÷2 = 250. The best estimate of the unbiased absolute error in the sum of

a number of items is given by the product of this average absolute error and the square root of the number of items. (The proof of this formula is outside the scope of this course.) The formula is as follows:

Absolute Error (unbiased type) = AAE x $\sqrt{N}$
Where, **AAE** is Average Absolute Error
N is Number of Items.

By using this formula, let us now estimate the absolute error in the present example.
Absolute Error : $250 \times \sqrt{4}$
= 500.

Similarly, we can also estimate the relative error with the following formula:
Relative Error = AAE $\times \sqrt{N} \div$ Approximated Total
= 250 $\sqrt{4} \div$ 51,000
= 0.0098 or 0.98%

Estimation when the Error is Biased

An item expressed in thousands can have an error between 0 and 999. So the average absolute error (AAE) in biased errors is $0+999 \div 2 = 499.5$. The formula for estimating the error, when the error is biased, is presented below:

Absolute Error (biased type) = AAE $\times$ N

Where, **AAE** is Average Absolute Error
N is Number of Items

By using this formula, let us now estimate the absolute error in the present example.

Absolute Error = $499.5 \times 4 = 1,998$.

Similarly, we can also estimate relative error with the following formula:

Relative Error = AAE $\times$ N $\div$ Approximated Total

Relative Error (when rounded down) = $499.5 \times 4 \div 50,000 = 0.0399$ or 3.99%

Relative Error (when rounded up) = $499.5 \times 4 \div 54,000 = 0.037$ or 3.7%.

In this example, you should note that the estimated relative and absolute errors are different from the actual relative and absolute errors (refer Illusiration 3) calculated when exact numbers (in column 1) were available. This difference will be quite small if the number of items is larger.

3.4.7 Sampling and Non-sampling Errors

You have learnt the meaning and the method of estimating the biased and unbiased errors. Let us now discuss about sampling and non-sampling errors.

Sampling Errors

The errors caused by drawing inference about the population on the basis of samples are termed as **sampling errors.** The sampling errors result from the bias in the selection of sample units. These errors occur because the study is based on a portion of the population. If the whole population is taken, sampling error can be eliminated. If two or more sample units are taken from a population by random sampling method, their results need not be identical and the results of both of them may be different from the result of the population. This is due to the fact that the selected two sample items will not be identical. Thus, sampling error means precisely the difference between the sample result and that of the population when both the results are obtained by using the same procedure or method of calculation. The exact amount of sampling error will differ from sample to sample. The sampling errors are inevitable even if utmost care is taken in selecting the sample. However, it is possible to minimise the sampling errors by designing the survey appropriately.

Sampling errors are of two types: (i) biased sampling errors, and (ii) unbiased sampling errors. Let us now discuss about them in detail.

1 **Biased Sampling Errors:** A bias may be said to exist when the values of the statistics obtained from the survey have a tendency to deviate only in one direction. Therefore, this type of error does not cancel out. These errors arise due to bias in the selection

of sample units, faulty collection of data, bias in analysis, etc. For example, possibility of biased sampling errors is more when the sample units are selected through deliberate sampling method instead of random sampling method.

Further, in case of difficulties in collecting information from some of the sampling units included in the random selection, the investigator might substitute them by some other units of the population. This also leads to bias if the substitute units are not selected randomly. Sometimes the respondents do not furnish all the information and if the investigator himself supplies the remaining information, this would also lead to bias. In some cases, the information supplied by the respondent itself may be biased, specially when the informant wants to conceal some facts from the investigator. Any consistent error in measurement will give rise to bias. Bias can also creep in as a result of improper data collection instruments and the incompetence of the investigator. Limitations of collection procedure, coding and methods of analysis also give rise to bias. This kind of errors tend to grow as the number of observations increase. Biased sampling errors are cumulative in nature.

2 **Unbiased** Sampling Errors: These errors arise due to chance differences between the units of population included in the sample and those not included. Errors which arise just on account of chance are called unbiased sampling errors. They are not the result of any bias. These errors do not accumulate with an increase in the number of observations. On the other hand these errors have a tendency to get neutralised with the increase in the number of observatibns. Therefore, these errors are also called compensating errors or non-cumulative errors.

Thus, the total sampling errors are made of biased as well as unbiased errors. The main objective of the statistical method in any survey is to devise sampling schemes so that biased errors are eliminated as far as possible and the unbiased errors are reduced to the minimum.

Non-sampling Errors

These non-sampling errors can occur in any survey, whether it be a complete enumeration or sampling. Non-sampling errors include biases as well as mistakes. These are not chance errors.

Most of the factors causing bias in complete enumeration are similar to the one described above under sampling errors. They also include careless definition of population, a vague conception regarding the information sought, inefficient method of interview and so on. Mistakes arise as a result of improper coding, computations and processing. More specifically, non-sampling errors may arise because of one or more of the following reasons:

i) Improper and ambiguous data specifications which are not consistent with the census or survey objectives.

ii) Inappropriate sampling methods, incomplete questionnaire and incorrect way of interviewing.

iii) Personal bias of the investigators or informants.

iv) Lack of trained and qualified investigators.

v) Errors in compilation and tabulation.

This list is not exhaustive, but it indicates some of the main possible reasons.

The sum of sampling errors and non-sampling errors is the total errors. In any survey, the objective is to minimise this total error. Non-sampling errors can be easily controlled by defining population precisely, constructing the questionnaire carefully and pre-testing it, training the investigators, proper checking and monitoring of every step carefully. But this is possible when the number of items is small. Otherwise, it will be very time consuming and costly. But by keeping the sample size small, the sampling errors increase. Hence, while planning a survey you have to be very careful in allocating your limited resources viz., money, time and human resources. The allocation should be done in such that the sampling and non-sampling errors are minimised and thereby maximum level of accuracy is achieved.

Check Your Progress C

1 Distinguish between biased errors and unbiased errors.

.....

.....

.....

.....

.....

Differentiate between sampling and non-sampling errors.

.....

.....

.....

.....

What are the causes of non-sampling errors?

.....

.....

.....

.....

State whether the following statements are True or False.

i) Unbiased errors are cumulative in nature.

ii) Biased **errors creep** in only when the data is approximated by rounding.

iii) **Sampling** errors arise only in case of sample studies,

iv) Non-sampling errors do not arise in case of a sample survey.

v) Two samples drawn randomly from a population may not yield identical results.

Nine figures are added. What is the absolute error in the sum, if they are rounded to (i) **nearest 000's** and (ii) lower **00's.**

.....

.....

.....

.....

15 LET US SUM UP

tatistical data may be obtained by counting or by measuring or by estimating. When **tems** are counted exact numbers can be found out. But in measuring and estimating, bsolute accuracy is not possible. Even in counting, where absolute accuracy is possible, may not be desirable always. It may not give the desired clarity. Extent of accuracy **equired** depends on purpose of the study and the situation. Spurious accuracy, which **nplies** accuracy greater than absolute accuracy or desired accuracy, should be avoided. **Desired** level of accuracy in figures can be achieved by approximation which is done by **ounding** off. There are three methods of rounding: (i) rounding up, (ii) rounding own, and (iii) rounding to nearest unit. The digits which show the extent to which a **gure** is accurate are called significant digits.

tatistical error means the difference **between** the true value **and** the estimated value. n statistical data errors are mainly of three types: (i) errors of origin, **i.e.,** errors due o limitations of the **measuring** instruments, selection of unsuitable statistical units, **nformants** giving incorrect answers, investigators' bias, etc., (ii) errors **of inadequacy** e., size of **sample** too **small** or sample not correctly representing the population, etc., **nd** (iii) **errors of manipulation i.e.,** unintentionally committed errors in counting, **easuring,** etc., or errors due to approximation.

n approximating a given figure by significant digits, there **will** always be some errors. here are three ways of measuring these errors: (i) absolute errors **i.e.,** true value inus estimated value, (ii) **relative** errors **i.e.,** absolute error divided by estimated **alue,** and (iii) percentage errors **i.e.,** relative error × 100.

When computations **are done** by approximate figures, errors go on increasing. The

total absolute error in the sum or difference of two numbers is the sum of absolute errors of individual figures. Similarly, the total relative error in multiplying or dividing twonumbers is the sum of the relative errors in the individual figures. Biased errors are those which occur only in one direction. They are additive in nature and go on increasing with the increase in the number of items.

Sampling errors are thosa which are caused by drawing inferences about the population on the basis of sample. They can be of two types: (i) biased sampling errors, i.e., errors due to faulty units, faulty collection of data, faulty method of analysis, etc., and (ii) unbiased sampling errors i.e., errors which arise due to chance difference or the difference between the population and the sample values. Biased sampling errors can be eliminated. Unbiased sampling errors cannot be eliminated, but they can be minimised. Non-sampling errors are those which can occur in the result of the population studies also. They are mostly of the nature of bias.

3.6 KEY WORDS

Absolute Error: The difference between the true value and.the estimated value.

Biased Errors: When errors are in one direction.

Non-sampling Errors: Those errors which can occur in the population studies also.

Percentage Error: Relative error expressed as a percentage i.e., relative error x 100.

Relative Error: The absolute error subtracted by the estimated value.

Rounding: Expressing large figures in a simplified form by dropping the last few digits.

Rounding Down: Reducing the figure to the next lower full unit.

Rounding Up: Raising the figures to the next full unit.

Sampling Errors: Errors caused by drawing inferences about the population on the basis of a sample.

Significant Digits: Presenting the figures in terms of only those digits which are accurate. These are digits that carry real information.

Spurious Accuracy: Accuracy which is greater than its real or desired accuracy.

Statistical Error: Difference between the true value and the estimated or approximated value of an item.

Unbiased Errors: Errors which tend to cancel each:other.

3.7 ANSWERS TO CHECK YOUR PROGRESS

A 4 i) True ii) False iii) True iv) False

5 i) 262 thousands ii) 88 thousands iii) 96 thousands iv) 4400 thousands v) 322 thousands

6 i) 23 hundreds ii) 76 thousands iii) 8.9 iv) 0.0064 v) 2.0

B 3 i) False ii) True iii) False iv) False v) True

4 i) 25,500 ±500 ii) 24,500 ±550 iii) 1,25,00,000 ± 12% iv) 50 ±12%.

C 4 i) False ii) False iii) True iv) False v) True

5 i) $\pm \sqrt{9} \times 250 = \pm 750$ ii) $\pm 9 \times 49.5 = \pm 445.5$

3.8 TERMINAL QUESTIONS/EXERCISES

Questions

1 What are the different types of errors in statistics? Explain the factors responsible for their occurrence.

2 What standard of accuracy is needed in statistical investigations? State the various methods of approximation and their utility in statistics.

3 Differentiate between biased and unbiased errors. Explain the methods of estimating them.

4 Write short notes on the following:
- i) Errors of approximation
- ii) Distinguish between sampling and non-sampling errors.
- iii) Distinguish between biased and unbiased errors
- iv) Errors of approximation

Exercises

1 Using the rules for rounding to the nearest digit, round the following numbers first to four significant digits, then to three significant digits and finally to two significant digits.
(i) 5.6994 (ii) 47.251 (iii) 3.0009 (iv) 0.0064251 (v) 5.34651

2 The population of a town has been estimated at 49,000 and actual population is 50,000. Calculate (i) absolute error, (ii) relative error, and (iii) percentage error
Answer: i) 1,000 ii) 0.02 iii) 2%

3 The distances of two petrol pumps from the city limit were measured as 225 ± 0.5 kms and 175 ± 0.3 kms. Find out the absolute error (AE) and relative error (RE) in (1) the total distance, and in (2) the difference of two distances.
Answer: (1) AE ± 0.8, RE ± 0.002; (2) ±0.8, RE ± 0.0166

4 There are two measures: A = 375 and B = 25. Each of these two measurements are subject to an error of ± 10%. Determine the range of absolute error and relative error under the following cases:
- i) The sum (A+B)
- ii) The product (A B)
- iii) The difference (A–B)
- iv) The quality A/B

Answer: i) AE ± 40 RE ± 10; ii) AE ± 1969 RE ± 21; iii) AE ± 40 RE ± 11.4; iv) AE ± 3.3 RE ± 22.2

5 The cost of 8 kg of sugar is Rs. 13. If the cost of 1 kg is given as Rs. 1.60 find absolute and percentage errors.
Answer: AE 0.025, PE 1.5%.

6 Find the absolute error and relative error in computing the average speed if total distance was rounded to 1,440 km. and the time was rounded to 70.5 hours.
Answer: AE 0.022, RE 0.0011

7 The sales made by the five regional offices of a company are shown below:

Regions	Units
North	13,434
South	54,682
East	36,482
West	17,804
Central	34,384

Approximate the above figures to the nearest thousand. Estimate the relative error of the total sales assuming that the actual sales values are not known. How the estimated relative error will change if the sale figures are approximated to the next thousand above the actual figure? Compare the estimated relative error with actual relative error.
Answer: 0.004.

Note: These questions and exercises will help you to understand the unit better. Try to write answers for them, But do not send your answers to the university. These are for your practice only.

UNIT 4 RATIOS, PERCENTAGES AND RATES

Structure

4.0 OBJECTIVES

After studying this unit, you should be able to

- explain the meaning of ratios, percentages, and rates
- discuss the computational aspects involved in working out ratios, percentages, and rates
- describe the precautions to be taken while using ratios, percentages, and rates
- illustrate the uses of ratios, percentages and rates in business and administration
- use logarithms in computations such as multiplications, divisions, roots, etc.

4.1 INTRODUCTION

You have already studied about various sources of statistical data and methods of collecting data. You also know about different types of errors which creep into data and methods of estimating such errors. Collection of reasonably accurate data alone does not help in drawing conclusions, as the figures do not speak for themselves. Data must be analysed and compared, so that meaningful and useful conclusions can be drawn. Quantitative data has to be condensed in a meaningful manner, so that it can be easily understood and interpreted. One of the common methods is to compute statistical derivatives. Ratios, percentages, rates, etc., are simple derivatives. These measures point out an existing relationship among factors and thereby help in better interpretation. In this unit you will learn about the meaning, purpose, computation and caution in use of different types of ratios, percentages and rates used in analysing statistical data. You will also learn how to use logarithms in lengthy computations.

4.2 MEANING OF VARIOUS STATISTICAL DERIVATIVES

As stated earlier, to draw meaningful conclusions, the mass of data collected must be analysed properly. Computation of statistical derivatives is one of the common approaches. Under the statistical derivatives, there are three important methods. They are: 1) ratibs, 2) percentages, and 3) rates. Now let us study the meaning of each of these methods.

4.2.1 Ratio

A ratio expresses the relationship between the magnitude of two quantities of the same kind and denotes how many times one of the quantities is contained in the other.

The relationship between the two quantities "A" and "B" is expressed in the form of a ratio as A:B and is read as "A" is to "B". In the ratio A:B, the first term "A" is called the antecedent and the second term "B" is called the consequent of the ratio. The ratio A:B may be expressed as A/B. Thus, the ratio between these two numbers "A" and "B" is simply the concept of "A" being divided by "B". Though the ratio is either an implied division or actual division of one number by another, it is not convenient to represent the ratio as division. For example, if one is comparing the ratio of male workers to female workers in a factory, stating 550/110 would not be easy to comprehend. This can be stated as 550:110 or 5:1 which is easier to understand. Thus, representation in the form of ratio also reduces the size of the number which facilitates easy comparison and quick grasp. When the two terms of a ratio are interchanged, the second ratio so obtained is called the inverse ratio of the first. Thus, the inverse ratio of A:B is B:A. For example, in a class of 80 students, there are 50 boys and 30 girls. Now the ratio of boys to the girls is 5:3 and the ratio of girls to the boys is 3:5. Notice that 5:3 is greater than one and 3:5 is less than one, A ratio can take a value greater than one or less than one or equal to one depending upon the situation. Thus, the ratio is the relation of one number or quantity to another, its value being expressed as the quotient of the first divided by the second.

But, however, the concept of ratio can be extended to three or more numbers. Three or more numbers may be compared and expressed in the form of A:B:C:D, etc. For example, in a class of 100 law students, 70 are from commerce, 20 are from science and 10 are from arts streams. Then the comparison of commerce, science and arts students can be represented as 7:2:1. As the number of categories increases, the proportion is a better derivative for presentation as it will be easy and less confusing. If there are a total of N items divided into three categories — N1 items in category 1, N2 items in category 2, and N3 items in category 3. Then the proportion of category 1, 2, 3, would be N1/N, N2/N, N3/N respectively. In this proportion, the denominator is the total number of items and the numerator is the number of items in the respective category. The proportion will always be less than one and the sum of all the proportions will be one. A ratio can be always converted into proportions.' For example, if the ratio of male to female is 3:2, then the proportion of male is $3/(3+2) = 0.6$, and that of female is $2/(3+2) = 0.4$.

4.2.2 Percentage

Ratios and proportions are often expressed as percentages, as relative measures can be visualised more concretely when expressed in percentages. The word per cent means per hundred. Ratios can be converted into percentages by taking one figure as the base and multiplying it by 100. For example, the yield of paddy crop in 1988 and 1989 is 2:3. To represent this in percentage, the yield of 1988 (i.e. 2) can be considered as the base. Then the yield of 1989 will be $312 \times 100 = 150\%$ of the yield of last year. You should know that the sum total of percentages will be equal to 100, only if all the categories are mutually exclusive and collectively exhaustive (i.e., an item belongs to only one category and no item from the total is left out).

4.2.3 Rate

Generally, when two quantities of the same kind are to be compared, the term ratio is used. For example, take the ratio of male and female workers in a factory. Here, both males and females are workers in the factory. So in this sense they are of the same kind. A ratio expressed in terms of different units is often called a rate, For example, in the case of per capita income, the numerator is the total income and the denominator is the total population. Other examples of rates are death rate, birth rate, accident rate, etc. Rate involves the concept of change. The rates are generally dynamic in nature and related to time. A rate of change is a quotient with the number representing the amount of change in its numerator and the denominator.

A rate is usually standardised in relation to the denominator. When one number is divided by another related number and the quotient is multiplied by 1,000 the resultant figure is known as rate per thousand. For example, if the number of deaths is divided

by the total population and the quotient is multiplied by 1,000, a crude death rate is obtained.

Coefficient: The rate per unit is called coefficient. Suppose the death rate is about 1.9% or 19 per thousand, then the coefficient of death will be 0.019. If this coefficient is multiplied by the total population, the total number of deaths can be obtained.

4.3 PURPOSE OF STATISTICAL DERIVATIVES

You have learnt the meaning of various statistical derivatives, viz., ratios, percentages, and rates. Now the question is: What is the purpose of computing these statistical derivatives? Comparison, as already stated, is the main purpose behind the computation of statistical derivatives. From the meanings of ratio, percentage and coefficient, it is very clear that all of them give a relative picture.

However, when one or more numbers are being compared with another number, the figure which is taken as the standard for comparison is known as the **base.** Which type of base should be chosen would depend upon the situation. Any derivative by itself is generally not meaningful for the analysis of a given problem. For instance, it is stated that a company earned 18% return on its investment during the current year. What does this signify? You may ask whether or not this is a high rate of return. Any meaningful use of derivatives requires comparison with some standard yardstick so that their significance can be evaluated. The return of 18% can be either compared with last year's return or with another competing firm's return on investment, if they are comparable.

While the derivatives are used to compare different groups, it is a common practice to reduce them to a common denominator and thereby the comparisons are made simple and more meaningful. Suppose, two business firms were started with a capital of Rs. 50,000 and Rs. 1,20,000 respectively. At the end of the year, the first business firm made a profit of Rs. 20,000 and the second business firm earned a profit of Rs. 40,000. It apparently shows that the second business has made double the profit of the first business. But by reducing them to a common denominator of 100, it can be seen that the first business has made a profit of 40% of the capital and the second business firm made a profit of 33% of the capital. The impression which you gather by looking at the absolute numbers is reversed now. Thus, profit as a percentage of capital is really more meaningful.

The derivatives are also useful in estimating the unknown quantity. For instance, the birth rate in a particular region is known and it can be assumed to be fairly constant over a period of time. If you know the total number of births, at a specific point of time, you can estimate the population at that point of time. **Thus, the derivatives are useful in the estimation of unknown quantities, over and above simplifying the data and increasing their comparability.**

Check Your Progress A

1 What is a ratio?

.....

.....

.....

.....

2 What is a percentage?

.....

.....

.....

.....

.....

3 What is a rate?

.....

.....

.....

.....

.....

4 Name the different types of ratios used in statistical work.

.....

.....

.....

.....

.....

5 State whether the following statements are True or False.

i) Ratio is always found between two quantities of same type.
ii) Base for computing percentages is always taken as 100.
iii) The ratio cannot be converted into a proportion.
iv) The rate per unit is called coefficient.

6 Match the items in Column A with the items in Column B.

	Column A	Column B
i)	In a business firm, the shares of Ramesh and Ranganathin capital are 2:3	a) Rate
ii)	In India, 20 persons out of 1,000 persons get heart attack	b) Proportion
iii)	In a city, population of males and females is 0.6 : 0.4	c) Percentage
iv)	In India, 70 persons out of hundred live in the villages.	d) Ratio

4.4 TYPES OF RATIOS

There are several types of ratios used in statistical work. The type of ratio used is basically dependent on the base, When one or more numbers are being compared with another number, the figure with which comparisons are made is known as the **base.** Now let us study about different types of ratios.

a) **The Distribution Ratio:** It is defined as the ratio of a part to a total which includes that part also. Suppose in a company there are 300 females out of 1,000 workers. Then the distribution ratio of females to the total is $300 : 1000 = 3 : 10$. This means 30% of the total labour force is females. In this example, 1,000 workers include 300 female workers also. The distribution ratio concept can be extended to more than two groups. This kind of ratio is also called total-to-parts ratio.

b) **Interpart and Interclass Ratio: A** ratio of a part in a total to another part in the same total is called interpart ratio. Here the base is one of the two parts as basically two parts are compared. For instance, sex ratio of a population is an example because sex ratio is usually expressed as number of females per 1,000 males, and it is not expressed as number of females per 1,000 population.

c) **Time Ratio:** This ratio is a measure which expresses the changes in a series of values arranged in a time sequence and is typically shown as percentage, Normally this is also known as **past to present ratio.** There are two main classes of time ratios : (i) those employing a fixed base period, and (ii) those employing a moving base. For instance you are interested in studying the production of tea in the current year. In the fixed base period method you would choose a particular year, say 1980 as the base year and compare the current year's production with the production of 1980. In the **moving base method** the base keeps changing. For calculating current year's tea production, last year's tea production would be assumed as the base. For calculating next year's production, current year production will be used as base. Thus, for

calculations of 1982, figures of 1981 will be used as base; for calculations of year 1983, figures of 1982 will be taken as base and so on. Such calculations give basically comparisons between data corresponding to two consecutive time periods.

d) Hybrid Ratio: A ratio between corresponding parts of different categories of data gives the hybrid ratio. The numerator and the denominator will usually be in different units, For instance, the statement that a car is travelling at "30 miles per hour", involves a hybrid ratio. Here the number of miles is divided by the number of hours and both the units (i.e., miles and hours) are mentioned in the statement of the result. Other common examples of hybrid ratios are per capita income, output per hour or per day, persons per square kilometre, number of children per family, cost per passenger mile, investment per mile, etc. Hybrid ratios are usually stated as per unit base and not as percentages. This is so because the numerator and the denominator are of different categories in this type of ratios. In this sense hybrid ratios can be viewed as rates also.

4.5 COMPUTATION OF RATIOS

You have already studied the meaning and types of ratios. Now let us discuss about the important aspects to be kept in mind while computing ratios. In every statistical ratio one must consider three points : (1) variables to be related, (2) choice of a logical base or denominator, and (3) the choice of units in the denominator. Now we study these three points in detail.

1 Variables to be Related: There must be a definite relationship between the numerator and the denominator. For instance, if you are interested in computing the earning of a company in the current year, the current year's investment must be taken into account and not the investment at the time of its inception. Another example could be the agricultural production per acre. In this ratio, agricultural production per acre of land cultivated is more meaningful than agricultural production per acre to total land (which includes wastelands, forests, deserts, etc.).

2 Choice of Base: The base or denominator of a statistical ratio is always a standard with which the numerator is being compared. As you know, through ratio we establish relationship between two items. Here it is very important to decide which of the two items is to be used as base. In some cases choice of the base is obvious, while in other cases choice of the base is not obvious. However, certain generalisations, can be made in the choice of the base.

i) In a comparison between a part and the whole, the whole is always the base. For example, in relating the number of unemployed to total labour force, the number of persons in the labour force would be the denominator of the ratio.

ii) In time comparisons between similar items (time ratios), the earlier event is taken as the base invariably. For example in comparing the percentage change of current year sales over the previous year, you should consider the previous year's sales as the base.

iii) If the relation is to be studied between two variables, one of which may be dependent upon the other, then the independent variable is generally used as the base of comparison. For instance, in relating the number of accidents to total passenger miles, the later would generally be taken as the base of comparison.

3 Choice of Units in the Denominator: The number of units in the denominator (i.e. base) may be determined by custom, convenience and effectiveness. We can illustrate some of the practices in this regard.

a) There are several cases in which the base of a ratio is expressed as a single unit. For instance, per capita income, per passenger mile, production per acre, etc.

b) Many times ratios are expressed in terms of percentages. For instance, the number of telephone lines in operation today is 150% of the number a year ago. In this case the number stated as a percentage indicates how many numerator units are there for every hundred denominator units. It is easy to visualise treating the base in units of 100.

c) Thousand, ten thousand or even a larger number of units may be used in the base. For example, a statement like 4.5 accidents per 1,000 manhours can also be stated as 45 accidents per 10,000 manhours or 0.0045 accidents per manhour.

Following are some guidelines that help in determining whether one or some higher power of units should be used as the base.

i) The number used as the base should be large enough so that the value of the numerator will appear mainly as a whole number but should not have more than two to three digits to the left of the decimal point. It is more convenient to say that there are 45 accidents per 10,000 manhours than to say that there are 4,500 accidents per 10,00,000 manhours.

ii) The number used as the base should be smaller than the number in the original data corresponding to the denominator. For instance, there are only 12 persons in a firm and nine of them had cars. In this case, it is better to use the original data as the relationship involved is clear without reducing the data into ratio form. If you say that 75% of the employees use cars means the same thing, but it may not give clear impression. Here, the denominator is 100 which is higher than the actual figure i.e., 12.

Thus in computing statistical ratios, one must first decide which variable should go into the numerator, which variable should go into the denominator, and what number of units the denominator of the desired ratio should contain.

4.6 APPLICATION OF RATIOS

Ratios, rates and coefficients are used in all types of studies. Per capita income, population per square kilometre, production per acre, turnover ratio, fixed assets ratio, intelligence quotient, freight revenue per mile, investment per mile, labour to output ratio, capital output ratio, etc., are examples of various popular ratios used. Details of some commonly used ratios are given below. These illustrations are, of course, not exhaustive but do serve to show the importance of ratios as a statistical measure in economics, business and population studies.

1 Per capita income of our country was Rs. 3,184 in 1987-88. This can be obtained by dividing national income of Rs. 2,49,905 crores by the total population of 785 millions.

2 Some ratios for the four metropolitan cities are listed in Table 4.1.

Table 4.1
Profile of Four Metropolitan Cities

		Bombay	Calcutta	Delhi	Madras
1.	Females per 1,000 males	772	781	808	930
2.	Population per sq. km. ('000s)	13.7	10.8	10.6	7.5
3.	Literacy rate (%)	68.2	65.6	62.7	67.4
4.	Working population ratio (%)	34.7	30.5	32.2	28.2
5.	Road accidents per 10,000 motor vehicles	607	279	74	417
6.	Per capita drinking water availability per day (litres)	112	259	239	73
7.	Hospital beds per 1,000 population (Nos.)	3.3	4.1	2.1	3.2

Source: Statistical Outline of India, 1988-89.

Table 4.1 presents some ratios for the four metros. Now you can compare the ratios between the four metros. These ratios are very helpful for the economists and planners to make demographic studies.

3 Among the accounting ratios, one of the well known ratios is the ratio of current assets to current liabilities. If the current assets of a corporation are Rs. 30 lakh and the current liabilities are Rs. 10 lakh, the current ratio is 3. Different standards of

current ratios have been fixed for different types of industries and businesses. These standard current ratios are treated as guidelines for individual enterprises in these industries and businesses.

4 Now-a-days the use of ratios has become very common in population studies. The current population growth rate is worked out at 2.2%. This growth rate of population is determined by mortality rates, fertility rates and the age composition of the population.

In 1985 the infant mortality rate (the number of deaths per 1,000 babies born) was 97 in India. According to 1986 figures the birth rate and death rate per 1,000 population are 33.2 and 12.2 respectively. All these ratios are refined and hence they are called refined ratios. A refined ratio is one in which the numerator or the denominator or both are adjusted so as to exclude the extraneous factors which tend to obscure the direct relationship between them. For instance, ratio of labour cost in a factory to total cost of manufacture is a useful ratio. But the denominator contains two kinds of costs, namely, fixed cost and variable cost. The ratio of labour cost to total variable cost gives a ratio which is more valuable to the management in analysing the operations. A ratio may be standardised by adjusting the component parts of a ratio for better comparability with other ratios. The use of standardised ratios is important in the field of vital statistics where standardised death rates, birth rates, etc., are employed in comparison with different cities or sections of the country. The calculations of standardised rates involve the concept of weighted average and is, therefore, out of scope of this unit.

4.7 CAUTION IN THE USE OF DERIVATIVES

Many of the errors in the use of derivatives spring from failure to express the meaning of derivatives correctly. Difficulties encountered in the computation and use of the derivatives can be generally traced to one or more of the following causes:

1 **Confusion** Regarding the **Base:** Suppose the price of a product has increased from Rs. 2,000 to Rs. 2,500 in the current year. The current price would be 125% of the last year's price. An alternative statement would be that price in the current year is 25% higher than that of last year. Such figures may be misinterpreted to mean either that price in current year is 25% of last year or this year price has increased by 125%. Continuing the same example, suppose the price declined to Rs. 2,000 by the next year. This means the decline is Rs. 500 which is equal to the increase in the last year. Note that in absolute terms the change is the same. i.e. Rs. 500. But, if we express it as a percentage in the first case the increase was 25% and now the decrease is only 20% (i.e., 500 in 2,500). The percentage of change in these two situations is not directly comparable since each percentage is computed from a different base. If an attempt to average them were made, an erroneous conclusion would easily have been reached. For example, the average of 25% and –20% would be +2.5%. This means, on an average, over these two years prices have increased by 2.5% per year. This kind of computation is not correct. Actually, after the two changes the price has returned to the original value. Successive changes of the same absolute amount of 500 produce a larger percentage of increase (i.e., 25%), than decrease of 20% as the increase is computed on the smaller base. A comparison of percentage changes cannot be valid without reference to their bases.

Consider another situation where price of a commodity is Rs. 2,000 which is increased by 20% and then decreased by 20%. After considering this 20% increase and 20% decrease, the value would be Rs. 1,920 which is lower than the original value. Thus, successive increases and decreases of the same per cent will depress the final value below the starting figure.

If any value declines by 100% it results in zero value. Greater than 100% decline cannot occur with quantities like costs, wages, labour, etc., and if it does, it indicates an error. For instance, if the price of Rs. 2,000 is reduced to Rs. 800, the decline of Rs. 1,200 is computed as 150% of the final price. This is an incorrect statement. The base is not correctly chosen. A decline of 150% should ordinarily make the price negative, which is not true here. The percentage decrease must be calculated with the original figure as base i.e., it should he stated as "price has dropped by 60%"

However, there are situations where percentage decline can be more than 100. Consider a firm which has earned a profit of Rs. 1 lakh in one year and sustained a loss of Rs. 50,000 in the next year. Then the total decline in the profit would be Rs. 1,50,000 or 150% of the original profit figure. Here there is nothing wrong because the final figure is negative as it is a loss. The data such as profit can assume a negative value. In such a situation, the percentage decline could be more than 100.

2 Distortions Caused by Small Bases: Consider another example where incorrect conclusions can be drawn due to the distortions caused by small bases. Suppose firm A's profits increased from Rs. 1,000 to Rs. 10,000 and firm B's profits increased from Rs. 50 lakh to Rs. 70 lakh. In the first case the increase is to the tune of 900% and in the second case the increase is only 40%. There is a very high percentage of increase in the case of firm A. Caution is to be exercised in the interpretation of these figures. Obviously a conclusion that the management of firm A is more efficient cannot be justified. Since the percentage indicates a relative magnitude only, no inference should be drawn from this regarding the absolute amounts. In such a situation, a correct picture can be obtained only if the absolute figures are shown.

3 Comparisons Based on Dissimilar Situations: The data should be homogeneous for the computation and the use of ratios and percentages. For instance, the general death rate of a town or a country may be computed by dividing the number of deaths by total population and multiplying by 1,000. But this general death rate, or crude death rate as it is called, relates to heterogeneous masses since the death rate varies with age, sex composition of the population, etc.

Consider another instance. In locality A, 500 out of 1,000 persons, suffered from cholera and in locality B, only 100 out of 1,000 persons are effected. If you compare the ratio of the people suffered from the disease with those not suffered from the disease, in locality A the ratio will be 1:1 and in locality B the ratio will be 1:9. From this ratio one could conclude that locality B is maintained more hygienically and people in that locality are healthier. This may not necessarily be true, as people in locality B might have taken inoculation against cholera.

Consider another statement "the number of crimes committed in the city is 45% more during the current year as compared to last year". From this statement the obvious conclusion is that crimes are increasing. Such a conclusion cannot bc drawn if the ratio compares two dissimilar situations. Suppose, earlier there was no proper reporting system and a new reporting system was introduced at the beginning of the current year. In such a situation one cannot conclude that the crimes have increased. It may be due to the new reporting system as all the crimes are being reported now, which was not done earlier. This means that comparison is made of dissimilar situations. Similarly, for companies having different definitions for costs and profits, comparison of percentage profits and costs should not be made. Before one can draw significant conclusions from the comparison, it is always necessary to find out whether the data analysed is comparable or not.

4 Calculations Based on Small Absolute Numbers: If the size of items is small, percentages may give misleading conclusions. Suppose, from a school only five students appeared for SSC examination and all of them passed. Then the result is stated to be 100%. In this example both the base and the magnitude to which it is compared are small. This kind of statement gives a false impression. Therefore, percentages should not be used if the size of items is very small. Sometimes there may be a situation where the base is too small and the magnitude to be measured against the base is very large. In such a situation one may arrive at a very high percentage figure which is not easy for comparison. For example, assume that price of a commodity is Rs. 10 per unit, Over a period there is a rise of Rs. 400 in this price. This implies prices have increased by 4,000%. This does not simplify but makes comprehension difficult. The statement that the bacteria in drinking water will cause discomfort to 0.0002% of the population, is less precise and less clearer than to say that about 1 person in 500,000 will have discomfort. In this situation the base is very large and the numerator is extremely small. Therefore, the statement of fact is much clearer than converting it into ratio or percentage.

5 Arithmetic Mistakes: Mistakes involving misplaced decimal points may lead to gross misinterpretations. For instance, take the statement "in the current year taxes have

increased by 14% compared to last year". Misplacing the decimal point and stating that "taxes have increased by 0.14% will have a totally different implication". The ratio of the population of two towns A and B was expressed as 8:1 where the population is eight lakhs and 1.5 lakhs respectively. In fact the actual ratio is 5.3 : 1. Thus, arithmetical mistakes in ratios also give an incorrect picture.

6 **Improper Averaging:** Averaging the percentages deserves some discussion as it is incorrectly done in several situations. For instance, take the case of a bolt manufacturing factory where there are three machines A, B, and C to manufacture bolts. Some of the bolts produced by these machines are defective. The defective rate of A, B and C are 3%, 2% and 4% respectively. A simple average of these three percentages would yield a defective rate of 3% (i.e., 3+2+4 ÷ 3). This may not be the real average defective rate, as the number of bolts produced by each machine may not be the same. To find appropriate average it is necessary to know the number of bolts produced by each machine. Suppose machine A produced 300 bolts, machine B produced 100 bolts and machine C produced 600 bolts. Now the appropriate average of defective bolts can be computed as follows:

Machine	Defective Bolts (percentage)	No. of Bolts Produced	Actual No. of Defective Bolts (Cols. 2×3÷100)
(1)	(2)	(3)	(4)
A	3	300	9
B	2	100	2
C	4	600	24
TOTAL		1,000	35

Average rate of defective bolts = (3511000) × 100 = 3.5%. Thus, 3.5% of the bolts are defective. This is substantially different from simple average calculated earlier.

From the above discussion it is evident that calculation of ratio and percentage must be done carefully, so that meaningful conclusions can be drawn. Whenever possible, the data from which these ratios are derived should also be given so that the reader can verify the relationship, and can detect the errors to make his own interpretation.

Check Your Progress B

1 List the guidelines to be followed in selecting denominator for calculating ratios.

.....

.....

.....

.....

2 List the causes which usually give rise to wrong interpretation of statistical derivatives.

.....

.....

.....

.....

3 State whether the following statements are True or False.

i) In time comparisons usually earlier event is taken as the base:

ii) The number used as the base for calculating a ratio has no relation to the size of figures to be compared.

iii) While comparing two figures of percentages it is not necessary to know the actual value of the denominators on which two figures are based.

iv) You cannot draw sound conclusions unless the method of analysing data is suitable to the purpose.

v) Statistical derivations cannot be used to estimate unknown quantities.

vi) The distribution ratio compares one part of the total with another part.

vii) To find time ratios it is not always necessary to use fixed base.

viii) Hybrid ratios are usually stated in percentage form.

4.8 LOGARITHMS

L i e ratios and percentages, logarithms also help us in making relative studies. Logarithms are very helpful, particularly in mathematical calculations. Multiplication, division, and finding roots and powers of large numbers can be done very easily with the help of logarithms.

4.8.1 The Meaning of Logarithms

Consider two numbers 4 and 16. They can be related to each other by the equation $4^2 = 16$. The exponent 2 is the logarithm of 16 to the base 4 and it is written as log, 16 = 2. It should be clear from this example that the logarithm is nothing but the power to which a base (4) must be raised to attain a particular number (16).

In general if $Y = b^t$, then t = log, Y which indicates that log of Y $(\log_b Y)$ is the power to which the base b must be raised in order to attain the value Y.

The following properties of logarithm must be remembered:

1 A negative number and zero cannot possess a logarithm.

2 Logarithm of unity with respect to a non-zero finite number base is zero for $1 = 3^0$, therefore $\log_3 1 = 0$.

3 Logarithm of any number with respect to the same number as base is unit ($\log^3 3 = 1$ as $3^1 = 3$).

The base of the logarithm need not be restricted to any particular number. But in actual applications two numbers are chosen as bases i.e., number 13 and number 'e'. When the base is 10, the logarithm is known as **common logarithm** symbolized by log!, or simply 'log'; With the Napier's constant 'e' (whose value is 2.71828), the loganthm is referred to as a **natural logarithm and is denoted by 'ln'.**

In analytical work, natural logarithms are more useful than common logarithms. In this unit, the emphasis is on the usage of logarithms for computational purposes. Common logarithms are frequently used in computational work. So we now discuss how they arc used in mathematical calculations.

$\log_{10} 1000 = 3$ (because $10^3 = 1000$)

$\log_{10} 100 = 2$ (because $10^2 = 100$)

$\log_{10} 10 = 1$ (because $10^1 = 10$)

$\log_{10} 1 = 0$ (because $10^0 = 1$)

$\log_{10} 0.1 = -1$ (because $10^{-1} = 0.1$)

$\log_{10} 0.01 = -2$ (because $10^{-2} = 0.01$)

From the above, it should be clear that the common logarithm of a number between 10 and 100 must be between 1 and 2, and the common logarithm of a number between 1 and 10 must be a positive fraction. Suppose, a number N is greater than 10 but less than 100 i.e., it has two digits. Then log N is between 1 and 2 i.e., 1 + a, where 1 is the integral part and 'a' is the fractional part. The fractional part is always denoted by a decimal point. If N were more than 100 and less than 1,000 then the logarithm of N will be 2 + b, where 2 is the integral part and 'b' the fractional part. That means that for a three digit number, integral part is 2 in logarithm. Continuing this arguments, when a **number is greater than 1 and contains 'n' digits in its integral part, the integral part of its common logarithm is n–1.**

Now suppose the number N has a value between 1 and 0.1 say 0.8, then its logarithm will be between 0 and –1 i.e., it can be taken as –1 + c where 'c' is a positive fraction. Similarly, if the number N is between 0.1 and 0.01 say 0.03, then its logarithm will be

between –1 and –2 i.e., it can be taken as –2 + d where 'd' is a positive fraction. Continuing further, if the number N is between 0.01 and 0.001 say 0.004 then its logarithm will be between –2 and –3 i.e., it can be taken as –3 + f where 'f' is a positive fraction. So, we notice that here **the integral part of the logarithm is negative and is one more than the number of zeros after the decimal and before the significant digits of the number 'N'.**

Continuing this argument, when a number is less than '1' and has 'x' zeros after the decimal place and before the first significant figure, the integral part of logarithm will be negative and one more than the number of zeros i.e., –(x + 1).

The integral part of a common logarithm is called **characteristic** and the fractional part is called **mantissa.** Note that the characteristics can be zero, positive or negative, but the mantissa is always positive.

4.8.2 Finding the Log Value of a Number

The procedure to find the log value of a number involves three major steps. They are: 1) finding characteristic, 2) finding mantissa, and 3) finding anti-logarithm. Now let us discuss these three steps in detail.

1. **Finding Characteristic:** In the first stage, we have to find out the characteristic. As discussed earlier, if the digits in the number are more than one, the characteristic will be one less than the number of digits to the left of the decimal place. For example, the characteristic of 415.42 is 2, as the number of digits to the left of the decimal place is 3. Similarly, characteristic of 17.23 is 1 and 7.23 is 0.

   In the case of thenumbcrs which are less than one, the characteristicis equal to one more than the number of zeros after the decimal point and before any significant digit. Thus, characteristic of 0.98 is –1, 0.098 is –2, 0.00908 is –3 so on and so forth.

2. **Finding Mantissa:** To find out the mantissa of a number, you have to use logarithm table. Logarithm tables are presented at the end of this unit. For example, you want to find mantissa of the number 3451. First you have to look at the log tables at the row corresponding to 34 (the first two digits of the given number) and the column corresponding to 5 (the third digit of the given number). The mantissa is 5378. Now look at the mean difference column 1 (the fourth digit in the given number) in the same row. The value is 1. Add this 1 to 5378 to obtain 5379. So, for the number 3451, the mantissa part is 0.5379. You already know that the characteristic is 3 for this number. So the log 3451 is 3.5379.

   Note that mantissa is always positive. It is not affected by the position of the decimal point. That is to say, the mantissa of 245, 24.5, 2.45, 0.245, 0.0245, 0.00245, 0.000245 would be the same. Looking at the table, it can be seen that the mantissa value of 245 is 0.3892. The characteristic of a number can be decided upon by looking at the digits in that number itself and themantissa can be obtained from the table using the first four significant digits. Look at the following table and observe how the characteristic is changing without a change in the mantissa value.

Number	Log Value
2450.0	2.3892
245.0	3.3892
24.5	1.3892
2.45	0.3892
0.245	$\bar{1}$.3892
0.0245	$\bar{2}$.3892
0.00245	$\bar{3}$.3892

Note: For some log values, you can find a bar over the characteristic. Putting bar over the characteristic indicates that the part where the bar appeared is negative and mantissa (the decimal part) is positive.

3 **Finding Anti Logarithms:** As you know the logarithm tables give the value of 'mantissa in the logarithms of a number. Whereas the antilog tablesgive the value of the number whose log value is known. Suppose in the above example, log value 3.3892 is known. We are now interested in finding out the corresponding actual number whose log value is 3.3892 i.e., the number 2450. Here, we can say that the

antilog of 3.3892 is 2450. Now let us learn how this antilog value is found from antilog tables.

In order to find the antilog of 3.3892, first consider only the mantissa part i.e., .3892. Look at the antilog tables at the row corresponding to .38 and column corresponding to 9. The number is 2449. Look at the mean difference column at 2in the same row, and the value is 1. By adding 1 to 2449, the digits in the antilog value will be 2450. The next task is to decide the decimal position. In the log value of 3.3892 the characteristic is 3. So according to rules discussed earlier, there should be four digits in the antilog number. Therefore, place a decimal value after four digits. That means, 2450.0 is the original value. To find the number corresponding to log 2.3892, the digits in antilog value obtained from the table will have to be the same as in the earlier case. Only the position of decimal point will change, which will have to be decided by the characteristic. In this case, characteristic is $\bar{2}$. So according to rules given earlier, the antilog must be less than '1' and there must be one zero after the decimal and before the first significant digit in the result. Thus antilog 2.3892 would be 0.0245.

4.8.3 Computation by Logarithms

We have discussed how to find log and antilog values. Now let us study how to use logarithms in different types of computations. Logarithms are normally used for multiplication, division, and finding the power and the root of a number. The calculations are based on a set of rules. Now let us study how it is done.

1 To Multiply Numbers: Find out the logarithms of the numbers to be multiplied, add them together and find out the antilog of the sum.

Thus a x b = Antilog (log a + log b)
This is based on the rule: log a × b = log a + log b.

Illustration 1

Multiply 84.5 by 32.8

Step I Find log of 84.5 and 32.8
log 84.5 = 1.9269
log 32.8 = 1.5159

Step II Add two log values
log 84.5 + log 32.8 = 1.9269 + 1.5159 = 3.4428

Step III Find the antilog of 3.4428
Antilog of 3.4428 = 2772.0
∴ 84.5 x 32.8 = 2772

Note: If we multiply 84.5 × 32.8 directly, we get 2771.6. Logarithm tables give figures only with 4 significant digits. Hence, the result obtained by logarithms is 2771.6 converted to 4 significant digits i.e., 2772.

Illustration 2
Multiply 59.3 by 0.0892
Following the same steps as in Illustration 1
log 59.3 = 1.7731
log 0.0892 = $\bar{2}$.9504

Adding log values, the value will be 0.7235

Note that the value carried forward from adding mantissa is '1'. So when adding characteristic, '2' positives and '2' negatives get cancelled. Therefore, the resultant characteristic of the sum will be '0'. The sign of the characteristic must be taken into account while adding the characteristics.

Now Antilog 0.7235 = 5.290
∴ 59.3 × 0.0892 = 5.290 (4 significant digits only)

2 To Divide: To divide one number by another, find out the log of the numerator and of the denominator. Then subtract the log of denominator from the log of numerator. Find the antilog of the difference, which will be the required answer,

$\frac{a}{b}$ = antilog (log a − log b)

This is based on the rule $\log \frac{a}{b} = \log a - \log b$

Illustration 3

Divide 1465.2 by 18.6

Step I Find log of 1465.2 and 18.6
log 1465.2 = 3.1659
log 18.6 = 1.2695

Step II Subtract log of 18.6 from log of 1465.2
i.e., 3.1659 – 1.2695 = 1.8964.

Step III Find the antilog of 1.8964
antilog 1.8964 = 78.77
∴ 1465.3 ÷ 18.6 = 78.77

Illustration 4

Divide 0.009 by 0.045

Following the same steps in Illustration 3
log (0.009) = $\bar{3}.9542$
log (0.045) = 2.6532
$\bar{3}.9542 - \bar{2}.6532$ = i.301
Antilog (i.301) = 0.2000
∴ 0.009 ÷ 0.045 = 0.2000

3 **To Raise a Number to a Power:** In order to raise a number to a power, multiply the log value of the number by the exponent of the power and find out the antilog. That is a^n = Antilog (n × log a)

This is based on the rule $\log a^n = n \times \log a$.

Illustration 5

Find $(0.4)^3$

Step I Find the log of the base that is 0.4
log 0.4 = $\bar{1}.6021$

Step II Multiply log value by the value to which it is to be raised i.e., $\bar{1}.6021 \times 3$ = 2.8063

Note that mantissa multiplied by 3 will have 1 to carry forward which is positive. Multiplying the characteristic by 3 will result in negative 3, so sum of $-3 + 1 = -2$ or $\bar{2}$.

Step III Find the antilog of the result i.e., $\bar{2}.8063$
,Antilog $\bar{2}.8063$ = 0.0640
∴ $(0.4)^3$ = 0.0640

4 **To Find the Root of a Number:** To get the root of a number, divide the log of the number by the index of the root and find out the antilog. That is

$$\sqrt[n]{a} = a^{1/n} = \text{antilog} \frac{\log a}{n}$$

This is based on the same rule as under (iii), which can also be written as

$$\log a^{1/n} = \frac{1}{n} \times \log a$$

Illustration 6
Find the value of $\sqrt[4]{85.6}$

Step I Find the log value of 85.6
i.e., log 85.6 = 1.9325

Step II Divide the log value of 85.6 by 4 as we are interested in finding the fourth root.
1.9325 ÷ 4 = 0.4831

Step III Find the antilog of 0.4831
antilog 0.4831 = 3.042
∴ $\sqrt[4]{85.6}$ = 3.042

Illustration 7
Find the value of $6\sqrt{0.00856}$

Following the same steps of the above Illustration
$\log 0.00856 = \bar{3}.9325$

To divide $\bar{3}.9325$ by 6, you have to write $\bar{3}.9325$ as $\bar{6} + 3.9325$ because in 3.9325 characteristic $\bar{3}$ which is negative when divided by 6 does not give an integer but gives –0.5. If the division has to be a logarithm of a number, the integral part may be negative but decimal part must be positive.

$$\therefore \quad \frac{\bar{3}.9325}{6} = \frac{\bar{6}+3.9325}{6} = \bar{1}.6554$$

Now Antilog $\bar{1}.6554 = 0.4523$

$\therefore 6\sqrt{0.00856} = 0.4523$

Illustration 8
Simplify $\sqrt{\frac{1.764 \times 89.727}{0.00406 \times 6584}}$

Let us assume the final answer as X.

$$\therefore X = \sqrt{\frac{1.764 \times 89.727}{0.00406 \times 6584}}$$

Taking logarithms of both sides

$$\text{Log } X = \log \left(\frac{1.764 \times 89.727}{0.00406 \times 6584} \right)^{1/2}$$

$$= \tfrac{1}{2} \log \frac{1.764 \times 89.727}{0.00406 \times 6584} \quad (\text{By rule } \log a^n = n \log a)$$

$$= \tfrac{1}{2} \log (\log 1.764 \times 89.727 - \log 0.00406 \times 6584)$$

$$(\text{By rule } \log \frac{a}{b} = \log a - \log b)$$

$$= \tfrac{1}{2} [\log 1.764 + \log 89.727 - (\log 0.00406 + \log 6584)]$$

$$(\text{By rule } \log a \times b = \log a + \log b)$$

$$= \tfrac{1}{2} [0.2465 + 1.9529 - (\bar{3}.6085 + 3.8162)]$$

Note: Logarithm tables can be read upto 4 significant digits only. But 89.727 has 5 digits. So it will be first changed to 4 significant digits i.e., 8973 and then the tables will be consulted. Simplifying the above expression.

Log X = ½ (0.7747) = 0.3874. (Taking only 4 significant digits)
$\therefore$ X = Antilog 0.3874
= 2.440
$\therefore$ The given expression has a value of 2.440.

Illustration 9
Find the value of $\tfrac{1}{2}^{10}$
Let $X = \tfrac{1}{2}^{10}$. Proceeding as in Illustration 8,

$$\begin{aligned} \text{Log } X &= \log \tfrac{1}{2}^{10} \\ &= \log 1 - \log 2^{10} \\ &= \log 1 - 10 \log 2 \\ &= 0 - 10\,(0.3010) \\ &= -3.010 \end{aligned}$$

Note: 1 While consulting the logarithm tables for finding the mantissa for '2', we will have to take it as 2.00 or tables will be consulted for 200.

2 We have log X = –3.010. This means both the integral part '3' and decimal part '.010' are negative. If –3.010 has to be a logarithm of some number X, then decimal part (i.e., mantissa) must be positive. So we rewrite –3.010 by adding and subtracting 1 in such a way that decimal part becomes positive.

Now log X $= -3.010$
$= -3.010 + 1 - 1$
$= -3 - .010 + 1 - 1$
$= -3 - 1 + 1 - .010$
$= -4 + .990$
$= \bar{4}.990$
$\therefore$ X = antilog $\bar{4}.990$
= 0.0009772

While consulting the antilog tables, we looked up for the mantissa 0.900 which has only 3 significant digits. So, the result obtained must also have accuracy upto 3 significant digits only.

Hence $½^{10} = 0.0009772$

Others Uses of Logarithms

The logarithms are very useful in statistical calculations. Logarithms are used in studying the proportionate changes. For example, 10 to 100 has the same degree of relative change as 100 to 1000. In both cases, figures become 10 times. This can be very easily seen from logarithms. The log of 10 is 1 and that of 100 is 2. So the change in logarithms is from 1 to 2. Similarly, change from 100 to 1000 results in change of 2 to 3 in log values. Increase from 1 to 2 is same as increase from 2 to 3. Thus, when changes in logarithms are equal, thisindicates that the relative changes in original numbers are identical. The concepts of logarithms are also used in graphs to show the rate of growth, comparison of series having different magnitudes or unlike units, comparing fluctuations within the same series, checking ratios, percentage changes, etc. The details of the use of logarithms in graphical presentation is out of scope of this any course.

Check Your Progress C

1 Add the following:

i) $\bar{1}.8346$ and 2.6213

ii) 3.1680 and 2.9431

.....

.....

2 Find the value of the following:

i) $2.2163 - 3.8304$

ii) $3.1694 - 2.4560$

iii) $0.3483 - 2.8631$

iv) $0.1294 - \bar{1}.0643$

.....

.....

.....

.....

.....

3 Simplify the following. Write your result in a form so that it can be taken as the logarithm of a certain number.

i) $8 \times \bar{1}.4631$

ii) $\bar{2}.1625/3$

iii) -8.864

iv) $0 - 2.4831$

.....

.....

.....

.....

4 Find the value of the following using logarithms.

i) 3.45 x 138.034

ii) 0.083 × 0.0048

iii) 183.45/3.56

iv) 0.013611.26

v) $(3.14)^5$

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

5 Obtain the value of the following with the help of logarithms.

i) $\frac{4.31 \times 68.453}{20.56 \times 9}$ ii) $\frac{83}{564 \times 35.05}$

.....

.....

.....

.....

.....

.....

.....

.....

4.9 LET US SUM UP

To draw meaningful and useful conclusions, collected data must be analysed in terms of statistical derivatives such as ratios, percentages, rates, etc. A ratio expresses the relationship between the magnitude of two quantities, It is generally stated as A:B or A/B. If number of categories increases, proportion is a better derivative to use. Proportion is the ratio of any one category to the total of all the categories. Ratios and proportions are often expressed as percentages also. A ratio calculated from two different kinds of data (i.e., one in numerator and the other in denominator) is called rate. A rate is usually expressed as per 1,000, per 100, etc. When the rate is expressed per unit, it is called *coefficient*. The main purpose of calculating these derivatives is to facilitate comparisons. They are also useful in estimating the unknown quantities of the data pertaining to either numerator or denominator.

There are several kinds of ratios, viz., Distribution Ratio, Interpart and Interclass Ratio, Time Ratio and Hybrid Ratio. .Time Ratio can be calculated in two ways; 1) by using fixed base or 2) by using moving base. Hybrid Ratio is calculated between corresponding parts of different categories of data. They are usually stated as per unit and not as per cent. In this sense they can be viewed as rates also. In calculating of these

ratios, the denominator or the base is always a standard with which the numerator is being compared. In time comparison earlier event is commonly taken as base. The number of denominator units used as base is determined by common practice, convenience and effectiveness. This number should be large enough so that value of numerator comes out in whole number having not more than two to three digits. It should also be smaller than the size of the original data to be taken in the denominator.

Misinterpretation or misuse of ratios, percentages, etc., is generally due to the following reasons: 1) confusion regarding the base, 2) distortion caused by small bases, 3) comparison of ratios reflecting dissimilar situations, 4) mistakes in computation, 5) calculations based on small absolute numbers, and 6) improper procedure of averaging. One must take care of these problems while computing statistical derivatives and drawing conclusions.

Logarithms are very helpful in mathematical calculations. Logarithms of a number 'Y' to the base 'b' is defined as the power to which base is raised to give rise to the given number. If $Y = b^t$ then $\log_b Y = t$. Logarithms to base 10 are called common logarithms and they are used in calculations. The integral part in the logarithm is called characteristic and the-decimal part is called mantissa. Characteristics can be positive or negative and can be found by just looking at the number of digits to the left of decimal or when the number is less than '1', by counting the number of zeros after the decimal and before the first significant digit. Mantissa is always positive and can be found with the help of log tables. Antilogarithm of a number 't' is equal to 10'. It is determined with the help of antilog tables. Calculations with the help of logarithms are made with the help of three rules.

1 $\log a \times b = \log a + \log b$

2 $\log a/b = \log a - \log b$

3 $\log a^n = n \times \log a$

4.10 KEY WORDS

Characteristic of Logarithm: It is the integral part of a common logarithm.

Common Logarithm: When the base of the logarithm calculation is 10, it is called common logarithm.

Distributive Ratio: Ratio of a part to total which includes that part. It is the same as proportion.

Hybrid Ratio: A ratio between corresponding parts of different categories of data. It is similar to rate.

Logarithm: Logarithm of a given number (logy) is the power to which a given base (say 'b') is raised to attain the given number. If $y = b^t$ then $t = \log_y y$.

Mantissa of a Logarithm: The decimal part in the value of a common logarithm.

Percentage: Gives the magnitude of the numerator when denominator of a ratio becomes 100.

Proportion: The ratio of the number of items in one category to the total number of items in all categories.

Rate: A ratio in which when numerator and denominatorare expressed in different units.

Ratio: Expresses the relationship between the magnitude of two quantities of the same kind and denotes how many times one of the quantities is contained in the other.

Time Ratio: Generally shown in percentages, expresses the change in a series of values relating to different time periods.

4.11 ANSWERS TO CHECK YOUR PROGRESS

A 5 (i) False (ii) True (iii) False (iv) True
6 (i) d (ii) a (iii) b (iv) c

B 3 (i) True (ii) False (iii) False (iv) True (v) False (vi) False (vii) True (viii) False

C 1 (i) 2.4559 (ii) 4.1111
2 (i) 4.3859 (ii) 6.7134 (iii) 3.4852 (iv) 1.0651
3 (i) 5.7048 (ii) 1.3875 (iii) 9.136 (iv) 1.5169
4 (i) 476.2 (ii) 0.0004 (iii) 51.53 (iv) 0.011 (v) 305.2
5 (i) 1.263 (ii) 0.5903

4.12 TERMINAL QUESTIONS/EXERCISES

Questions

1 Explain the meaning and purpose of ratios, percentages and rates.

2 Explain the factors which lead to the misinterpretation of statistical derivatives.

3 What are the various types of ratios? Explain the factors to be kept in mind while computing the ratios.

Exercises

1 Year-wise number of births in a hospital are given. Compute the following:
a) Percentage of males and females every year.
b) Time ratios using i) fixed base approach (1982), and ii) moving base.

Year	Males	Females
1982	1,847	1,754
1983	1,915	1,816
1984	1,823	1,733
1985	1,670	1,588
1986	1,608	1,529

2 One is interested in comparing the sales of three TV brands viz., X, Y, Z in a town during the last month. Four retailers who deal in these three brands are identified and their sales figures are as follows:

Retailer	No. of TV sets sold		
	Brand X	Brand Y	Brand Z
Retailer 1	4	3	1
Retailer 2	8	4	3
Retailer 3	8	6	1
Retailer 4	10	7	0

Calculate the following:
a) The distribution ratio and percentage number of X brand sets to the total number of TVs sold.
b) The distribution ratio and percentage of the number of X and Y brands together to the total sales.
c) The interclass ratio of the number of X TVs to the number of Y TVs sold.
d) The interclass ratio of the number of X plus Y TVs to Z TVs.

Answer: a) 0,545 ; 54.55%
b) 0.909 ; 90.91%
c) 3:2
d) 10:1

3 Mr. A is a new salesman in a business firm and Mr. B is an experienced salesman in the same firm. Mr. A increased his sales in the second month by more than 50% over his first month's sales. In the same period, Mr. B's sales increased by 5%. Does this prove that Mr. A is a better salesman than Mr. B? Explain.

4 What refinement is desirable in the denominator of each of these ratio's:
a) Employees killed in a bus accident to the total number of employees of railways.
b) The number of unemployed in a community to the total number of persons in the community.
c) The number of students passing the B.A. degree to the total number of students in the city.
d) The number of measles cases in the town to the town population.

5 Given the following information about the credit societies in two different states, compute different ratios to enable interpretation of the data. What are your findings?

State	No. of Societies	No. of Members (in 000s)	Loans given dudng the year	
			Number (000s)	Amount (000s)
A	962	2,240	1,340	4,55,800
B	481	546	240	28,900

6 Find the value of the following using logarithms:

i) $100 \times \sqrt{\frac{280 \times 234}{232 \times 194}}$

ii) $32 \div \sqrt{176.5 \times 60}$

iii) $(34.1)^{5.1}$

Answer: (i) 120.7 (ii) 0.9874 (iii) 6,56,60,000

7 Find the value of r when

$$(1 + \frac{r}{100})^5 = 2$$

Answer: 14.9

Note: These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not send them for assessment to the university. They are for your practice only.

SOME USEFUL BOOKS

Elhance, D.N., and veena Elhance, 1988, ***Fundamentals of Statistics,*** Ketab Mahal: Allahabad (Chapters 1, 2 & 7).

Gupta, C.B., ***An Introduction to Statistical, Methods,*** Vikas Publishing House: New Delhi, (Chapter 2, 4).

Gupta, S.P., 1989, ***Elementary Statistical Methods,*** Sultan Chand & Sons: New Delhi. (Chapters 1-4).

Sancheti, D.C., and Kapoor, V.K., 1989, ***Statistics Theory Metods and Applications, Sultan Chand & Sons*** : New Delhi. (Chapter 1 & 2).

Shenoy, G.V., Srivastava V.K., and Sharma, S.C., 1989. ***Business Statistics,*** wiley Eastern: New Delhi, (Chapter 1 & 2).

Simpson, G, and Kafka, F. ***Basic Statistics,*** Oxford & ***IBH*** Publishing: New Delhi (Chapters 1, 2 & 7).

UNIT 5 COLLECTION OF DATA

Structure

5.0 OBJECTIVES

After studying this unit, you should be able to :

- describe the factors affecting choice of data
- explain the problems of collecting primary data
- narrate the different methods of collecting primary data
- state the sources of secondary data
- explain the precautions to be taken while using secondary data, and
- describe the advantages and disadvantages of using secondary data.

5.1 INTRODUCTION

In Unit 2 we have briefly discussed the meaning of primary and secondary data, different methods of collecting primary data and the sources of obtaining secondary data. In this unit we will discuss in detail the various factors determining the choice of data, the problems in collecting primary data, and different methods of collecting primary data. You will also study the merits and limitations of secondary data, sources of secondary data, and the precautions one should observe in using secondary data.

5.2 FACTORS AFFECTING CHOICE OF DATA

As you know, statistical data can be categorised as primary and secondary data. Primary data refers to the data collected for the first time by the investigator as original data. Primary data are generally in the shape of raw material to which statistical methods have to be applied for analysis, If the investigator collects data which has been collected and p r o d by someone else also, such data is referred to as secondary data. Secondary data is generally in the shape of finished product since it has already been treated in some form or the other. When you plan a statistical survey, you have to indicate whether you should collect primary or secondary data. This is a basic question which must be settled in advance before you start data collection. The choice of data depends on several factors which are stated below :

1) Object of the enquiry
2) Scope of the enquiry
3) Financial resources

4) Time factor
5) Status of the investigating agency
6) Human resources
7) Availability of secondary data
8) Degree of accuracy desired

1) Object of the Enquiry : This is the most important factor affecting the choice of data. You should collect the data which can serve the object of the enquiry. The object will indicate the type of information required for the survey. If the objective of the study is served by the primary data, you have to go for primary data.

2) Scope of the Enquiry : Scope refers to the coverage of the survey with regard to the type of information, the subject matter, geographical area covered, etc. If the information to be obtained happens to be quite comprehensive and basic, the primary data would prove to be more suitable.

3) Financial Resources : Finance, in fact, is a big constraint in statistical survey and you have to act within this limitation. The availability of funds determines, to a large extent, the type of data to be used in an investigation. In most of the cases, collection of primary data requires more funds than the secondary data. When funds are very limited, we usually go for secondary data. If adequate finance is available, then we can plan for collecting primary data.

4) Time Factor : The required time for the collection of primary data and secondary data should also be considered in deciding whether to select primary data or to use secondary data. Collection of primary data requires relatively more time than the collection of secondary data. If there is sufficient time for the accomplishment of the investigation, we may use primary data. But if the time is a constraint, then we should consider secondary data. The time factor. thus, affects the type of data to be collected for the survey.

5) Status of the Investigating Agency : This is another important factor in making choice of data to be collected. Much depends upon whether the investigating agency is the Government or some public organisation/institution or an individual. In the first case one can think of collecting primary data on a large scale. However, for individuals, it is very difficult to collect primary data on a large scale. For an individual it is economical and practically feasible to use secondary data. Public organisations or institutions may also take-up field surveys for obtaining relevant information but that may not be the case with private organisations. The Government or the public institutions can afford to spend more money and employ adequate number of trained and competent staff for the collection of primary data. But individuals or private organisations have lots of constraints in this regard.

6) Human Resources : Availability of human resources also affects our choice concerning the data. As you know, for collecting primary data you require more persons. If you have competent and well trained staff you can easily organise field surveys and may collect primary data. If you do not have enough human resources, you can plan to use secondary data for your investigation.

7) Availability of Secondary Data : Secondary data can be made use of only if they are available. If the secondary data are not available or if they are not adequate or not suitable, there is no alternative except to collect primary data.

8) Degree of Accuracy Desired : The choice of data also depends upon the degree of accuracy desired. Before making the choice of data, we have to decide on the degree of accuracy desired. If the secondary data attain the same degree of accuracy as desired in the present investigation, you can use secondary data. Otherwise it is advisable to plan for collecting primary data that fulfil the requirement of the desired level of accuracy.

Among all the factors discussed above, no single factor can form the basis of our choice for data. Keeping all the factors in view, a decision has to be taken whether to use primary or secondary data.

5.3 PROBLEMS IN COLLECTING PRIMARY DATA

We have discussed various factors which influence the decision relating to the choice of data. Now, let us discuss the problems involved in the collection of primary data. In fact, the problems in collecting primary data are different from the problems we face when we collect

secondary data. When you go for primary data, you generally confront the following problems:

1) When you decide to collect primary data, you have to plan the field work quite comprehensively. This is necessary because the quality of result of an enquiry depends, to a large extent, on the preparations made before starting the data collection. Various steps in planning the enquiry were already discussed in Unit 2.

2) To collect the primary data, the unit in terms of which the data have to be gathered must be very clearly and unambiguously stated. The unit must possess the entire characteristics of a good statistical unit. Properly defined unit is a prerequisite for smooth collection of primary data.

3) The problem concerning the technique of data collection is also to be looked into. As you know, there are two techniques of data collection : (i) Census method, and (ii) sample method. The former method requires the collection of information from all the units in the population whereas the latter method obtains information from only a part of the universe. The investigator must decide which technique he will use. The choice would depend upon the availability of resources, the time factor, nature and scope of enquiry and similar other factors. In case of sampling technique, the sampling design has to be carefully specified.

4) Developing the frame is another problem which has to be set prior to the collection of data. 'Frame' refers to a list, map or other specification of the units which constitute the available information relating to the population designated for a particular enquiry. Data can be collected easily if a suitable frame already exists. Otherwise a suitable frame has to be developed before embarking on the collection of data.

5) Before collecting primary data it is necessary to decide on the degree of accuracy desired. As you know, a reasonable level of accuracy is desired in all statistical enquiries. The desired level of accuracy is to be determined keeping in view the object and purpose of enquiry.

6) Designing the forms for the collection of data is another problem in the context of collecting primary data. Careful attention should be given to the designing of various forms (viz., questionnaire, schedule, etc.) that will be used. They should be designed in such a way that required information can be collected through them. Before the commencement of data collection, the forms should be pre-tested,in order to examine their effectiveness. If someshortcomings are identified in the pre-testing, the same should be eradicated before finalising the instrument i.e., the form.

7) The selection, training and supervision of the field staff is more important for the collection of primary data than is necessary in the case of secondary data. Since the success of survey depends upon the field staff, it is essential that they are properly selected, thoroughly trained and their work closely supervised. The enumerators selected should be honest, intelligent and hard working. They should be able to elicit the needed information from the respondents.

8) Then comes the problem of exercising control over the quality of field work. Occasional field checks should be made to ensure that the interviewers are doing their assigned job sincerely and efficiently. A careful watch should be kept for unanticipated factors in orders to keep the data collection work as much realistic as possible. In other words, steps should be taken to ensure that the data collection work. is under control so that the information . collected is in accordance with the pre-defined standard of accuracy. The collected data should be checked for omissions, inconsistencies, and other errors before they are passed on for further processing.

9) In spite of the best effart, the problem of non-response may remain. Some suitable method(s) should be designed to tackle this problem. One method of tackling the problem of non-response is to make a list of non-respondents and take a small sample of them and efforts can be made for securing response with the help of experts. But in any case enumerators should not be allowed to substitute for anyone who is not considered to be a good respondent. Otherwise the bias would creep in the collected information. Proper organisation has to be set up so that the data collection work proceeds smoothly. This problem is more serious particularly in big enquiries or investigations. You must judiciously select the method(s) of data collection (like observation, questionnaire, schedule, interview, etc.) which may prove appropriate for your study. Of course, while making a choice of the method to be used, you should pay attention to the nature and object of enquiry, availability of funds, the time factor and precision required,

Check Your Progress A

1) Differentiate between primary and secondary data.

.....

.....

.....

.....

2) List the factors that affect the choice of data to be used in an investigation.

.....

.....

.....

.....

3) Mention five problems with which the investigator is generally contronted while collecting primary data.

.....

.....

.....

.....

4) State whether the following statements are True or False.

i) It takes more time to collect secondary data compared to primary data.

ii) National income data collected by the Government is a secondary data in the hands of a researcher who uses it for his study.

iii) Any single factor, considered in isolation, should not form the basis of our choice of data.

iv) There is no difference in the problems confronted by the investigator whether he collects primary data for himself or uses secondary data.

v) Secondary data should be used after it has been ensured that it is reliable, adequate and suitable.

vi) The problem of developing a suitable frame arises while using secondary data.

vii) The problem of selection, training and supervision of the field staff arises whenever we collect primary data irrespective of the method we use for ,the purpose.

viii) Problem of non-response is a major problem in case of secondary data.

5.4 METHODS OF COLLECTING PRIMARY DATA

There are many methods for the collection of primary data and any one of them can be employed depending upon the nature of the survey. In Unit 2 you studied in brief about these methods. Now let us study about them in more detail.

5.4.1 Observation

Observation is a systematic viewing, coupled with consideration of the seen phenomena. The Concise Oxford Dictionary defines observation as accurate watching and noting of phenomena as they occur in nature with regard to cause and effector mutual relations. The required information is obtained directly through observation rather than through the reports of others. In the case of behaviour one finds out what the individual does, rather than what the individual says he does. If the informants are unable to provide the information or can give only very in exact answer, questioning is not useful and observation is the only way to proceed. For instance, when you are studying the behaviour of small children who cannot speak, you can collect the information by abserving the children under different circumstances. You

should remember that all phenomena are not open to observation. Even if a phenomenon is open to observation, it may not find a ready observer at hand.

Observation may be participant observation or non-participant observation. In Participant Observation Method the observer joins in the daily life of the group or organisation he is studying. He watches what happens to the members of the community and how they behave. He also engages in conversation with them to find out their reactions to, and interpretations of, the events that have occurred. In the **Non-Participant** Observation Method in order to collect information the observer will not join the group or organisation he is studying but will watch it from outside.

Merits : This method has the following advantages :

1) This is the best suitable method when the informants are unable to provide information or can give in exact information.
2) This method provides first hand information and provides deeper insights into the problem. Therefore, this is useful for intensive studies.

Limitations : This method suffers from the following limitations :

1) In many cases, you cannot predict when the events occur. So a phenomenon which is open to observation may not find a ready observer at hand.
2) The observer should be very objective in interpreting the events he has observed. Otherwise, the bias of the observer may creep into the results.
3) This is not suitable for large scale extensive studies.
4) The presence of the observer may influence the behaviour he is observing. In such cases he may not get the actual information.

5.4.2 Personal Interviewing

Under this method data are collected by the investigator himself through interviews. Therefore, the enquiry is intensive rather than extensive. Under this method the investigator meets the informants personally, asks them questions pertaining to enquiry and collects the desired information. Thus, if a person wants to collect data on the wages of workers of the National Ball Bearing Company, he would go to the factory site of this company, contact the workers and collect the relevant information. Thus, this method is generally used in small size surveys confined to a small locality.

Interviews can be formal or informal. In **Formal Interviewing,** set questions are asked and the answers are recorded in astandardisedform. This is the practice in large scale interviews where a number of investigators are assigned to the job of interviewing. In a formal interview, the interviewer's bias is minimised. This type of interview is most suitable when you know very clearly what type of information you require for your survey. In the case of Informal Interviewing, the investigator may not have a set of questions but have only a number of key points around which to build the interview. The interviewer is at liberty to vary the sequence of questions, to explain their meaning, to add additional ones and even to change the wording. Informal interviews are preferred in the case of an explorative survey where you are not sure about the type of data you collect.

Merits : The major advantages of this method are as follows :

1) The response of the persons interviewed is more encouraging as most people are willing to supply information when approached personally.
2) The information obtained is likely to be more accurate because the doubt of any of the informants can be cleared by the investigator himself.
3) Additional information about the personal characteristics of informants which are helpful in interpreting the results later on may as well be collected.

Limitations : The limitations of this method are as follows :

1) Major limitations of this method are the subjective factors or the biases of the investigator coming in either consciously or unconsciously.
2) It is a costly and time consuming method especially when the number of persons to be interviewed is large and they are spread over a wide area. So, this is not suitable for big surveys.

5.4.3 'Through Local Reports and Correspondents

Under this method the investigator appoints local agents or correspondents in different places of the field of enquiry and the relevant information is obtained through them. These correspondents collect and transmit the information to the office of the investigator. Newspaper agencies generally adopt this method. This method is also used by various departments of the Government in cases where regular information is to be collected from a relatively wide area. In case of making crop estimates or for obtaining regular information regarding prices of different commodities for the preparation of price index this method is generally used.

Merits : The chief merit of this method is that it is comparatively cheap and also gives approximately good results. The method is equally appropriate for extensive enquiries.

Limitations : The danger of entering personal bias of the correspondents in the reports submitted by them at more or less regular intervals is, however, great in this method. Thus, it is necessary that the correspondents or the agents appointed for the purpose must be selected very carefully and trained properly.

5.4.4 Questionnaire

Collection of data through questionnaires is the most popular method for collecting primary data. A questionnaire is a list of questions pertaining to the enquiry. Under this method a questionnaire is sent to various informants with a request to answer the questions and return the questionnaire. The questionnaire is mailed to the respondents who are expected to read the questions and record their response in the space meant for the purpose on the questionnaire itself. The respondents have to answer the questions on their own. This method is extensively employed in various economic and business surveys.

Merits : The merits of this method are as under :

1) This method is very economical particularly when the universe is large and spread geographically on a vast area.
2) Since the answers happen to be in the respondent's own words, he/she is free from the bias of the interviewer.
3) Respondents can take their own time to answer the questions. So they give well thought out answers.
4) Respondents that are at remote places and are not easily approachable can also be reached conveniently.
5) Large samples can be covered and thus the results can be more dependable and reliable.

Limitations : This method also suffers from the following limitations :

1) Sometimes the respondents do not bother to return the questionnaires. So there is the problem of low rate of return of the duly filled in questionnaires. And also bias due to non-response cannot often be determined.
2) Questionnaires can be circulated only among the respondents who are educated and cooperative.
3) Once the questionnaires are sent to the respondents, the investigator cannot change or modify the questions for individual respondents.
4) There is no flexibility because of the difficulty of amending the approach once the questionnaires have been despatched.
5) There is also the possibility of ambiguous replies or omission of replies to certain questions. Interpretation of omissions is difficult.
6) It is difficult to know whether willing respondents are truly representative.
7) This method is likely to be the slowest of all, because the respondents take their own time to return the filled in questionnaires.

Before sending them to the respondents, it is advisable to conduct a 'Pilot Survey' for pre-testing it. Pilot Survey is in fact the replica and rehearsal of the main survey. From the experience gained in this sort of survey, changes can be made in the questionnaire for the final collection of data. The pre-testing is necessary particularly in case of a big enquiry.

Features of a Good Questionnaire : In order to make the questionnaire more effective, it must be very carefully drafted. The' form and tone of the questionnaire must be designed so as to bring in the personal element which is lost in the mailed questionnaire. The following are the qualities of a good questionnaire :

1) It should be short and simple.

2) Questions should proceed in logical sequence starting with easy questions and then moving on to more difficult ones. Personal questions should generally be avoided or may be left to the end.

3) Questions may be dichotomous (yes or no type), or multiple choice. Open ended questions are difficult to analyse and should be avoided to the extent possible.

4) In order to ensure the reliability of respondent there should be some control questions. They introduce a cross-check to see whether the information collected is correct or not.

5) Adequate space for answers should be provided in the questionnaire itself. There should always be provision for indications of uncertainty e.g., "do not know", "no preference". and so on.

6) Layout and design of the questionnaire should also be attractive so that it may attract the attention of the respondents.

5.4.5 Schedule

This method of data collection is similar to that of the questionnaire. The schedule is also a proforma containing a set of questions. The'difference between the questionnaire and the schedule is that the schedule is being filled in by the enumerators who are specially appointed for the purpose. These enumerators go to respondents with the schedules and ask them the questions from the schedule in the order they are listed. The enumerator records the replies in the space meant for the same in the schedule itself. In certain situations, schedules are handed over to respondents and the enumerators help the respondents in recording the answers. Enumerators explain the objectives of the investigation and also remove the difficulties which the respondent may feel in understanding the implications of a particular question(s) or the definition or concept of difficult terms. Thus, the essential difference between the questionnaire and schedule is that the former (i.e., questionnaire) is sent to the informants by post and in the latter case the enumerators carry the schedule personally to informants and fill them in their own handwriting. This method is usually adopted in investigations conducted by governmental agencies or by some big organisations. For instance, population census all over the world is conducted through this method.

Data collection through schedules requires enumerators for filling up schedules and as such they should be very carefully selected. They should be trained to perform their job well. They should be intelligent and must possess the capacity of cross-examination in order to find out the fact. Above all, they should be honest, sincere, hard working and should have the patience and perseverance. **In** drafting the schedules, **all** points stated for a good questionnaire, **must** as well **be** observed.

Merits : The main advantages of this method are as follows :

1) It can be adopted in those cases where informants are illiterate.

2) The problem of non-response is avoided as the enumerators go personally to obtain the information.

3) The method is very useful in extensive enquiries and can lead to fairly reliable results.

4) The identity of the respondent is known which is not always clear in case of a questionnaire.

Limitations : This method has the following limitations :

1) This method is very expensive as enumerators are generally paid persons. Money also has to be spent in training them.

2) Another limitation is that if the investigator is not good in interviewing, most of the information collected by him may be unreliable.

3) Since the investigator is present when the respondent is giving the answers, the respondent may not give answers to some personal questions freely.

5.4.6 Choice of Method

As discussed above, there are several methods for the collection of primary data. You have to choose the best suitable method for your study. You must make your choice of method very judiciously so that the method chosen will be quite appropriate and effective. For this purpose, you should consider (i) nature, scope and object of enquiry, (ii) availability of funds, (iii) the time factor, (iv) the precision required, (v) other relevant factors as stated in Section 5.2 above.

But you should always remember that each method of data collection has its own advantages and limitations and none of them is suitable in all situations. For instance, the observation method is suitable for intensive field surveys to be conducted when the incident is really happening. The interview method is considered suitable in cases where indirect sources of information are required to be tapped because direct observation is not possible. Information through local reports and correspondents is considered a suitable method when information is to be obtained at regular intervals from a relatively wide area. The questionnaire method is appropriate in extensive enquiries where informants are spread over a wide area. This method, however, can be adopted only when the respondents are educated and capable of filling in their responses themselves. Data collection through schedules is suitable in case of extensive enquiries spread over a wide area wherein informants may not always be literate. This method, however, requires lot of funds, relatively more time and a team of dedicated enumerators. Due to its high rate of response this method is usually adopted by the Government in extensive enquiries such as the population census.

In case funds are available and more information is desired, personal interview method can easily be adopted provided the enquiry is confined to a limited area. In case there is sufficient time and limited funds the questionnaire method will be more suitable. Where a wide geographic area is to be covered, the use of questionnaires supplemented by personal interview will yield more reliable results. In short, the most desirable approach while making a choice of method depends on the nature of the particular problem and on the time and resources (financial as well as human) available along with the desired degree of accuracy. Above all those, much depends upon the ability and experience of the investigator. In this context A.L. Bowley's remark "in collection of statistical data commonsense is the chief requisite and experience is the chief teacher", is quite appropriate.

Check Your Progress B

1) Differentiate between an interview and a schedule.

.....

.....

.....

.....

.....

2) Distinguish between a questionnaire and a schedule.

.....

.....

.....

.....

.....

3) Distinguish between an observation and a interview.

.....

.....

.....

.....

.....

4) Name the appropriate method of primary data collection under each one of the following situations :

	Situation	Appropriate Method
i)	Intensive survey is to be conducted when the incident is really happening.	.....
ii)	When the information is to be obtained at regular intervals over a wide area.	.....
iii)	In extensive enquiries where literate informants are spread over a wide area and you have enough time.	.....
iv).	When time is ample, funds are limited, much information is to be gathered, area to be covered is wide consisting of literate persons.	.....
v)	When time is ample, funds are available area is wide, all persons may not be literate, team of sincere and honest enumerators is available.	.....

5.5 SOURCES OF SECONDARY DATA

We have discussed various problems and methods for the collection of primary data. Now let us discuss the secondary data. In case you decide to collect secondary data, you have to look into various sources from where you can obtain it. The source from which you actually collect the data depends upon the nature of the problem. The sources of secondary data can broadly be classified into two categories : i) pubished, and ii) unpublished sources. Let us discuss these two sources in detail.

5.5.1 Published Sources

Data is published and made available to all the interested parties. Usual sources of pubiished data are the following :

1) Reports and official publications of the central and state governments.

2) Various publications of foreign governments or of international bodies and their subsidiary organisations such as the World Bank, International Monetary Fund, United Nations Organisation, etc.

3) Semi-official publications of various local bodies such as municipal corporations, district boards, etc.

4) Private publications, such as :

i) Technical and trade journals such as Commerce, Capital, etc.
ii) Publications of professional bodies like the Institute of Chartered Accountants of India, Institute of Company Secretaries, Institute of Bankers, etc.
iii) Publications of trade and industry organisations like the Federation of Indian Chambers of Commerce, Stock Exchanges, etc.
iv) Annual reports of banks and joint stock companies.
v) Reports prepared by research scholars, universities, economists, etc.
vi) Public records and statistics, historical documents and other sources of published information such as books, magazines, newspapers, etc.

5.5.2 Unpublished Sources

All statistical material is not necessarily available in published form. There are various sources of unpublished data which can also be used wherever necessary. Unpublished data may generally be found in diaries, letters, unpublished biographies and autobiographies. Unpublished data may also be available with scholars and research workers, trade associations, labour bureau and other public/private organisations and individuals.

Thus, there is a vast amount of information available both in published and unpublished sources which constitute the basis for several statistical studies. The investigator may use one or more sources suitable for his project.

5.6 PRECAUTIONS IN USING SECONDARY DATA

As you know, the secondary data has been collected and analysed by someone else. Therefore, while using it you should be very careful. You have to study the data carefully because it may be unsuitable or may be inadequate in the context of your study. It is never safe to take published statistics at their face value without knowing their meaning and limitations. You should always keep in mind the following precautions before using secondary data :

1) Reliability of Data : Secondary'data should only be utilised if they are found reliable. The reliability can be tested by examining the following aspects :

   i) Who collected the data?
   ii) What were the sources of data?
   iii) Were they collected in a proper manner?
   iv) At what time were they collected?
   v) Was the compiled biased?
   vi) What level of accuracy was desired? Was it achieved?

   If the collecting agency happens to be some government institution or international organisation or other competent authority, the secondary data can be taken as more reliable compared to the data collected by individuals or by some private organisation that is not well reputed. Secondary data collected from published sources of government departments and corporations established under the Act of Parliament, and international institutions are reliable.

2) Suitability of Data : The data which may be suitable in one enquiry may not necessarily be suitable in another enquiry. So you should examine whether the data is suitable for your study or not. If the available data is found to be unsuitable, it should not be used. So, you must carefully scrutinise the definition of various terms and units of data collected. Similarly, the object, scope and nature of the original enquiry must also be studied. If these aspects are not found sound, the data will not be suitable for the relevant enquiry and should not be used. For example, you are conducting a survey on wage levels including allowances of workers. If the secondary data is available only on basic wages, such data is not suitable for the present enquiry.

3) Adequacy of Data : Adequacy of the data has to be judged in the light of the requirements of the survey and the geographical area covered by the available secondary data. For example, if our object is to study the wage rate of workers in cotton textile industry of India and the published reports provide the data on wage rates of workers in all industries together, then the data would not serve the purpose. The question of adequacy may also be considered in the light of the time period for which the data are available. For example, for studying price trends we may require data for the last 20 years but the secondary data is available for the last 4 years only. Here the available data would be inadequate and would not serve our object. Similarly, if the level of accuracy achieved in a given data is found inadequate for the purpose of a relevant enquiry, such data should not be used by the investigator.

Thus, we should use given secondary data if it is reliable, suitable and adequate. If secondary data is available from authentic sources and also suitable and adequate for the particular study, it will not be economical to spend time, energy and money in organising field survey for collecting primary data. Thus, if the suitable secondary data is available, as discussed above, it should be utilised with due precaution.

5.7 ADVANTAGES AND DISADVANTAGES OF SECONDARY DATA

There are certain advantages and disadvantages in using the secondary data. Let us now discuss them.

Advantages

1) It is much more economical to use secondary data as we do not need to spend money on printing data collection forms, and hiring large numbers of enumerators.

2) Secondary data, if available, can be obtained more quickly compared to primary data. Secondary data can be collected in a few days whereas it may take months to complete field work for obtaining primary data. As such the investigation may be accomplished in lesser time with the help of secondary data.

3) Secondary data facilitates the work of individual investigator or research organisation when they find it impossible to collect primary data with regard to several subjects. Census data, national income data, etc., cannot be collected by an individual but they can be easily obtained from government publications.

4) The worldwide data concerning diverse phenomena like world trade, industry, population, health, etc., are usually obtainable through secondary sources published by international agencies like United Nations Organisation, World Bank, International Monetary Fund. etc.

5) At time, there may be lots of usable information in the already available data which can well be utilised by the investigator and he can even have new insights concerning the problems he is studying.

6) Most statistical analysis in practice rest upon secondary data since they are readily available in many cases in diverse fields. We use primary data only when secondary data do not provide an adequate basis for analysis.

Disadvantages

1) Secondary data is very risky and is to be used only when their reliability, suitability and adequacy have been ensured. If this is not done, the results of the investigation may not be fully correct.

2) It is difficult to find secondary data which exactly fit the needs of your investigation.

3) There is also the problem of finding secondary data which is sufficiently accurate. Due to bias, inadequate size of sample, errors of definition, etc., the secondary data may be erroneous.

4) Many times, secondary data are not available and in such situations we have to compulsorily collect primary data.

Check Your Progress C

1) Write names of five published sources of secondary data.

.....

.....

.....

.....

2) What are the basic factors to be kept in mind while using the secondary data.

.....

.....

.....

.....

3) State whether each of the following statements is True or False.

i) Using secondary data is always economical than using primary data.

ii) Secondary data is always available in published sources.

iii) Secondary data when used without verifying the suitability, reliability and adequacy may not yield correct results.

iv) Secondary data which is not published is not reliable.

5.8 LET US SUM UP

Statistical data can be either primary data or secondary data. Primary data means the data which is collected for the first time as original data. Secondary data refers to the data collected and processed by someone else earlier and is being used now in the present enquiry. Important factors which affect the choice of data are : 1) object and scope of enquiry, 2) financial resources, 3) the time factor, 4) status of the investigator, 5) human resources, 6) precision required, and 7) availability of secondary data.

There are several problems in the collection of primary data. Important among them are : The problem of defining the unit of data collection, the problem concerning the technique of collection, the problem of developing the frame, the problem of deciding the degree of accuracy, the problem of designing questionnaire/schedule, selection and training of enumerators, the problem of tackling non-response, the problem of control of the field work and other administrative aspects.

There are several methods of collecting the primary data viz., observation, interview, information from correspondentsand local reports, questionnaire and schedules. Each of these methods has its own merits and limitations. As such no one method is appropriate in all situations. The investigator must make a choice of one or the other method as per the needs of the situation after taking into consideration the object and nature of enquiry, availability of funds, the time factor and the precision required.

Secondary data may be obtained from both published and unpublished sources. Secondary data should always be used with precaution. It should be used after ensuring reliability, suitability and adequacy, otherwise it may result in misleading conclusions. Secondary data have certain advantages and disadvantages.

5.9 KEY WORDS

Enumerators : Persons who go to respondents and gather information from them through pre-designed schedules.

Interview : A method of collecting primary data by meeting the informants personally and asking them questions.

Non-response : Problem of not getting the questionnaire duly filled in by the informants,

Observation : A method of collecting the information by observation when the incident is actually happening.

Pilot Survey : Replica and rehearsal of the main survey. The experience gained through the pilot 'survey is made use of in finalising the questionnaire.

Primary Data : Data that is collected for the first time as original data.

Published Sources : Sources which contain published statistical information.

Questionnaire : An instrument for collection of primary data containing a list of questions pertaining to enquiry, generally sent by post to informants and the respondent himself writes the answers.

Schedule : An instrument for'the collection of primary data which contains a set of questions to be filled in by the enumerators who arespecially appointed for the purpose.

Secondary Data : Data which were collected and processed by someone else but are being used in the present enquiry.

5.10 ANSWERS TO CHECK YOUR PROGRESS

A) 4) i) Fasle ii) True iii) True iv) False v) True vi) False vii) False viii) False.

B) 4) i) Observation
ii) Through correspondents and local reports
iii) Questionnaire
iv) Questionnaire
v) Schedule

C) **3)** i) False ii) False iii) True iv) False

5.11 TERMINAL QUESTIONS

1) Differentiate between primary and secondary data. Describe the factors that affect the choice of data to be **used** in an investigation.

2) Describe the problems that are usually confronted by the investigator when he decides to collect primary data.

3) Explain various methods of collecting primary data and also narrate their merits and demerits.

4) "It is never safe to use secondary data without proper scrutinisation" explain.

5) What are the sources of secondary data? Explain the advantages and disadvantages of using secondary data.

Note : These questions will help you to understand the unit better. Try to write answers for them. But do not'submit your answers to the University. These are for your practice only.

UNIT 6 CLASSIFICATION OF DATA

Structure

6.0 OBJECTIVES

After studying this unit, you should be able to :

- explain the meaning of classification
- state the objectives and methods of classification
- describe various terms relating to frequency distribution; and
- construct frequency distribution.

6.1 INTRODUCTION

You have learnt about various sources and methods of collecting primary data and secondary data. As the collected data is in the raw form, you cannot interpret it and draw useful conclusions. Therefore, to draw meaningful conclusions on the basis of collected data, it is essential to present it in summarised and simple form. Classification of data helps us in presenting the mass of data in summarised and simple form. In this unit you will learn the meaning, objectives and different methods of classification. You will also learn about different kinds of frequency distributions and the method of constructing them.

6.2 MEANING OF CLASSIFICATION

Classification means arranging the mass of data into different classes or groups on the basis of their similarities and resemblances. All similar items of data are put in one class and all dissimilar items of data are put in different classes. Statistical data is classified according to its characteristics. For example, if we have collected data regarding the number of students admitted to a university in a year, the students can be classified on the basis of sex. In this case, all male students will be put in one class and all female students will be put in another class. The students can also be classified on the basis of age, marks, marital status, height, etc. The set of characteristics we choose for the classification of the data depends upon the objective of the study. For example, if we want to study the religions mix of the students, we classify the students on the basis of religion.

6.3 OBJECTIVES OF CLASSIFICATION

Classification helps in achieving the following objectives :

1) It helps in presenting the mass of data in a concise and simple form.

2) It divides the mass of data on the basis of similarities and resemblances so as to enable comparison.

3) It is a process of presenting raw data in a systematic manner enabling us to draw meaningful conclusions.

4) It provides a basis for tabulation and analysis of data.

5) It provides us a meaningful pattern in the data and enables us to identify the possible characteristics in the data.

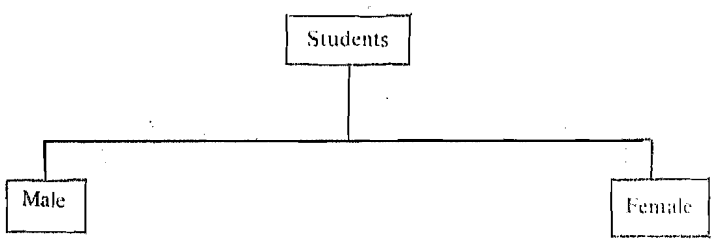
6.4 METHODS OF CLASSIFICATION

You have studied the meaning and objectives of classification. Now let us study the methods of classification. Broadly, there are two methods of classification : i) classification according to attributes, and ii) classification according to variables.

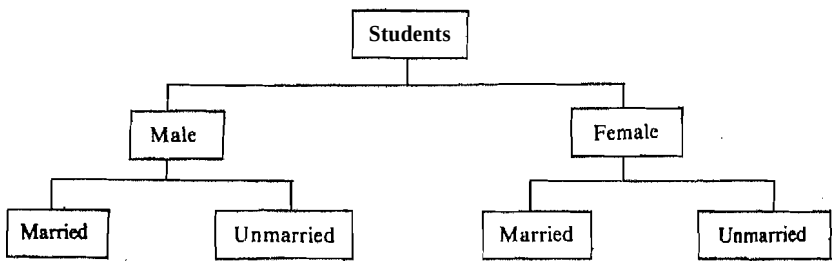
6.4.1 Classification According to Attributes

An attribute is a qualitative characteristic which cannot be expressed numerically. Only the presence or absence of an attribute can be known. For example, intelligence, religion, caste, sex, etc., are attributes. You cannot quantify these characteristics. When classification is to be done on the basis of attributes, groups are differentiated either by the presence or absence of the attribute (e.g. male and female) or by its differing qualities. The qualities of an attribute can easily be differentiated by means of some natural line of demarcation. Based on this natural difference, we can determine the group into which a particular item is placed. For instance, if we select colour of hair as the basis of classification, there will be a group of brown haired people and another group of black haired people. There are two types of classification based on attributes.

1) **Simple Classification :** In simple classification the data is classified on the basis of only one attribute. The data classified on the basis of sex will be an example of simple classification. It can be shown as under :



2) **Manifold Classification :** In this classification the data is classified on the basis of more than one attribute. For example, the data relating to the number of students in a university can be classified on the basis of their sex and marital status as shown below :



6.4.2 Classification According to Variables

Variables refer to quantifiable characteristics of data and can be expressed numerically. Examples of variable are wages, age, height, weight, marks, distance, etc. As you know, all these variables can be expressed in quantitative terms. In this form of classification, the data is

shown in the form of a frequency distribution. A frequency distribution is a tabular presentation that generally **organises** data into classes, and shows the number of observations (frequencies) **falling** into each of these classes. Based on the number of var'ables used, there are three categories of frequency distribution : 1) un'-variate frequency distribution, 2) bi-variate'f equency distribution, and 3) multivariate frequency distribution.

1) Uni-variate Frequency Distribution : The frequency distribution with one variable is called a uni-variate frequency distribution. For example, the students in a class may be classified on the basis of marks obained by them. This is pres'nted in Illustration 1.

Illustration 1

An example of Uni-variate Frequency Distribution.

Marks in Statistic.	No. of Students
0-10	15
10-20	25
20-30	30
30-40	20
40-50	10
Total	**100**

The following points should be noted about the frequency distribution in Illustration 1.

1) The marks in statistics have been div'ded into various classes of 0-10, 1(-20, 20-30, etc.

2) The first class 0-10 marks signifies that the students secur'ng 0 marks or above but less than 10 marks will be put in this class. Similarly, the class 10-20 denotes that the students securing 10 marks or above but less than 20 wi l be placed in this class.

3) The students falling into these classes have been put in the respective classes, which means that there are 15 students in the class 0-10, 25 students in the class 10-20 and so on. The num ber of students falling in a particular class is known as the frequency of that class.

2) Bi-variate Frequency Distribution : The frequency distribution with one variable is called bi-variate frequency distribution. The uni-variate frequency distribution given in Illustration 1 shows only the marks of the students in statistics. If a frequency distribution shows two variables i.e., marks in statistics and age, it is known as bi variate freq ency distribution. J ook at Illustration 2 for an example of bi-variate frequency distribution.

Illustration 2

An example of Bi-variate Frequency Distribution.

Marks i n Statistics	Number of Students with the Age (in Years on Last Birthday)				
	18 years	19 years	20 years	21 ears	Tota:
0-10	1	4	8	2	15
10-20	4	8	9	4	25
20-30	—	3	17	10	30
30-40	—	5	10	5	20
40-50	—	—	1	9	0
Total	5	20	45	30	100

The following points should be noted about the bi-variate distribution presented i n Illustration 2 :

1) The marks in statistics have been divided as 0-10, 10-20, 20-30, etc., whereas age in years-on the last birthday has been taken as 18 years, 19 years, 20 years, etc.

2) The students securing 0 marks or above but less than 10 marks and 18 years of age have been put against 0-10/18 years. This number is 1. Similarly, the number of students falling in 0-10 class but with 19, 20 and 21 years of age are 4, 8, and 2 respectively. The total number of students with 18 years of age is five among which one person is placed 0-10 marks class and the remaining four in 10-20 marks class.

3) Multi-variate Frequency Distribution : The frequency & distribution with more than two variables is called multivariate frequency distribution. For example, the students in a class may be classified on the basis of marks, age and sex. Now let us take the example presented in Illustration 2 and further classify the students based on sex. Study Illustration 3 carefully and examine how it is done.

Illustration 3

Example of Multi-variate Frequency Distribution.

Marks in Statistics	Number of Students with the Age (in Years on Last Birthday)											
	18 years		19 years		20 years		21 years		22 years		Total	
	Male	Fem.	M	F	M	F	M	F	M	F	M	F
0-10	—	1	2	2	5	3	2	—	—	—	9	6
10-20	3	1	4	4	4	5	4	—	—	—	15	10
20-30	—	—	1	2	7	10	7	3	—	—	15	15
30-40	—	—	3	2	5	5	3	2	—	—	11	9
40-50	—	—	—	—	—	1	2	2	3	2	5	5
Total	3	2	10	10	21	24	18	7'	3	2	55	45

Check Your Progress A

1) What is classification?

.....

.....

.....

2) What are the objectives of classification?

.....

.....

.....

.....

3) What is the distinction between simple classification and manifold classification based on attributes'?

.....

.....

.....

.....

4) Distinguish between classification according to attributes and classification according to variables?

.....

.....

.....

.....

5) What is the distinction between uni-variate and bi-variate frequency distributions?

.....

.....

.....

.....

6) State whether the following statements are True or False.

i) Classification of data means arranging the data in different classes on the basis of similarities and resemblances.

ii) Classification helps in presenting the mass of data in concise and simple form.

iii) No comparison is possible with the help of classified data.

iv) The data classified on'the basis of sex and age is an example of simple classification.

v) A frequency distribution is a tabular presentation that generally organises data into classes and shows the observations falling into each of these classes.

vi) A bi-variate frequency distribution shows the distribution based on two variables.

7) From among the words given in the brackets, select the appropriate word and fill in the blank.

i) Attribute refers to ..... aspect of data. (quantitative/qualitative)

ii) The classification of the students of a university on the basis of their height is an example of classification according to ..... (attribute/variable)

iii) The classification of the students of a university on the basis of their religion is an example of classification according to ..... (attribute/variable)

iv) A frequency distribution prepared on the basis of weight of persons is an example of ..... frequency distribution. (uni-variate/bi-variate)

v) A frequency distribution prepared on the basis of age, weight and height is an example of ..... frequency distribution. (bi-variate/multi-variate)

6.5 TERMS RELATING TO FREQUENCY DISTRIBUTION

After studying the meaning and methods of frequency distribution, we should now discuss about the terms relating to frequency distribution. There are several terms relating to frequency distribution. It is essential for us to understand them first.

1) Discrete **and** Continuous Variables : A discrete variable is one which can take only isolated values or it appears by limited gredations. Normally, it does not carry any fractional value. Usually it is the result of counting something. The examples of discrete variable are the number of workers in a factory, number of machines in a factory, number of children in a family, number of accidents on a particular day, etc. A discrete variable is also called a discontinuous variable.

On the other hand a continuous variable is capable of assuming any fractional value within a specified range of values. It is the result of measuring something. The examples of continuous variables are weight of the students of a class, income of the employees of an organisation, age of the residents of a city, etc.

A frequency distribution prepared for discrete variables is called a discrete distribution whereas a frequency distribution prepared for continuous variables is known as continuous distribution. The two types of distribution are shown in Jllustration 4 and Illustration 5.

Illustration 4

An example of Discrete Frequency Distribution.

No. of Goals Scored in a Hockey Match	No. of Matches
0	20
1	15
2	10
3	3
4 & above	2
Total	**50**

Illustration 5

An example of Continuous Frequency Distribution.

Height of Students (in Cms.)	No. of Students
130-140	5
140-150	15
150-160	20
160-170	5
170-180	5
Total	**50**

In Illustration 4, the variable of number of goals scored in a hockey match is discrete variable. Similarly, in Illustration 5, the variable of height of students is a continuous variable. However, it should not be inferred from the above illustrations that a frequency distribution with class intervals can only be prepared for continuous variables. In fact, frequency distribution with class intervals can also be prepared for discrete variables as shown in Illustration 6.

Illustration 6

An example of Frequency Distribution with Class Intervals for Discrete Variables.

No. of Students Present in a Class	No. of Working Days in the Year
10-14	8
15-19	12
20-24	116
25-29	114
Total	**250**

In Illustration *6* the variable on number of students present in a class is a discrete variable. But it is presented here with class intervals like a continuous variable. Such distributions are discrete in nature. They look like continuous distributions in presentation,

2) Class-lipits: Every class or class interval has two limits :

1) lower limit and 2) upper limit. The smallest possible measurement in a class is known as **lower limit** whereas the highest possible measurement is known as the **upper limit.**

Study the Illustration *6* carefully. In the frequency distribution presented in this illustration, the class limits of the first class i.e., 10–14, are 10 and 14. The smallest possible measurement for this class is 10 whereas the largest possible measurement is 14. Similarly, the lower and the upper limits for the second class i.e., 15–19, are 15 and 19 respectively.

3) Mid-point of a Class Interval : The mid-point of a class interval is the point lying half-way between the lower limit and the upper limit. Symbolically, it can be expressed as follows :

$$\text{Mid-point or Mid-value} = \frac{L + U}{2}$$

Whereas, L is the lower limit of the class interval
U is the upper limit of the class interval

In other words, the mid-point is obtained by adding the lower limit and upper limit of a class and then dividing it by two. The mid-points of the various classes in the example on marks in statistics have been calculated in Illustration 7.

Illustration 7

Calculation of Mid-points of Class Intervals.

Marks in Statistics	No. of Students	Mid-point of a Class Interval
0-10	15	$\frac{0+10}{2} = 5$
10-20	25	$\frac{10+20}{2} = 15$
20-30	30	$\frac{20+30}{2} = 25$
30-40	20	$\frac{30+40}{2} = 35$
40-50	10	$\frac{40+50}{2} = 45$

4) Magnitude of a Class Interval : The magnitude of a class interval means the difference between the upper limit and lower limit of a class interval. Look at Illustration 8 carefully and study how the magnitude of the various class intervals has been calculated.

Illustration 8

Calculation of Magnitude of Class Intervals.

Marks in Statistics	No. of Students	Magnitude of Class Intervals
0-10	15	10-1 = 10
10-20	25	20-10 = 10
20-30	30	30-20 = 10
30-40	20	40-30 = 10
40-50	10	50-40 = 10

In Illustration 8 presented above, you must have noticed that the magnitude of the various class intervals is equal. But it is not always necessary that the magnitude of different classes in a distribution must be equal. We may have a distribution with unequal magnitude for different classes.

5) Frequency of a Class : The number of observations falling into the limits of a class is called the frequency of that class. Study lllustration 8 presented above. The frequencies for the first class interval (i.e., 0-10) is 15. Similarly, the frequency for the second class interval (i.e. 10-20) is 25. This implies that 15 students are falling into the limits of the first class interval, 25 students into the limits of the second class interval and so on.

6) Number and Width of Classes : It is not possible to lay down any hard and fast rule regarding the number and width of classes. The number of classes should neither be very large nor too small. Sturges has suggested a formula for determining the number of classes. According to him :

$$K = 1 + 3.3 \log N,$$

Where, N = Number of items to be classified
K = Number of classes ordinarily to be used

Due to other reasons if it is not possible to take the number of classes as K, then usually the number of classes may be between K – 2 to K + 2. For example, if N = 500, we can calculate the value of K as follows :

$$\begin{aligned} K &= 1 + 3.3 \log N \\ &= 1 + 3.3 \log (500) \\ &= 1 + 3.3 (2.6990) \\ &= 1 + 8.9067 \\ &= 9.9067 \\ &= 10 \end{aligned}$$

So the number of classes for 500 items may be 10 or we can go for any number falling in between K – 2 and K + 2, i.e., anywhere between8 to 12. However, this formula only provides a guideline. At the time of determining the number of classes, it is also essential to take 'into account the decision regarding the width of the classes. As stated earlier, it is not necessary that the width of the various class intervals should be the same. If, however, the width of the various classes is kept equal, it provides a sound basis for comparison of data.

7) Open-end Distributions : An open-end distribution is one in which one or two classes lack one class limit. It is possible that the first class interval may not have the lower limit and the last class interval may not have the upper limit in a frequency distribution. Such a distribution is known as an open-end distribution. Look at the Illustration 9 carefully for the example of an open-end distribution.

Illustration 9

Example of an Open-end Distribution.

Marks in Statistics	No. of Students
Less than 10	15
10-20	25
20-30	30
30-40	20
40 and above	10
Total	100

In Illustration 9 presented above, the first class interval is presented as 'less than 10'. In this case there is no lower limit. Similarly, there is no upper limit for the last class interval as 'it is stated as '40 and above'. The open-end distribution creates certain problems such as the calculation of mid-point and computations based on mid-point.

8) Exclusive and Inclusive Frequency Distribution : The frequency distribution can be either exclusive distribution or inclusive distribution. In an exclusive frequency distribution, the upper limit of a given class interval is excluded from that class interval and is taken as the lower limit of the next class interval. Such a distribution has overlapping class limits. In an inclusive distribution, the upper limit of a particular class is included in that class interval. Such a distribution has non-overlapping class limits. The examples of exclusive and inclusive distributions are given in Illustrations 10 and I I.

Illustration 10

An example of Exclusive Frequency Distribution.

Marks in Statistics	No. of Students
0-10	15
10-20	25
20-30	30
30-40	20
40-50	10
Total	100

Illustration 11

An example of an Inclusive Frequency Distribution.

Marks in Statistics	No. of Students
0-9	15
10-19	25
20-29	30
30-39	20
40-49	10
Total	100

In exclusive distribution presented in Illustration 10, in case of the first class interval (0-10) the upper limit (i.e., 10) will be excluded from this class interval and will be taken as the lower limit of the next class interval (i.e. 10-20). Therefore, the meaning assigned to the first class interval will be 0 and below 10, second class interval is 10 and below 20, and so on. However, in inclusive distribution presented in Illustration 11, the upper limit (i.e. 9) of the first class interval (i.e., 0-9) will be included in the first class interval itself. In fact, both the lower and upper limits of a class are included in the same class interval.

In the inclusive method of presenting the class intervals, it is not possible to classify items having values between the upper and lower limits of two consecutive class intervals. For example, in the inclusive distribution given in Illustration 11, items with values 9.3, 9.5, 9.9, 19.1, 19.4, etc. cannot be classified in any of the classes. There is no such difficulty in exclusive method of presenting the class intervals. Therefore, the distributions are generally expressed in exclusive form. However, it is a common practice to use exclusive method for continuous variables and inclusive method for discrete variables.

9) **Class** Boundaries or Real Limits of a Class, Interval : Measurements relating to continuous variables are always recorded correct upto a reasonable degree of accuracy. When height of a person is recorded as 160 cm., it means that the actual height of the person may be anywhere between 159.5 cm and 160.5 cm. If such data (recorded in integral values only) is classified by using inclusive type class intervals e.g., 155-159, 160-164, 165-169, etc., then the group 160-164 will include all persons with actual height between 159.5 cm and 164.5 cm. The limits 159.5 and 164.5 are called the lower and the upper class boundaries or real limits of the inclusive class interval 160-164. The lower and upper class limits of this class are only 160 and 164, the figures used in writing the class interval. Thus, lower class boundary of any inclusive type class interval is 0.5 less than its lower class limit and upper class boundary is 0.5 higher than the upper class limit.

Conversion of class limits to real limits is necessary when continuous variables recorded in discrete form are classified by using inclusive type class intervals. You will come across with such situations at the time of analysis of data discussed in Units 11 and 12.

10) Cumulative Frequency Distribution : The cumulative frequency distribution shows cumulative frequencies and not the actual frequencies for the various classes. The cumulative frequency of the first class interval is the same as its actual frequency. The cumulative frequency of the second class interval is obtained by adding the frequency of the first class interval and the frequency of the second class interval. Similarly, the cumulative frequencies of the various other class intervals are determined. Look at Illustration 12 carefully and study how the frequency distribution is converted into cumulative frequency distribution.

Illustration 12

Calculation of Cumulative Frequency Distribution.

Marks in Statistics	No. of Students (Frequency)	Cumulative Frequency
0-10	15	15
10-20	25	25 + 15 = 40
20-30	30	30 + 40 = 70
30-40	20	20 + 70 = 90
40-50	10	10 + 90 = 100

A cumulative frequency distribution can be expressed as shown in Illustration 13.

Illustration **13.**

Presentation of Cumulative Frequency Distribution.

Marks in Statistics	No. of Students (Cumulative Frequency)
Less than 10	15
Less than 20	40
Less than 30	70
Less than 40	90
Less than 50	100

A cumulative frequency distribution can either be a "less than" cumulative frequency distribution or a "more than" cumulative frequency distribution. The frequency distribution presented in Illustration **13** is an example of a "less than" cumulative frequency distribution. In this type of distribution, the cumulative frequencies are in ascending order.

In a more than cumulative frequency distribution, the frequency of the last class interval is taken as the cumulative frequency of that class. The cumulative frequency of the class before the last class interval is obtained by adding the frequency of that class to the cumulative frequency of the succeeding class interval. In this type of distribution the cumulative frequencies are in descending order. Look at Illustration 14 carefully and study how a "more than" cumulative frequency distribution is calculated.

Illustration **14**

calculation of "More Than" Frequency Distribution.

Marks in Statistics	No. of Students (Frequency)	Cumulative Frequency
0-10	15	15 + 85 = 100
10-20	25	25 + 60 = 85
20-30	30	30 + 30 = 60
30-40	20	20 + 10 = 30
40-50	10	10

The "more than" frequency distribution can also be expressed as shown in Illustration 15.

Illustration **15**

Presentation of "More Than" Frequency Distribution

Marks in Statistics	No. of Students (Cumulative Frequency)
More than 0	100
More than 10	85
More than 20	60
More than 30	30
More than 40	10

You should note that "less than" cumulative frequencies are related to the upper limits of the class intervals and "more than" cumulative frequencies are related to lower limits of the class intervals.

Check Your Progress B

1) Distinguish between a discrete variable and a continuous variable.

.....

.....

.....

2) What is the magnitude of a class interval?

.....

.....

.....

3) What is the mid-point of a class interval?

.....

.....

.....

4) What is the frequency of a class interval?

.....

.....

.....

5) What is an open-end distribution?

.....

.....

.....

.....

6) Distinguish between exclusive and inclusive frequency distributions.

.....

.....

.....

7) Distinguish,between "less than" and "more than" cumulative frequency'distributions.

.....

.....

.....

8) State whether the following statements are True or False.

i) A discrete variable can be measured.

ii) A discrete variable carries a fractional value.

iii) A continuous variable can take a fractional value.

iv) The smallest possible measurement in a class is known as the upper limit.

v) The mid-value of a class-interval is half way between the lower limit aad upper limit of a class.

vi) The magnitude of a class is the differencebetween the upper and lower limits.

vii) The number of observations falling into the limits of a class is called the frequency of the class.

viii) The number of classes in a class interval can be either large or small.

ix) The open-end distribution creates problems in the calculation of mid-point.

x) In a "more than" cumulative frequency distribution, the cumulative frequencies are in ascending order.

9) Fill in the blanks with the appropriate word given in the brackets.

i) A continuous variable is the result of ..... (counting/measurement)

ii) Height of the students of a class is an example of..... variable. (continuous/discrete)

iii) The number of runs scored in a cricket match is an example of..... variable. (discrete/continuous)

iv) It is not always necessary that the magnitude of different classes must be ..... (equal/unequal)

v) In an open-end frequency distribution, the first class interval lacks ..... class limit(s). (one/two)

vi) In an'inclusive frequency distribution, class limits are ..... (overlapping/non-overlapping)

vii) In inclusive frequency distribution, the upper limit of a class is,..... in that class. (included/excluded)

viii) The "less than" cumulative frequency distribution has cumulative frequencies in ..... order. (ascending/descending)

6.6 FORMATION OF A FREQUENCY DISTRIBUTION

You have studied various terms relating to frequency distribution. Now let us study the detailed procedure relating to the formation of frequency distribution.

6.6.1 Data Array

In the process of the formation of a frequency distribution, data array is the first step. **Data array is an orderly arrangement of data either in ascending order or in descending order.** The data array is useful as it gives us information about the range of data and throws some light on the nature of the data. However, the data array is very useful when the number of observations is not very large. In case of a very large number of observations, the data array will be very unwieldy. Study Illustration 16 carefully and understand how data array is prepared.

Illustration 16

Array the following data relating to the marks secured in statistics by the students of a class.

17	21	9	26	17	10	9	26	44	17
27	28	23	35	45	36	20	13	39	29
30	35	29	39	41	48	40	43	33	48
30	15	16	47	31	49	46	48	36	14

Solution

The marks secured in statistics by the students of a class can be presented in the ascending order as follows:

39	36	9	14	17	23	28	31	48	44
40	36	9	15	18	26	29	33	48	45
40	36	10	16	20	26	29	33	48	45
43	39	13	17	21	27	30	35	49	47

From the data array prepared in Illustration 16, we can clearly understand the minimum and the maximum marks secured by the students in statistics.

6.6.2 Steps in Constructing a Frequency Distribution

As stated earlier, the data array becomes unwieldy when the number of observations is very large. In such cases frequency distribution is constructed to condense the size of the data. The following steps are necessary before constructing a frequency distribution.

i) Formation of the class intervals with class limits clearly identified.

ii) Each individual item is taken up and a tally bar is placed against the appropriate class interval. Tally bar is a vertical bar. One tally bar is drawn for every value.

iii) After the four tally bars have been drawn the fifth bar is drawn across as shown here.

iv) When all the observations have been placed against the appropriate classes, these are totalled up and recorded as frequency of that particular class interval.

Illustration 17

From the data given in Illustration 16, form a frequency distribution taking 0-10 as the first class interval.

Solution

The first class interval is of size 10 and the largest item is 49. Taking all the class intervals of equal length, the other class intervals will be 10-20, 20-30, 30-40, 40-50. You should note that the given data is discrete. But still exclusive class intervals have been taken which is the most common method in use. The meaning of the class interval 0-10 is 0 and below 10. Therefore, it includes items with values ranging from 0 to 9. Similarly, other class intervals will include items in ranges 10 to 19, 20 to 29, etc.

Frequency Distribution of Marks Secured in Statistics

Marks in Statistics	Tally	Bars	No. of Students (Frequency)
0-10	11		2
10-20	𝍸	111	8
20-30	𝍸	1111	9
30-40	𝍸	𝍸	10
40-50	𝍸	𝍸 1	11
Total			**40**

Illustration 18

Prepare a discrete frequency distribution from the following data relating to the number of typing errors on a page committed by a typist.

0 0 2 1 3 0 4 ~ 0 4 3 3 1 2 0 3 2 3 3
1 2 1 3 4 2

Solution

The smallest number of errors per page is 0 and the largest number of errors is 4. We have to construct a discrete frequency distribution so that the number of errors will be taken as 0, 1, 2, 3 and 4.

Frequency Distribution

No. of Errors per page	Tally	Bars	No. of Pages (Frequency)
0	𝍸		5
1	1111		4
2	𝍸	1	6
3	𝍸	11	7
4	111		3
Total			**25**

6.6.3 Guidelines for Selecting the Class Intervals

If the class intervals are not given, then their number and width may be determined on the basis of the following guidelines :

i) Find the smallest and the largest items of the data. Calculate their difference. This gives the range of the data.

ii) Count the number of items and decide the number of classes "K", by Sturge's rule or as per convenience.

iii) Divide the range by number of classes 'K'. Let us call it 'C'. This value 'C' will be the basis for determining the width of the class interval.

iv) Usually the width of the class intervals should be same throughout and generally the limits of the class interval should be a round number i.e., a multiple of 5 or 10. So the possible limits of class intervals, 'i' will be the lowest limit, a round number, just less than 'C' and the upper limit, a round number, just more than 'C'.

v) The starting point of writing the first class interval will be taken as a round number just less than the smallest item of the data or equal to it if this smallest item itself is a round number.

vi) Now write the full sets of class intervals with the help of suitable values of 'i' (determined under step iv taking the starting point decided under step v above). Out of these sets, the set which has the number of class intervals as 'K' will be taken as the ideal set. If none of the sets has 'K' class intervals, then a set which has a number of class intervals between the two numbers K–2 to K+2 is selected.

Illustration 19

Prepare a frequency distribution by inclusive method for the following data on weight (in Kgs.) of students of a class.

42	47	48	50	41	61	57	52	57
47	41	59	60	63	42	44	45	46
57	47	57	53	48	47	56	61	56
62	49	58	52	55	42	42	56	51

Solution

The number of class intervals is not given, so let us first determine the number and width of class intervals.

i) The minimum weight is 41 Kgs. and the maximum weight is 63 Kgs. So the range is 22.

ii) The number of items is 45. So by Sturge's rule, the number of class intervals may be calculated as follows :

$K = 1 + 3.3 \log 45$
$= 1 + 3.3 \times 1.65$
$= 6.4$
$= 7$ (approximated)

iii) Possible width of the class intervals could be

$$C = \frac{\text{Range}}{K}$$
$$= \frac{22}{7}$$
$$= 3.1$$

iv) A round number higher than 3.1 is 5. Round number less than 3.1 is '0'. So the length of class interval 'i' is 5. Here we have only one value of 'i'.

v) Since smallest value is 41, the first class interval will start from 40. So the different class intervals of inclusive type and size 5 will be 40-44, 45-49, 50-54, 55-59 and 60-64. Here the last class interval is 60-64 because the largest item is 63.

vi) The number of class intervals in this set is 5. This is not equal to its ideal number K which is 7, but it is within the limits of K–2 to K + 2 (i.e., 5 to 9). So we can proceed with the frequeqcy distribution.

Frequency Distribution (Inclusive Method)

Weight (in Kgs)	Tally Bars	No. of Students (Frequency)
40-44	𝍸 111	8
45-49	𝍸 𝍸	10
50-54	𝍸 111	8
55-59	𝍸 𝍸	8
60-64	𝍸 𝍸 1	11
Total		45

Illustration 20

Prepare a bi-variate frequency distribution for the following data of 20 candidates regarding the marks in internal assessment (X) and percentage in final examination (Y).

X : 10 11 13 12 12 14 12 13 14 13 14 10 12 11 11 12 13 12 13 13

Y : 12 20 52 43 50 60 58 54 78 81 69 38 47 41 45 49 63 68 56 72

Solution

Marks in internal assessment (X) have only five different values 10, 11, 12, 13 and 14. So each of them will be taken to represent a group. The percentage of marks in final examination (Y) has a lowest figure of 12 and highest figure of 81, giving a range of 69. By Sturge's rule for 20 items the ideal number of classes will be 5. Thus probable length of each class will be 69/5 or 14. This may be rounded to 15. As the minimum percentage (Y) is 12, the lower limit of the first class interval will be taken as 10. Thus the various class intervals for 'Y' will be 10-25, 25-40, 40-55, 55-70, and 70-85. Writing 'X' variable in the first row and 'Y' variable in the first column, the outline of the bi-variate distribution will be as follows :

Percentage In Final Examination (Y)	Marks in the Internal Assessment					Total
	10	11	12	13	14	
10-25						
25-40						
40-55						
55-70						
70-85						
Total						

Now to get the bi-variate frequency distribution the number of frequencies will be determined by allocating corresponding values of 'X' and 'Y' to various cells and drawing tally bars. The first set of corresponding values of X and Y has values of 10 and 12. So a tally bar will be drawn in the cell representing columns loaded by 10 and the row headed by 10-25. In the same way tally bars for other cells will be drawn. When all the values have been allocated to various cells, and the respective tally bars have been drawn, the required bi-variate frequency distribution will be as follows :

Bi-variate Frequency Distribution

Percentage in Final Examination	Marks in the Internal Assessment					Total
	10	11	12	13	14	
10-25	1(1)	1 (1)				2
25-40	1(1)					1
40-55		11(2)	1111(4)	11(2)		8
55-70			11 (2)	11(2)	11(2)	6
70-85				11(2)	1(1)	3
Total	2	3	6	6	3	20

Check Your Progress C

1) What is data array?

.....

.....

.....

2) What are the steps in the formation of a frequency distribution?

.....

.....

.....

3) State whether the following statements are True or False.

i) Data array is an orderly arrangement of data either in ascending or in descending order.

ii) Data array does not provide any information regarding the range of data.

iii) Data array is very useful when the number of observations is very large.

iv) A frequency distribution helps to condense the data.

v) Vertical bar placed against the appropriate class interval is called tally bar.

6.7' LET US SUM UP

Classification means arranging the data into different classes on the basis of similarities and resemblances. It helps in presenting unwieldy data in a concise and simple manner. Classification provides us a basis for tabulation and analysis and helps us to identify the possible characteristics of data. Classification may be done according to attributes or variables. Classification according to attributes is based on qualitative characteristics of data. Simple classification is done on the basis of one attribute whereas manifold classification is done on the basis of more than one attribute. Classification according to variables is based on the quantifiable characteristics of data. The data is shown in the form of frequency distribution. A frequency distribution can be uni-variate or bi-variate or multi-variate.

There are several terms relating to frequency distribution. A discrete variable takes an isolated value whereas a continuous variable can take any fractional value within a specified range of values (except open-end). Every class interval (except open-end) has two limits : 1) lower limit, and 2) upper limit. Mid-point of a class lies half-way between the two limits of a class. Magnitude of a class interval means the difference of the two limits. Frequency of a class implies the number of observations falling into limits of a class.

The number of classes in a distribution can be decided on scientific basis. The decision regarding the width of classes should also be taken while determining their number. The lack of lower limit of the first class and the upper limit of the last class will indicate open-end class interval. In exclusive class interval, the upper limit of a class is excluded from the class interval and is taken as the lower limit of the next class interval. In inclusive class interval the upper limit of a class is also included in the same class. In exclusive class the limits are overlapping whereas in inclusive class they are non-overlapping. In a cumulative frequency distribution, the classes indicate cumulative frequencies which may be either a "less than" or a "more than" cumulative frequency distribution.

Data array is an orderly arrangement of data either in ascending or descending order. It provides information about the range and the characteristics of data. A frequency distribution is prepared to condense the size of the data.

6.8 KEY WORDS

Attribute : An expression of a qualitative characteristic of facts.

Bi-variate : Having two variables.

Classification : Division of data on the basis of similarities and resemblances.

Class **Limits** : The lower and upper limits of class interval.

Continuous Variable : A variable which can take any fractional value within a specified range of values.

Cumulative Frequency Distribution : A distribution which shows cumulative frequencies instead of actual frequencies.

Data **Array** : An orderly arrangement of data either in ascending or descending order.

Discrete Variable : A variable which can take any isolated value.

Exclusive Class : A class in which its upper limit is excluded from that class and included as lower limit in the next class.

Frequency : Number of observations falling into a particular class.

Inclusive Class : A class in which both its lower and upper limits are included in that very class.

"Less Than" Distribution : A cumulative frequency distribution showing cumulative frequencies in ascending order.

Magnitude of a Class : The difference between upper limit and lower limit of a class.

Mid-point : The point which lies half-way from the lower limit and higher limit.

"More Than" Distribution : A cumulative frequency distribution in which the cumulative frequencies are in descending order.

Multi-variate : Having multiple variables.

Open-end : A distribution in which the end of a class is open (not closed).

Tally Bar : Vertical bar representing an observation placed against the relevant class.

Uni-variate : Having one variable.

Variate : Refers to quantifiable characteristics of data.

6.9 ANSWERS TO CHECK YOUR PROGRESS

A) 6) (i) True, (ii) True, (iii) False, (iv) False, (v) True, (vi) True

7) (i) qualitative, (ii) variable, (iii) attribute, (iv) uni-variate, (v) multi-variate

B) 8) (i) False (ii) False (iii) True (iv) False (v) True (vi) True (vii) True (viii) False (ix) True (x) False

9) (i) measurement, (ii) continuous, (iii) discrete, (iv) equal, (v) one, (vi) overlapping, (vii) included, (viii) ascending

C) 3) (i) Ttue (ii) False (iii) False (iv) True (v) True

6.10 TERMINAL QUESTIONS/EXERCISES

Questions

1) Explain the meaning and objectives of classification. Also, discuss the various methods of classification.

2) Write a short note on each of the following:
   i) Discrete variable
   ii) Class limits
   iii) Mid-point of a class
   iv) Number of classes in a distribution
   v) Open-end distribution
   vi) Cumulative frequency distribution
   vii) Data array

Exercises

1) Array the followingdata in ascending order.

3 19 17 25 23 21 19 17 15 13 11 9 10 26 34
32 30 28 26 24 22 20 18 16 15 31 29 27 25
23 21 19 17 15 13 11 18 28 26 24 22 20 18
16 14 12 10 8

2) Array the following data in ascending order.

51 28 29 75 33 25 73 75 67 27 61 48 81 66 45
37 61 55 47 39 53 61 53 55 34 49 49 45 36
54 47 73 21 44 36 61 37 35 29 61

3) Marks in accounts for the second year B.Com. students are given below. Form a frequency distribution by inclusive method taking 0-19 as the first class interval.

55 33 35 5 23 37 73 75 87 29 97 80 66 53 87
71 4 25 93 66 47 93 81 29 58 66 59 62 29
61 21 37 46 27 42 71 52 78 27 47 16 49 91
9 3 8 7 1 1 6 1

4) Form a frequency distribution for the following data by exclusive method.

2 18 25 16 21 21 16 15 12 13 11 9 9 25 28
34 28 28 22 24 20 18 18 14 14 30 27 25 23
21 19 15 11 6 9 16 27 15 22 20 17 18 13
14 10 10 7

5) Draw a frequency distribution for the following data related to the number of goals scored by a football team in different matches.

0 2 4 1 3 0 3 2 3 0 0 1 5 2 0 3 3 4 2
1 1 3 1 3 2 1 0 1 3 1 2 1 2 1 2 2 0 1
2 1 0 2 1 0

6) Draw a "less than" cumulative frequency distribution for the following data on monthly wages and number of workers.

Wages: (Rs.)	450-475	475-500	500-525	525-550	550-575	575-600
No. of Workers :	30	45	60	40	15	10

7) Draw a "more than" cumulative frequency distribution for the following data :

Marks:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Students :	12	17	31	40	28	22	12	8

Note : These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.

UNIT 7 TABULAR PRESENTATION

Structure

7.0 OBJECTIVES

After studying this unit, you should be able to :

- describe the meaning and objectives of tabulation
- list the parts of a statistical table
- outline the essentials of a good statistical table, and
- construct a statistical table.

7.1 INTRODUCTION

In Unit *6* we have discussed the objectives of classifying the mass of data so as to render comparison of data possible. We have also explained the procedure for the construction of a frequency distribution involving one or two variables. When two variables are given, the arrangement in rows and columns is ordinarily known as a statistical table. Such tables can be constructed even when the given data relates to attributes. In this unit, you will study in detail the meaning and objectives of tabulation and the procedure of constructing statistical tables.

7.2 MEANING OF TABULATION

The tabular presentation of data is one of the techniques of presentation of data, the two other techniques being diagrammatic presentation and graphic presentation. **The tabular presentation means arranging the collected data in an orderly manner in rows and columns.** The horizontal arrangement of the data is known as **rows,** whereas the vertical arrangement is called **columns.** The classified facts are recorded in rows and columns to give then tabular form.

7.3 OBJECTIVES OF TABULATION

Tabular presentation serves the following objectives:

1) **Systematic** Presentation of Data : Generally the collected data is in fragmented form. The mass of data is presented in a concise and simple manner by means of statistical tables. Thus, tabulation helps in presenting the data in an orderly manner.

2) **Facilitates** Comparison of Data : If the data is in the raw form, it is very difficult to compare. Comparison is possible when the related items of data are presented in simple and concise form. The presentation of complete and unorganised data in the form of tables facilitates the comparison of the various aspects of the data.

3) Identification of the Desired Values : In tabulation, data is presented in an orderly manner by arranging it in rows and columns. Therefore, the desired values can be identified without much difficulty. In the absence of tabulated ,data, it would be rather difficult to locate the required values.

4) Provides a Basis for Analysis : Presentation of data in tabular form provides a basis for analysis of such data. The statistical methodology suggests that analysis follows presentation of data. A systematic presentation of data in tabular form is a prerequisite for the analysis of data. Statistical tables are useful aids in analysis.

5) Exhibits Trend of Data : By presenting data in a condensed form at one place, tabular presentation exhibits the trend of data. By looking at a statistical tables, you can identify the overall pattern of the data.

7.4 DISTINCTION BETWEEN CLASSIFICATION AND TABULATION

Several people consider classification and tabulation as synonyms. The two also appear to convey the same meaning and also serve the same objectives. However, there is a difference between the two. In classification, the data is divided on the basis of similarity and resemblance, whereas tabulation is the process of recording the classified facts in rows and columns. Here, the two belong to the same chain. Tabulation begins where classification ends. In fact, classification provides a basis for tabular presentation. In 'Unit 6 we have stated that the frequency distribution is a tabular presentation of the number of observations falling against different sizes or classes. Therefore, after classifying the data into various classes, they should be shown in the tabular form.

7.5 KINDS OF TABLES

Depending upon the use and objectives of the data to be presented, there are different types of statistical tables. They can be classified under the following broad heads :

1) Information or Classifying tables

2) General Purpose or Reference Tables

3) Special Purpose or Summary Tables

7.5.1 Information or Classifying Tables

This type of tables is prepared to show the important characteristics of the collected facts. The tables are prepared on the basis of similarities in the collected data. The main purpose of preparing this type of tables is to present the data in a condensed and simple form. These tables can be further classified as : i) simple tables, and ii) complex tables.

1) Simple Tables : This type of tables is also known as one way tables. These tables are prepared on the basis of only one characteristic of the collected data. The table showing the data relating to the number of students in a college in different years will be an example of simple or one way table. Look at Illustration 1 for an example of a simple table.

Illustration **1** Example of a simple table

Number of Students in a College from 1982-83 to 1988-89

Year	No. of Students
1982-83	1500
1983-84	1550
1984-85	1600
1985-86	1650
1986-87	1600
1987-88	1675
1988-89	1700

Similarly, the students of the college can be divided on the basis of their age, and separate simple tables for each year can be prepared.

2) Complex Tables : As you know simple tables present only one characteristic of the data. When the tables show more than one characteristic of the data, they are called complex tables. We may have a two-fold table showing three characteristics or a many-fold table showing several characteristics of the data. The table showing the number of students in a college on the basis of their sex and marital status during different years is an example of a complex table. Look at Illustration 2 for an example of a complex table.

Illustration 2 An example of a complex table

Sex and Marital Status of Students in a College during 1982-83 to 1988-89

Year	No. of Students				Total
	Male		Female		
	Unmarried	Married	Unmarried	Married	
1982-83	950	50	475	25	1,500
1983-84	975	55	490	30	1,550
1984-85	1,000	55	510	35	1,600
1985-86	1,035	60	520	35	1,650
1986-87	1,010	50	510	30	1,600
1987-88	1,080	50	510	35	1,675
1988-89	1,090	55	515	40	1,700
Total	7,140	375	3.530	230	11,275

7.5.2 General Purpose or Reference Tables

This type of tables are prepared to store information and they contain wide range of information relating to a specified subject. Such tables are complex tables and are generally found as appendices to various reports. These tables should be prepared in a systematic manner so as to render references easier. The tables appended to the census reports are good examples of general purpose or reference tables.

7.5.3 Special Purpose or Summary Tables

These tables show a specific point relating to data and are helpful in statistical analysis. They provide a basis for comparison by indicating specific answers to given questions. These tables are also called text tables as they arecomplementary to a given text. These tables indicate rates, percentages, averages, etc. For instance take the study discussing the increasing rate of industrial accidents in a country and the number of persons killed in these accidents. The table shown in Illustration 3 can follow the text to show high rate of persons killed in accidents in coal mines.

'Illustration3

An example of special purpose or summary tables.

Relationship Between the Total Number of Persons Died in Industrial Accidents and Persons Died in Coal Mines

Year	Persons Died in Industrial Accidents	Persons Died in Coal Mines	Persons Died in Coal Mines as a % in Total Deaths in Industrial Accidents
1976	930	150	16.1
1977	1,154	285	24.7
1978	1,250	115	9.2
1979	930	108	12.0
1980	1,350	270	20.0

Check Your Progress A

1) What is tabulation?

.....

.....

.....

.....

2) What are the objectives of tabulation?

.....

.....

.....

.....

3) What is the difference between classification and tabulation?

.....

.....

.....

.....

4) Distinguish between simple and complex tables.

.....

.....

.....

.....

5) What is a general purpose or reference table?

.....

.....

.....

.....

6) What is a special purpose table?

.....

.....

.....

.....

7) State whether the following statements are True or False.

i) Tabulation is the technique of analysis of data.

ii) Tabulation means arranging the collected data in an orderly manner in rows and columns.

iii) A column is a vertical arrangement of data.

iv) Tabulation facilitates comparison of data.

v) In the absence of tabulated data it is simple to identify the desired values.

vi) Statistical tables are useful in analysis.

vii) Overall pattern of data can be identified with the help of tabulation.

viii) Simple table shows more than one characteristic of data.

ix) The information relating to the age, religion and marital status will be presented by means of a simple table.

x) General purpose table contains wide range of information.

xi) Appendices to the Census 1981 are an example of reference tables.

xii) Special purpose tables provide a basis for comparison.

8) Fill in the blanks with the appropriate word given in the brackets.

i) A row is a .....arrangement of data. (horizontal/vertical)

ii) Recording of classified facts in rows and columns is known as ..... (classification/tabulation)

iii) Statistical tables help in presenting data in a ..... manner. (simple/complex)

vi) Tabulation provides a basis for .....(collection/analysis)

v) Classification and tabulation convey .....meaning. (same/different)

vi) Information tables are prepared to show data in .....form. (condensed/unorganised)

vii) Complex table presents .....characteristics of the given data. (one/more than one)

viii) General purpose tables are given as .....(part of the main body of the report/appendices to the report)

ix) Special purpose tables are helpful in ..... . (collection/analysis)

x) Special purpose tables a n ..... to a given tent. (independent/complementary)

7.6 CONSTRUCTION OF A STATISTICAL TABLE

You have studied that the basic objective of tabulation is to present unorganised data in orderly form so that analysis of the data becomes easier. To achieve this objective, we should be very careful in tabulating the data. To ensure this, it is very important to have a clear understanding of the rules and practices followed in the construction of statistical tables. Before we proceed for the construction of a statistical table, we should discuss the major parts and features of a good statistical table.

7.6.1 Parts of a Statistical Table

A statistical table, in general, should have the following parts. While studying about each part, you refer to Figure 7.1 for illustration.

Figure 7.1 : Different Parts of a Statistical Table

TABLE NUMBER **TITLE**

Head Note

Stub	Caption		Total
	Caption Subhead	Caption Subhead	
Stub-Entries	Field		
Total			

Footnotes:

Source:

1) **Title :** There should be a title at the top of every statistical table. The title should be clear, concise and adequate. It should clearly indicate the description of the data being presented in the table and should also hint at the usefulness of the information being depicted in the table. The title should answer the questions : What is the data? where is the data? how is the data classified? and, what is the time period of data?

2) **Table Number :** Every table should be identified by a number. It facilitates easy reference. Whenever you refer to the table in the text, you can give the number of the table only. The number may be either Arabic or Roman. It can be placed at the beginning of the title of the table or can be centred above the title of the table.

3) **Head Note :** Head note is written just below the title, preferably on the right hand corner. Read note'indicates the unit in which the data have been given.

4) **Stub :** As you know, a statistical table is an arrangement of data in rows and columns. The titles given to the rows are called 'stubs'. The stubs should be clearly stated. Stubs should clearly describe the data presented in the rows of the table.

5) **Caption :** The title of the columns are referred to as captions. They are also called 'box heads'. The caption labels the data presented in a column of the table. The caption should be clearly stated. There may be **sub-heads or sub-captions** in each caption.

6) **Body or Field :** The body of the table is the most important part. The information given in the rows and columns forms the body of the table. It contains the quantitative information to be presented.

7) **Footnote :** Any explanatory notes concerning the table itself, placed directly beneath the table, is called 'footnote'. The main purpose of footnote is to clarify some of the specific items given in the table or to explain the ambiguities, omissions, if any, about the data shown in the table.

8) **Reference Note :** If the data is collected from secondary sources, a'note is given to disclose the sources from which the data is collected. Such'note is called reference note. This reference note is placed beneath the table along with the footnotes. The reference note should be clear and specific as this becomes the basis of verifying the reliability of'data.

9) **Totals :** The totals and sub-totals of all rows and columns should also be given in the table.

7.6.2 Requisites of a Good Statistical Table

You have studied the parts of a statistical table. Now let us discuss the features of a statistical table. There are certain general guidelines in preparing a good statistical table. They are as follows :

I) A good table must present the data in a clear and simple manner.

2) It should have a brief and clear title. The title should be self-explanatory and should represent the description of the contents of the table.

3) The stub, stub entries, captions and caption heads should be brief and clear. The columns may be numbered to facilitate easy reference in the text.

4) The headnote should be precise and complete as it relates to the unit of the data.

5) The totals and sub-totals should be given at the appropriate places.

6) The references should be clearly stated so that the reliability of the data could be verified if needed.

7) If necessary, the derived data (ratios, percentages, averages, etc.) may also be incorporated in the tables.

8) .As far as possible abbreviations should be avoided in a statistical table. If it is essential to use abbreviations, their meaning must be explained in footnotes.

9) Wherever necessary, proper ruling should be provided in a table. Normally, the columns are separated from one another by lines. These lines make the table more readable and attractive, and also show the relations of the data more clearly. Always lines are drawn at the top and bottom of the table, and also below the captions.

10) Use of ditto mark should be avoided.

11) Columns and rows which are to be compared with one another should be placed side by side.

12) If it is necessary to emphasise the relative significance of certain categories, different kinds of type spacing and indentation should be used.

13) All the column figures should be properly aligned. Decimal points and plus-minus signs also should be in perfect alignment.

14) Generally not more than four to five characteristics may be shown at a time in a table, otherwise it will become too complex.

7.6.3 Preparation of Statistical Tables

You know the parts ana features of a good statistical table. Now let us take some illustrations and learn the, construction of tables.

Illustration 4

Prepare a blank table showing the marks, age and sex of the students of a college in 1988, The mark groups should be taken as 0-10, 10-20, 20-30, 30-40 and 40-50 whereas the age should be taken in years as 17, 18, 19 and 20.

Solution

The table is to show marks, age and sex. The marks can be represented by means of rows (stub) and the age as column (caption) and the sex as sub-column (caption head). The blank table can be presented as follows :

Marks, **Age and Sex of the Students in 1988**

Marks	Age in Years								Total	
	17 Years		18 Years		19 Years		20 Years			
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
0-10										
10-20										
20-30										
30-40										
40-50										
Total										

Footnote:

Source:

Illustration 5

The data on manufacture of cars by different manufacturers in India showed that in 1985-86 the share of Maruti Udyog is 48.4%, Premier Auto 28.5%, and Hindustan Motors 22.6%. In 1986-87, the share of Hindustan Motors dropped to 18%, whereas it went upto 59.2% for Maruti Udyog. The share of other car manufacturers was 1.2% in 1986-87. Present this data in a tabular form.

Solution

The data is given in percentages. In 1985-86 the total share of Maruti Udyog, Premier Auto and Hindustan Motors is 99.5%. This implies that 0.5% of the cars were manufactured by other manufacturers. In 1986-87, the share of Hindustan Motors dropped to 18% from 22.6% whereas it went upto 59.2% from 48.4% in case of Maruti Udyog and the shares of other manufacturers increased to 1.2% from 0.5%. It implies that the balancing figure 21.6% is the share of Premier Auto. The data can be presented in the tabular form as below :

Shares of Different Manufacturers in the Production of Cars in India

(Percentages)

Name of the Manufacturer	1985-86	1986-87
I) Maruti Udyog	48.4	59.2
2) Premier Auto	28.5	21.6
3) Hindustan Motors	22.6	18.0
4) Other Manufacturers	0.5	1.2
Total	**100.0**	**100.0**

Illustration 6

The Central Government outlay on labour and labour welfare sector during the Seventh Five Year Plan was Rs. 9,510 lakhs, out of which Rs. 4,184 lakhs was allocated to training, Rs. 546 lakhs on employment services, Rs. 3,180 lakhs on labour welfare, and Rs. 1,600 lakhs on rehabilitation of bonded labour. The total outlay for the year 1986-87 was Rs. 1,847 lakhs of which Rs. 465 lakhs was on training, Rs. 107 lakhs on employment services, Rs. 755 lakhs on labour welfare and the balance on rehabilitation of bonded labour. Compared to the previous year, during 1987-88 the total outlay increased by Rs. 49 lakhs. As compared to 1986-87 the outlay on employment services decreased by Rs. 11 lakhs, on labour welfare it decreased by Rs. 148 lakhs and on rehabilitation of bonded labour decreased by Rs. 327 lakhs. It should be noted that the outlay does not include coaching-cum-guidance scheme of scheduled castes and scheduled tribes. This was collected from Annual Plan 1987-88, page 335, Planning Commission, Government of India. Present this data in a tabular form.

Solution

The total outlay of the Seventh Five Year Plan period is given item-wise. For the outlay of 1986-87, the outlay on rehabilitation of bonded labour is not given. It is a balancing figure. The total outlay for the year 1986-87 was Rs. 1,847 lakhs and the outlay on all the other heads together was Rs. 1,327 lakhs (i.e., Rs. 465 lakhs on training, Rs. 107 lakhs on employment services, and Rs. 755 lakhs on labour welfare). Now the outlay on rehabilitation of bonded labour was Rs. 520 lakhs (balancing figure). It is said that the total outlay for 1987-88 increased by Rs. 49 lakhs as compared to the outlay of 1986-87. This means the outlay for 1987-88 is Rs. 1,896 lakhs. The outlay on employment services decreased by Rs. 11 lakhs, meaning thereby that it came down to Rs. 96 lakhs from Rs. 107 lakhs. The outlay on labour welfare decreased by Rs. 148 lakhs from Rs. 755 lakhs, which means that the amount came down from Rs. 755 lakhs to Rs. 607 lakhs. On rehabilitation of bonded labour the outlay decreased by Rs. 327 lakhs which means that it came down to Rs. 193 lakhs. This implies that the outlay on training in 1987-88 was Rs. 1,000 lakhs (balancing figure). This data can be shown in tabular form as follows :

Central Government Outlay on Labour and Labour Welfare Sector

(Rs. in Lakhs)

Head of Expenditure	Seventh Five Year Plan	1986-87	1987-88
1) Training	4,184	465	1,000
2) Employment Services	546	107	96
3) Labour Welfare	3,180	755	607
4) Rehabilitation of Bonded Labour	1,600	520	193
Total	9,510	1,847	1,896

Note: The outlay does not include coaching-cum-guidance scheme for scheduled castes and scheduled tribes.
Source: Annual Plan 1987-88, Planning Commission, Government of India, page 335.

Check Your Progress B

1) A statistical table is presented below. Identify the major parts on it.

Table 7.4

Kerala Soaps & Oils Limited
Production, Sales and Net Profits (-) Net Loss 1980-81 to 1983-84

(Rs. in Lakhs)

Year	Production		Sales		Net profits/ (-) Net loss
	Quantity (Mts.)	Value	Quantity (Mts.)	Value	
(1)	(2)	(3)	(4)	(5)	(6)
1980-81	8319	862.00	7439	747.36	11.59
1981-82	8659	925.48	9125	897.03	09.62
1982-83	9212	1016.93	9141	952.57	(-) 50.60
1983-84	6013	793.00	6094	757.61	(-) 190.12
1984-85	N.A.	939.00	N.A.	882.80	(-) 112.92

Compiled from the Annual Reports & Accounts, 1980-81 to 1983-84, Kerala Soaps & Oils Limited, Calicut, Kerala

2) Distinguish between the title and the headnote.

.....

.....

.....

.....

3) Distinguish between a caption and a stub.

.....

.....

.....

.....

.....

4) Distinguish between a footnote and a reference note.

.....

.....

.....

.....

5) What is the body of the table?

.....

.....

.....

.....

6) State whether the following statements are True or False.

i) There should be a clear, precise and self-explanatory title on the top of every statistical table.

ii) Headnote is to be written just below the body of the table.

iii) The headnote should indicate the unit in which the data has been given.

iv). Caption is also called box head.

v) The body of the table contains the quantitative information to be presented.

vi) A reference note tries to clarify the ambiguities, if any, about the data shown in the table.

vii) A footnote is given to show the sources of data.

viii) The stubs and captions of a table should be brief and clear.

ix) Ruling is not very important in a statistical table.

x) Use of abbreviations, as far as possible, should be avoided in a statistical table.

7) Fill in the blanks with the appropriate words given in the brackets.

i) The title of a statistical table should be as ..... as possible. (large/small)

ii) The headnote should be written just below the title of the table, preferably on the ..... hand corner. (right/left)

iii) The title given to rows of a table is called a ..... . (stub/caption)

iv) Caption labels the data to be given in the ..... (columns/rows)

v) It is ..... to give table number below every table. (essential/ not essential)

vi) The captions ..... be numbered. (may/may not)

vii) The headnote relates to ..... of the data. (unit/characteristics)

viii) While constructing a statistical table, ditto mark should be ..... . (used/avoided)

7.7 LET US SUM UP

Tabular presentation is a technique of arranging data in an orderly manner in rows and columns. Horizontal arrangement of data is called rows and the vertical arrangement is known as columns. Tabular ,presentation facilitates comparison of data, helps in identifying the desired values, provides a basis for analysis and exhibits the trend of data. Some people treat classification and tabulation as synonyms. In fact, it is not so. Classification divides the data on the basis of similarities and resemblances, whereas tabulation is the process of recording classified facts in rows and columns.

Statistical tables can either be information tables or general purpose tables or special purpose tables. Information tables, on the basis of the number of characteristics shown, can be either simple tables or complex tables. Simple tables show only one characteristic whereas the complex tables show more than one characteristics of data. Complex tables could be two fold tables, three fold tables or mani-fold tables. General purpose tables are the store house of information and are generally found as appendices to various reports. Special purpose tables are helpful in statistical analysis and indicate rates, percentages, averages, etc.

A statistical table has several parts known as title, headnote, stub, caption, body, foonote, reference note, totals, table number, etc. A good statistical table must present data in a clear

and simple manner, should have a brief and clear title, brief and clear stubs and captions, precise and complete headnote, a clear reference note, and totals and sub-totals at appropriate places. It should show derived data wherever necessary and avoid use of abbreviations and "ditto" mark. It should be provided with proper ruling, wherever necessary.

7.8 KEY WORDS

Caption : Tables the data in the columns of the table.,

Complex Table : A statistical table prepared on the basis of more than one characteristics of the .data.

Column : Vertical arrangement of data in a statistical table.

Field (Body) : The main part of the table where the quantitative information is presented.

Footnote : A note presented just under the main part of the table clarifying the ambiguities of data, if any.

General Purpose Table : A statistical table containing a wide range of information on specific subject and found generally as appendices to the reports

Head Note : A note presented just below the title, preferably on the right hand corner, indicating the unit in which the data is presented.

Reference Note : A note presented beneath the main part of the table disclosing the source from which the data is collected. It is also known as source note.

Row : Horizontal arrangement of data in a statistical table.

Simple Table : A statistical table prepared on the basis of one characteristic of the data. It is also known as one-way table.

Special Purpose **Table** : A statistical table which is helpful in analysis of data and indicates derived statistics.

Stub : Title given to a row in a statistical table.

Title : Note written on the top of a statistical table indicating the characteristics of the data presented in that table.

7.9 ANSWERS TO CHECK YOUR PROGRESS

A) 7) i) False ii) True iii) True iv) True v) False vi) True vii) True viii) False ix) False x) True xi) True xii) True.

8) i) horizontal ii) tabulation iii) simple iv) analysis v) different vi) condensed vii) more than one viii) appendices to the reports ix) analysis x) complementary

B) 1)

Table 7.4
Kerala Soaps & Oils Limited
Production. Sales and Net Profits/(–) Net Loss 1980-81 to 1983-84.

(Rs. in Lakhs)

Year	Production		Sales		Net pro&/ (-) Net loss
	Quantity (Mts.)	Value	Quantity (Mts.)	Value	
(1)	(2)	(3)	(4)	(5)	(6)
1980-81	8319	862.00	7439	747.36	11.59
1981-82	8659	925.48	9125	897.03	09.62
1982-83	9212	1016.93	9141	952.57	(-) 50.60
1983-84	6013	793.00	6094	757.61	(-) 190.12
1984-85	N.A.	939.00	N.A.	882.80	(-) 112.92

compiled from the Annual Reports & Accounts, 1980-81 to 1983-84, Kerala Soaps & Oils Limited, Calicut, Kerala.

6) i) True ii) False iii) True iv) True v) True vi) False vii) False viii) True ix) False x) x) True

7) i) small ii) right iii) stub iv) columns v) essential vi) may vii) unit viii) avoided

7.10 TERMINAL QUESTIONS/EXERCISES

Questions

1) What is tabulation? What are the objectives of statistical tables?

2) Draw the format of a statistical table and indicate its various parts.

3) Distinguish between simple andcomplex statistical tables and give examples of the two types of tables.

4) Describe the requisites of a good statistical table?

Exercises

1) Prepare a blank statistical table to show the age, sex and literacy levels of the residents of a town.

2) In an organisation there are 1,000 employee. in which 40%are ladies. Of the total, 30%of the employees are smokers and the number of smokers among ladies is 10.

3) The following figures relate to the number of crimes (nearest-hundred)in four metropolitan cities of India. In 1961, Bombay recorded the highest number of crimes i.e. 19,400 followed by Calcutta with 14,200, Delhi 10,000, and Madras 5,700. In the year 1971, there was an increase of 5,700 in Bombay over its 1961 figure.The corresponding increase was 6,400 in Delhi and 1,500 in Madras. However, the number of these crimes fell down to 10,900 in case of Calcutta for the corresponding period. In 1981, Bombay recorded a total of 36,300 crimes. In that year, the number of crimes was 7,000 less in Delhi as compared to Bombay. In Calcutta the number of crimes increased by 3,100 in 1981 as compared to 1971. In the case of Madras the increase in crimes was by 8,500 in 1981 as compared to 1971. Present this data in tabular form.

Note : These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.

UNIT 8 DIAGRAMMATIC PRESENTATION

Structure

8.0 OBJECTIVES

After studying this unit, you should be able to :

- state the importance of visual presentation
- describe the usefulness of diagrams for the presentation of data
- explain the principles of constructing diagrams
  identify the different types of diagrams, and
- prepare the different types of diagrams.'

8.1 INTRODUCTION

In Unit 7 you have studied that presentation of data in tabular form facilitates comparison as it presents mass of data in simple and orderly manner. It is easier to establish trend and patterns when the data is in tabular form. Besides presenting in tabular form, the data can also be presented in the form of diagrams and graphs. The presentation of data in the form of diagrams and graphs is also called visual presentation of data. Compared to tabular presentation, data presented in diagrams and figures is more impressive and it is easier to draw conclusions. In this unit, you will study the importance, principles and different types of diagrammatic presentation of data.

8.2 IMPORTANCE OF VISUAL PRESENTATION OF DATA

Visual presentation of data means presentation of data in the form of diagrams, curves and straight line. Visual presentation of data is desirable for following reasons :

1) Visual presentation of data eliminates the dullness of the numerical data. From a large mass of numerical data it is often difficult to draw any conclusion. Besides, it also causes undue strain on the mind. The data when presented in the form of diagrams and graphs, 'creates interest and leaves an impression on the mind of the reader for a longer period.

2) Comparison of data is much easier if it is presented in the form of diagrams and graphs. In several cases, careful glance at the diagram or graph renders the comparison of the complex data much easier.

3) The location of various statistical measures is possible with the help of graphs. Several measures of central value such as Median, Quartiles, Mode, etc., can be located with the help of graphs (ogives and histogram).

4) The trends of the past performance can be established with the help of graphs. Presentation of such trends on graphs helps in forecasting.

5) Diagrams and graphs have become an integral part of the advertisement campaign of several business firms. An advertisement, without visual effect looks incomplete.

8.3 PRINCIPLES OF PREPARING DIAGRAMS

You have studied the importance of diagrammatic presentation of data. Let us now discuss the guidelines to be followed while preparing diagrams. The following guidelines should be kept in mind while preparing diagrams :

1) A diagram is to be prepared on the graphic axes—'X' axis and 'Y' axis. However, it is not necessary to use a graph paper. While taking scales on these two axes, it must be ensured that the data is being presented in a meaningful manner. The scale on the two axes should be clearly set up.

2) Whenever the data are to be presented on the 'Y' axis (vertical scale), the scale should start from zero. Generally, the vertical scale is not broken.

3) A diagram must always have a concise and self-explanatory title.

4) Colours and shades should be used to exhibit various components of a diagram and a key be provided.

5) To make the diagram attractive, leave reasonable margin on all sides of the diagram. The diagram should not be too small or too big.

6) If a number of diagrams are to be prepared, it is desirable to number them for the purpose of reference.

8.4 TYPES OF DIAGRAMS

Diagrams are generally classified on the basis of length, breadth and height. Broadly, diagrams are classified as : 1) one dimensional diagram, 2) two dimensional diagram, and 3) three dimensional diagram. Besides these diagrams, the data can also be presented in the form of maps and pictographs. However, in this unit, we discuss the one dimensional and two dimensional diagrams only.

The one dimensional diagrams can be further classified as : I) simple bar diagrams, 2) multiple bar diagrams, 3) sub-divided bar diagrams, and 4) percentage sub-divided bar diagrams. Similarly, two dimensional diagrams also can be classified as : I) rectangles, 2) sub-divided rectangles, 3) squares and circles, and 4) pie diagrams. Now let us study about all these types in detail.

8.5 ONE DIMENSIONAL DIAGRAMS

One dimensional diagrams are prepared only on the basis of one dimension i.e., length. The other two dimensions have no significance in this type of diagrams. This type of diagrams take the shape of bars or column charts. Bars or column charts enable the magnitude to be compared visually. The length of the various bars is proportionate to the magnitude of the given data. However, their thickness is not related to the magnitude of the data. Their thickness is only to make the diagram attractive. For example, the production figures of a business concern are 10,000 units and 20,000 units for the years 1988 and 1989 respectively. If we draw a bar diagram for this data, the length of the two bars should be in the ratio of 1: 2 and normally, the thickness would be the same. One dimensional diagrams can be further classified as discussed below.

8.5.1 Simple Bar Diagrams

In a simple bar diagram, one bar is prepared to represent one given value. The length of various bars is in the ratio of the magnitude of the given data. As the width of the bars is not significant, width is uniform for all bars. Though the bars may look like rectangles, but they are not rectangles as they represent only the length. The gap may be left in between the different bars. The gap between the bars is normally identical. Simple bars can be prepared either vertically or horizontally. Both positive and negative values can be presented through this simple bar diagram.

This simple bar diagram is generally prepared when the data indicates different values of a variable over a time period or when the data represents different situations. It is very easy to prepare the simple bar diagrams. Study Illustrations 1 and 2 carefully and learn the presentation of simple bar diagrams.

Tllustration 1

Prepare a simple bar diagram for the following data relating to the value added per employee in Maruti Udyog Ltd.

Year	1983-84	1984-85	1985-86	1986-87	1987-88	1988-89
Value Added Per Employee (Rs. in '000) :	153	192	234	396	474	498

Solution

The value added per employee (Rs in '000) is given for different years. One bar will be prepared for every year in the ratio of the magnitude of the given data. Identical gap will be left in between the different bars. The bars have been prepared vertically by showing the years on horizontal axis and the value added per employee on vertical axis. Now study Diagram 8.1 carefully.

Diagram 8.1 : Simple Bar Diagram Showing the Value Added per Employee in Maruti Udyog Limited

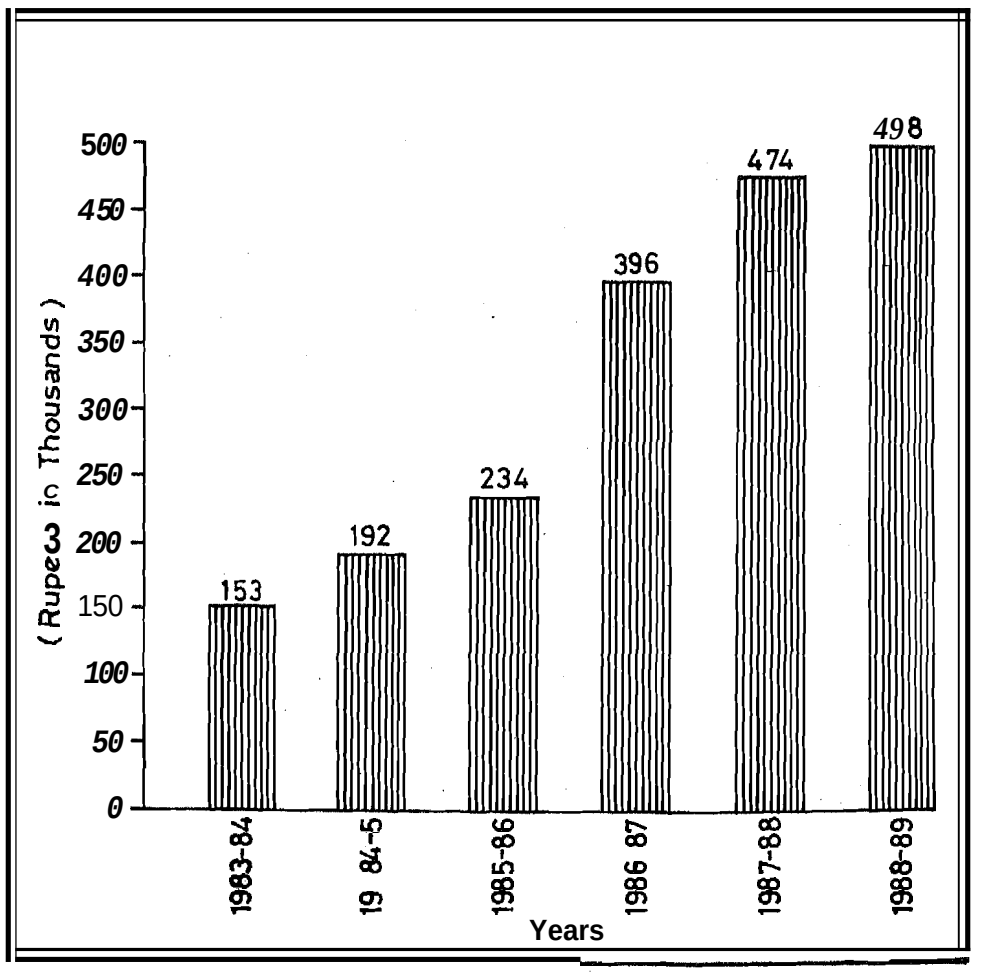
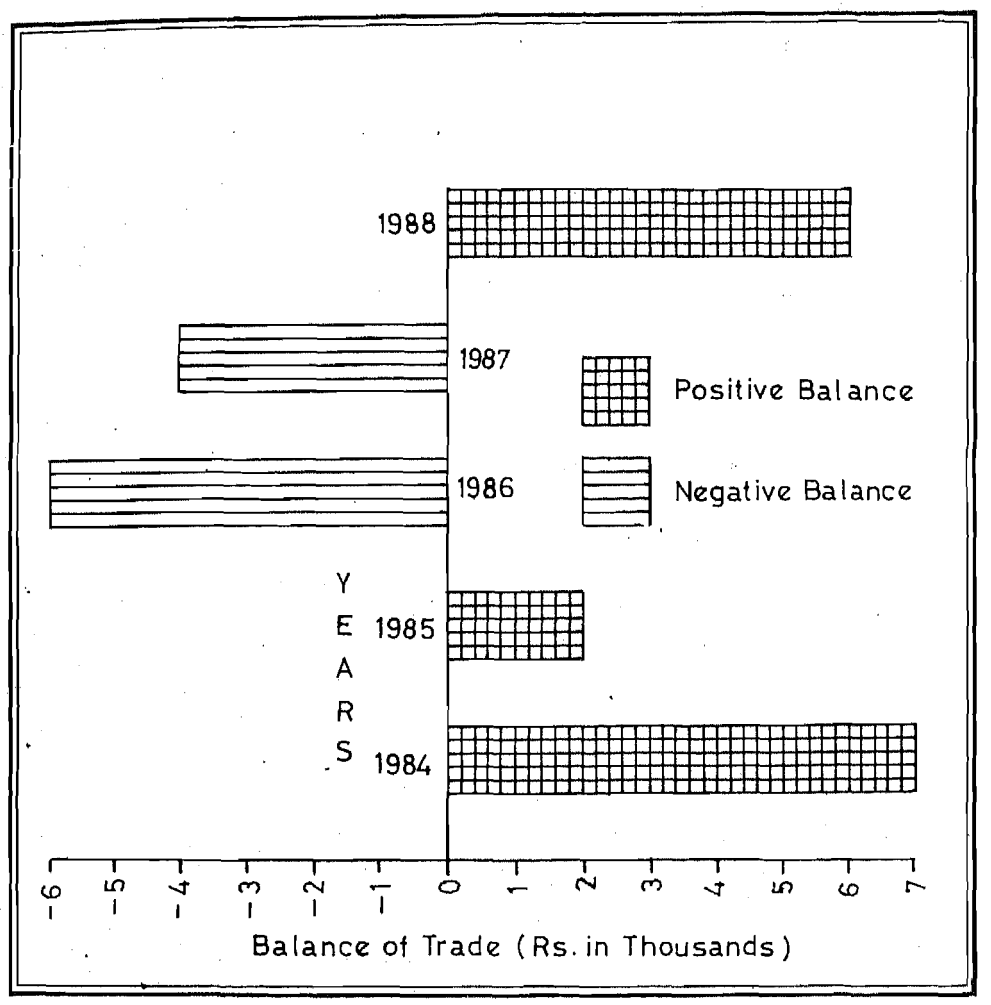


Illustration 2
Present a bar diagram for the following data relating to the balance of trade of a country.

Year	1984	1985	1986	1987	1988
Balance of Trade : (Rs. in '000)	7	2	—6	—4	6

Solution
The given data represents positive as well as negative values. This data can be represented either vertically or horizontally. If the bars are constructed horizontally, the positive values are taken on the right hand side of the vertical axis and the negative values on its left side. Similarly, when the bars are constructed vertically, the positive values are taken on the upper side of horizontal axis while the negative values are taken on the lower side. Since we have shown vertical presentation in the previous illustration, now this diagram is presented horizontally. Study Diagram 8.2 carefully and understand how the bars are drawn.

Diagram 8.2 : Simple Bar Diagram Showing the Positive and Negntive Values



8.5.2 Multiple Bar Diagrams

In this type of diagram two or more bars are constructed adjoining each other. These bars either represent different variables or various components of the same variable. The bars for a given set 'K' are constructed adjoining each other and identical gap is left in between the bars of differentsets. Just like simple bar diagrams, the length of the various bars varies in the ratio of the magnitude of the given-values. The width of the different bars is identical. This diagram, on the one hand, facilitates comparison of the values of different variables in a set and on the other it facilitates the comparison of the values of the same variable over a period of time. To facilitate easy comparison, the different bars of a set may be coloured or shaded differently. But the colour or shade for the bars representing the same variable or a component in different sets should be the same. Now let us take up Illustration **3** and learn about the preparation of multiple bar diagrams.

Illustration 3

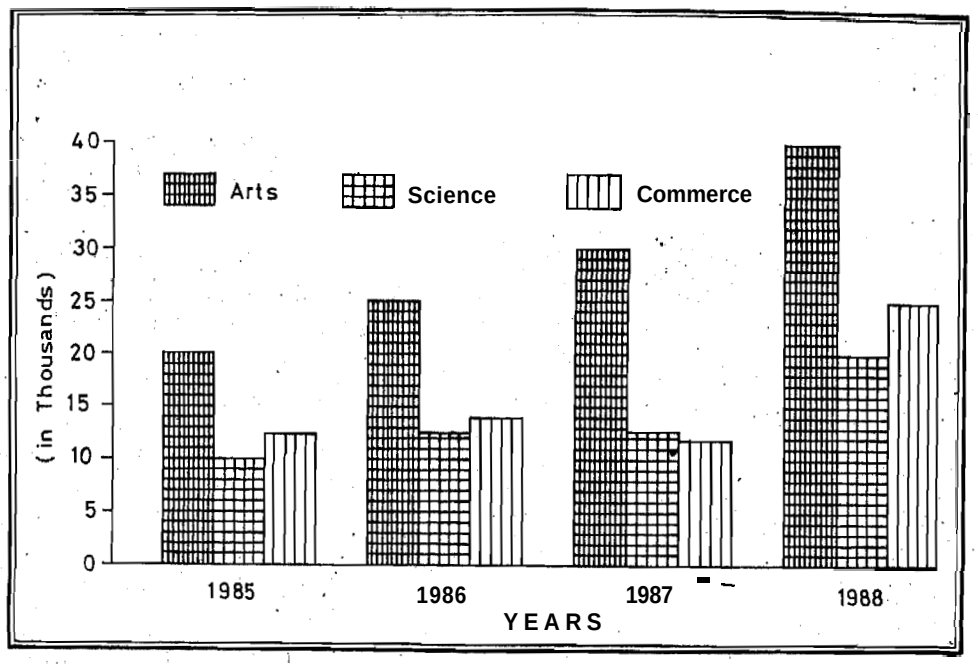
The following data relate to the enrolment of students in a university in different faculties. Present this data as a multiple bar diagram.

Year	Arts	Science	Commerce
1985	20	10	12
1986	25	12	14
1987	30	15	14
1988	40	20	25

Solution

The data relates to the enrolment of students in a university in the faculties of arts, science and commerce. This data relates to four years—1985 to 1988. Therefore, four sets of bars should be drawn, each set representing one year. Within each set there should be three bars representing three faculties. Within a set all the three bars should be constructed adjoining each other and their width should be identical. Identical gap should be left in between all the four sets. Study Diagram 8.3 carefully and understand how the diagram is prepared.

Diagram 8.3 ; Multiple Bar Diagram Showing the Enrolment of Students in Different Faculties in a University



8.5.3 Sub-divided Bar Diagrams

This type of bar diagram is prepared to represent the different components of the same variable. It is also called component diagram. In this diagram one bar is constructed for the total value of the variable, and then the bar is sub-divided in proportion to the values of the various components of that variable. In fact, the values of the different components are cumulated for constructing this bar diagram and the bar is to be sub-divided at these cumulated points. Study Illustration 4 carefully to understand the method of preparing subdivided bar diagram.

Illustration 4

The following data relate to the number of students admitted to first year class in different courses in a college. Show this data by means of a sub-divided bar diagram

Year	Arts	Commerce	Science
1986	300	200	100
1987	250	250	200
1988	250	300	200

Solution

For constructing a sub-divided bar diagram, first of all, we have to identify the points where the bars for different years are to be sub-divided. For this purpose, we should calculate the cumulative values. The length of the bar for any specific year should be in the ratio of the total

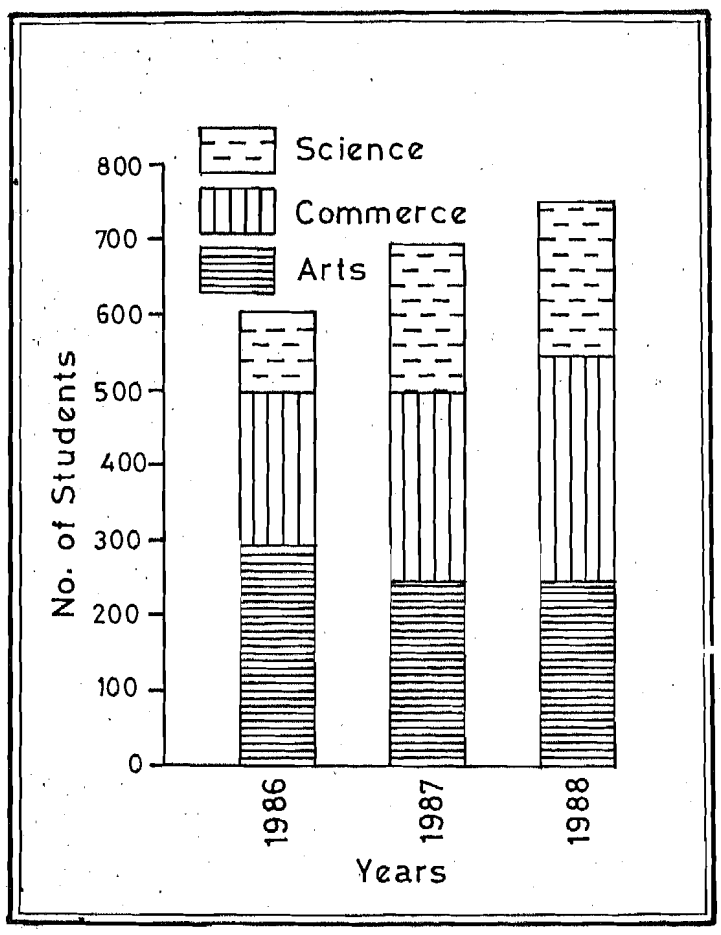
number of students admitted to the college in that year. In the present case, every bar should have three components viz., arts, commerce and science. Now let us first compute the cumulative figures.

Number of Students Admitted to First Year In Different Courses

Courses	1986		1987		1988	
	Number	Cumulative Number	Number	Cumulative Number	Number	Cumulative Number
Arts	300	300	250	250	250	250
Commerce	200	500	250	250	300	550
Science	100	600	200	700	200	750
Total	600		700		750	

From the cumulative figures computed in the above table, we know the points where the bars are to be sub-divided. For example, take the case of 1986 cumulative figures. The bar should be sub-divided at 300 and 500. Now study Diagram 8.4 carefully and understand how the sub-divided bar diagram is constructed for the above data.

Diagram 8.4 : Sub-divided Bar Diagram Showing the Students Adrnittcd to First Year in Different Courses in a College.



8.5.4 Percentage Sub-divided Bar Diagrams

A sub-divided bar diagram can be prepared on the basis of percentage figures also. In this type of diagram, the length of a bar is taken as 100 and the length of each component is

represented by the percentage share of that in the total. The length of different bars in this diagram will be the same. It is supposed to represent the relative changes and not the absolute changes in the values of the different components.

This type of bar diagram is also constructed in the same manner as the simple sub-divided bar diagram. Since each component is expressed as a percentage in the total, first of all these percentages are calculated. Then these percentages are cumulated and a bar is sub-divided at these cumulated points. Study Illustration 5 carefully and understand the method of preparing the percentage sub-divided bar diagram.

Illustration 5

The following data relate to the number of students admitted to the first year class of a college in different courses in the year 1987-88. Represent this data as a percentage sub-divided bar diagram.

Courses	1987	1988
Arts	200	200
Science	40	100
Commerce	160	200
Total	400	500

Solution

For constructing a percentage sub-divided bar diagram, first we have to calculate the percentages of different components to the total. Then these percentages should be cumulated and the bars should be sub-divided at these cumulative percentage points. The length of the two bars will be taken as 100. In the present illustration. every bar will have three components viz., arts, science and commerce.

Students Admitted to First Year in Different Disciplines

Courses	1987			1988		
	Number	%	Cumulative %	Number	%	Cumulative %
Arts	200	50	50	200	40	40
Science	40	10	60	100	20	60
Commerce	160	40	100	200	40	100
Total	400	100		500	100	

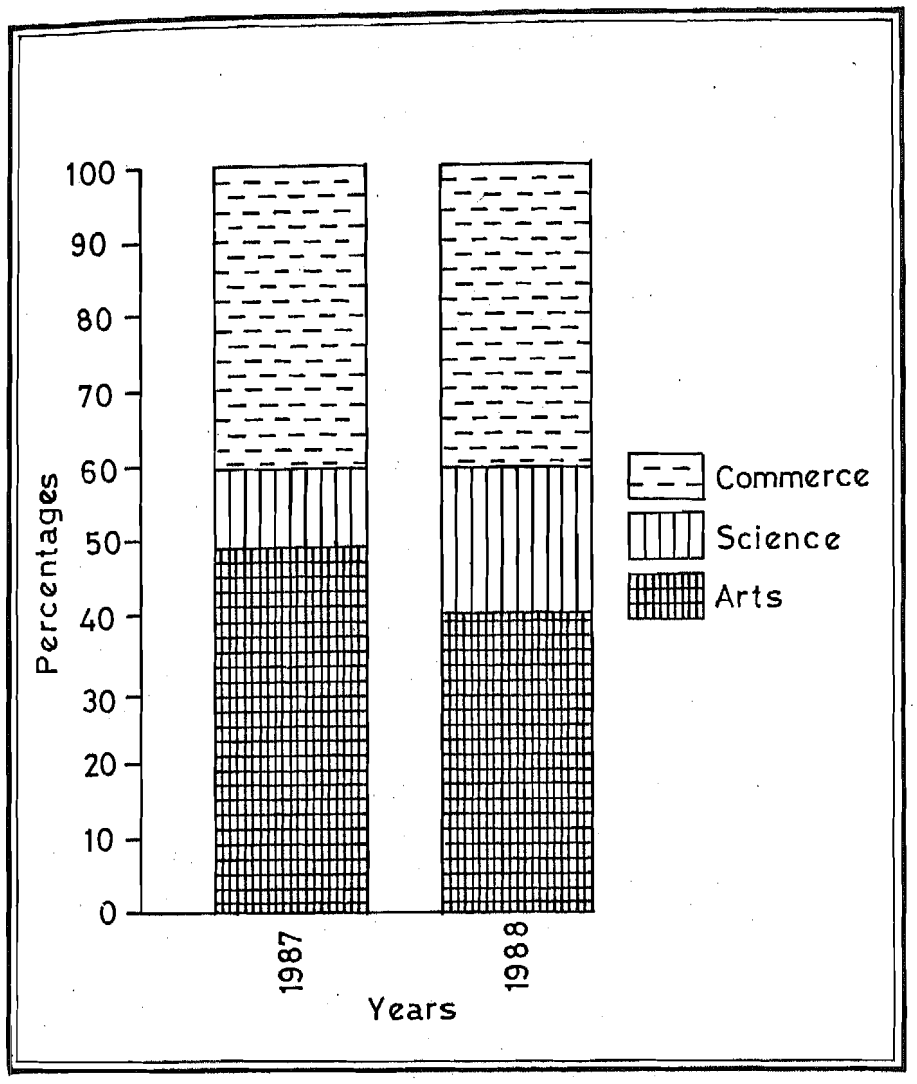
From the cumulative percentage figures computed in the above table, we know the points where the bars are to be sub-divided. For instance, take the cumulative percentage figures for 1987. In this case, you have to draw the bar for the figure 100 with sub-divisions at 50 and 60. Look at Diagram 8.5 carefully and study how the bars are drawn.

Illustration 6

The following data relate to the cost of production and selling price of a TV Cabinet. Prepare percentage sub-divided bar diagram for this data.

Item	Cost and Selling Price per Cabinet (in Rs.) 1988
Raw Material	500
Wages	200
Polishing	100
Selling Price	1000
Profit (+) or Loss (–)	+200
	660
	330
	110
	1000
	–100

Diagram 8.5 : Percentage Sub-divided Bar Diagram **Showing the** Students Admitted **to** First **Year** in **Different** Courses in a College.



Solution

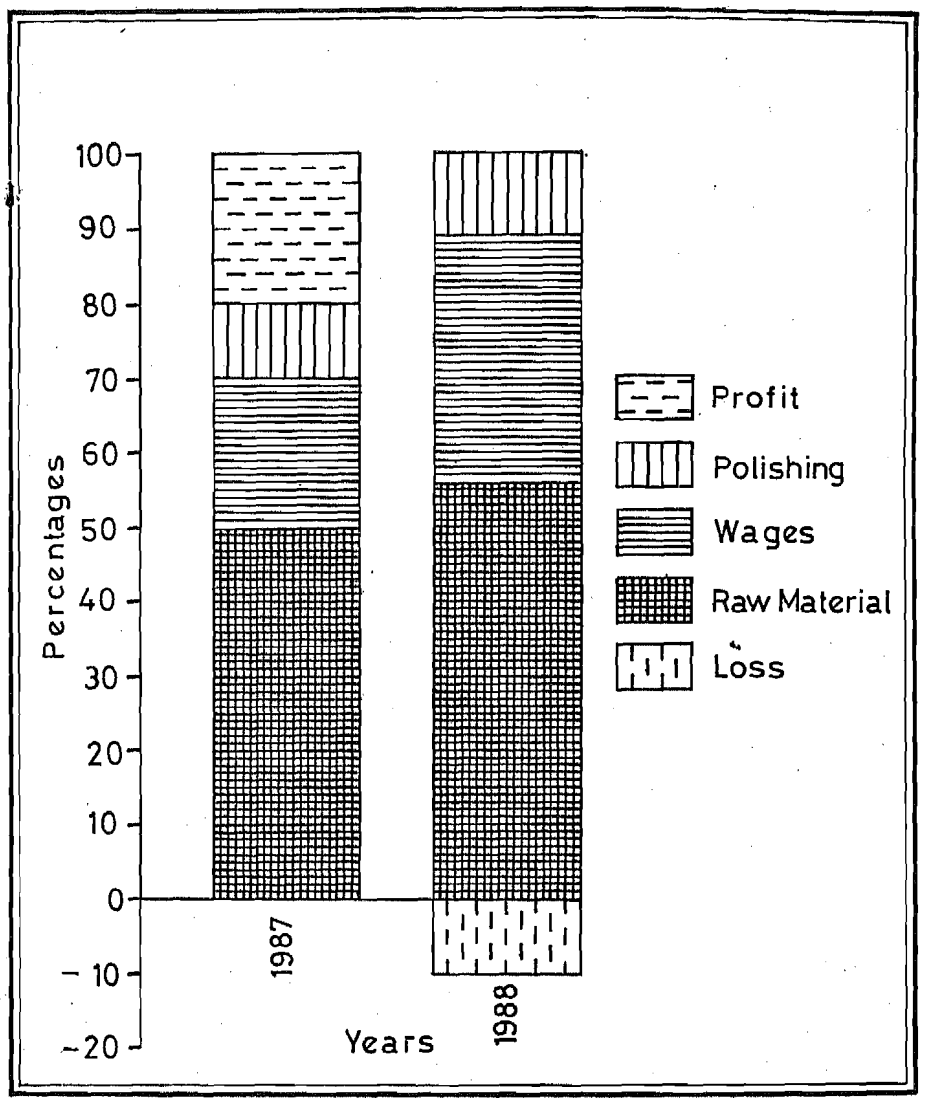
In this problem one factor is showing both positive and negative values. For constructing **percentage** sub-divided bar diagram in such cases, first of all it is necessary to decide which of the variable is to be taken as 100%so that positive and negative percentages may properly be represented. In the present illustration selling price should be taken as 100% and all the other factorsshould be expressed as percentage to selling price. So, first **we** have to compute the percentages of different components to selling price. Then these percentages will be cumulated and the bars will be sub-divided at these cumulative percentage points,

Cost and Selling Price per TV Cabinet

Items	1987			1988		
	Amount **Rs.**	%	Cumulative %	Amount Rs.	%	Cumulative %
Raw material	500	50	50	660	66	66
Wages	200	20	70	330	33	99
Polishing	100	10	80	110	11	110
Profit(+)or **Loss** (-) per Cabinet	200	20	100	100	10	100
Selling Price (Total)	1000	100		1000	100	

The bar for 1987 (when there is profit) will start from zero line on X-axis. Since there is loss in 1988, the position of this loss should be shown below X-axis. For the year 1988 marking will start from below the X-axis. So in the bar for 1988 the first portion representing raw materials (66%)will be shown 10%below X-axis indicating loss and the remaining 56%(i.e., 66%– 10%)above the X-axis. Now wages portion will start from 56%and go upto 89%(i.e., 99%– 10%).Similarly, other parts will be marked. Now the percentage sub-divided bar diagram will look like the same as presented in Diagram 8.6.

Diagram 8.6 : Percentage Sub-divided Bar Diagram Showing the Cost and Selling Price per a TV Cabinet.



Check Your Progress A

1) What is the difference between tabular presentation and visual presentation of data?

.....

.....

.....

.....

2) Differentiate between one dimensional and two dimensional diagrams.

.....

.....

.....

.....

3) What is a simple bar diagram?

.....

.....

.....

.....

.....

4) State the circumstances under which a multiple bar diagram is prepared.

.....

.....

.....

.....

.....

5) What is the distinction between multiple bar diagram and sub-divided bar diagram?

.....

.....

.....

.....

.....

6) State whether the following statements are True or False.

i) It is easier to draw conclusions from diagrams as compared to statistical tables.

ii) Visual presentation of data makes them interesting and creative.

iii) No statistical measures can be located with the help of graphs.

iv) The trends of past data can be established with the help of visual presentation of data.

v) While preparing a diagram, it must be ensured that the data is presented in a meaningful manner.

vi) A key should always be provided in a diagram.

vii) One dimensional diagrams are prepared only on the basis of length.

viii) In a multiple bar diagram, the length of various bars is proportionate to the magnitude of the given data.

ix) There is no difference between a bar and a rectangle.

x) Normally, identical gap should be left in between the different bars.

xi) In case of multiple bar diagrams, the length of the various bars varies in the ratio of magnitude of the given values.

xii) A subdivided bar diagram is prepared to show the different components of the same given variable.

7) Fill in the blanks with the appropriate words given in the brackets.

i) Visual presentation of data. . . . . .the dullness of numerical data. (eliminates/increases)

ii) A careful glance at diagrams renders the comparison of complex data . . . . . (more difficult/easier)

iii) A simple bar diagram is prepared to represent . . . . . given value(s). (one/two)

iv) Simple bars . . . . . represent negative values. (can/cannot)

v) In multiple bar diagrams, the width of different bars is kept. . . . . (different/identical)

vi) In a sub-divided bar diagram, each bar is sub-divided on the basis of . . . . . values. (cumulative/actual)

vii) A sub-divided bar diagram . . . . . be prepared on percentage basis. (can/cannot)

viii) Negative values . . . . . be shown in the sub-divided bar diagram. (can/cannot)

8.6 TWO DIMENSIONAL DIAGRAMS

As you know, the diagrams may be classified as : 1) one dimensional diagrams, 2) two dimensional diagrams, and 3) three dimensional diagrams. We have already discussed in detail about one dimensional diagrams. Now let us discuss about the two dimensional diagrams.

As you know, there are three dimensions i.e., length, width and height, on the basis of which the diagrams are constructed. One dimensional diagrams are prepared only on the basis of length, and the width is not at all significant. The length of the bars is proportionate to the magnitude of the given data. The two dimensional diagrams. however, are prepared on the basis of two dimensions i.e., length and width. As the product of length and width indicates the area, this type of diagram is also called Area Diagram.

The two dimensional diagrams may be classified as : 1) rectangles, 2) sub-divided rectangles, 3) squares and circles, and 4) pie diagrams. Now let us discuss each of these categories in detail.

8.6.1 Rectangles

A rectangle is prepared on the basis of length and width and indicates the area. Before constructing a rectangle, we have to identify the variable to be represented by the length and the variable to be represented by the width of the rectangle. For example, a manufacturing enterprise has produced 10,000 units of product X during a particular month and the cost of production per unit is Rs. 20. For preparing the rectangle, the number of units produced can be considered to represent length and the cost of production per unit to represent width of the rectangle. Since a rectangle indicates area, in this case rectangle will show the total cost of production (i.e. 10,000 × Rs. 20 = Rs. 2,00,000). Now, let us take up an illustration and learn the preparation of rectangles practically.

Illustration 7

The following data relate to the number of units produced and cost per unit of a manufacturing enterprise during first two months of 1989. Prepare a rectangle for this data.

	January	February
Number of Units Produced	1,00,000	80,000
Cost per Unit	Rs. 8	Rs. 10

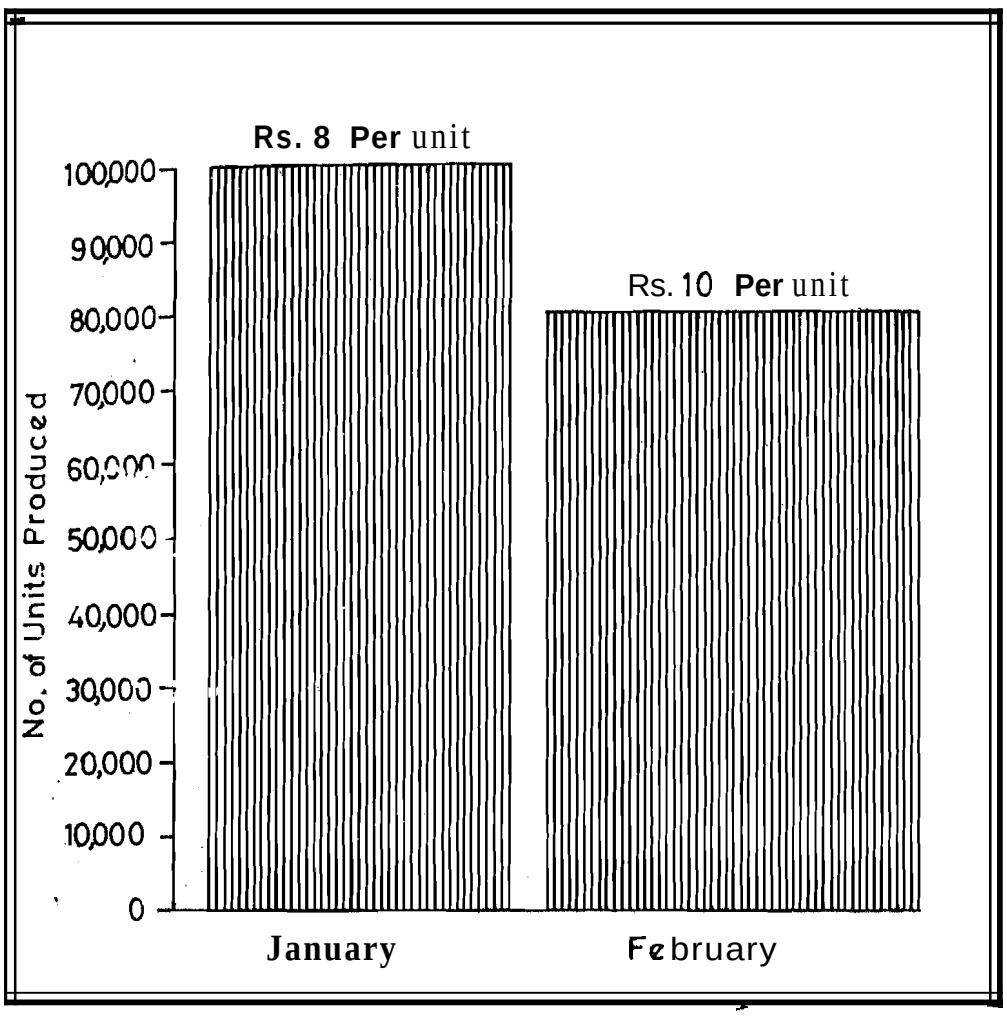
Solution

For this data, we have to prepare two rectangles on the following basis :

	Rectangle-I	Rectangle-II
Length (Units Produced)	1,00,000 units	80,000 units
Width (Cost per Unit)	Rs. 8	Rs 10

The length of the two rectangles will be in the ratio of 10:8 and the width in the ratio of 8:10. Now look at Diagram 8.7, the two rectangles have been presented. However, you should note that the area in case of both the rectangles is the same.

Diagram 8.7 : Rectangles Showing the Cost of Production.



8.6.2 sub-divided Rectangles

A rectangle can also besub-divided. The sub-divided rectangles will also show the area in respect of different components. For sub-dividing a rectangle, the cumulative values of the various components are calculated and the rectangle is sub-divided according to these cumulative values. Let us take some illustrations and prepare sub-divided rectangles.

Illustration 8

The following data relates to the number of units produced by a manufacturing concern and the cost per unit on various items. Prepare a sub-divided rectangle for this data.

i)	**No. of units produced**	**Rs.** 5,000
ii)	**Cost of Raw Material**	**Rs. 30,000**
iii)	**Wages**	**Rs. 15,000**
iv)	**Other costs**	Rs. **5,000**

Solution

First, we have to determine the length and width of the rectangle. The length may be represented by the cost per unit, which is the total cost divided by units produced i.e., Rs. 50,000 ÷ 5,000 = Rs. 10. The width may be represented by the number of units produced i.e. 5,000 units. Since a subdivided rectangle is to be prepared, we have to calculate per unit cost for various heads of expenditure and then calculate their cumulative values. Look at the following table where these cumulative values have been computed and presented.

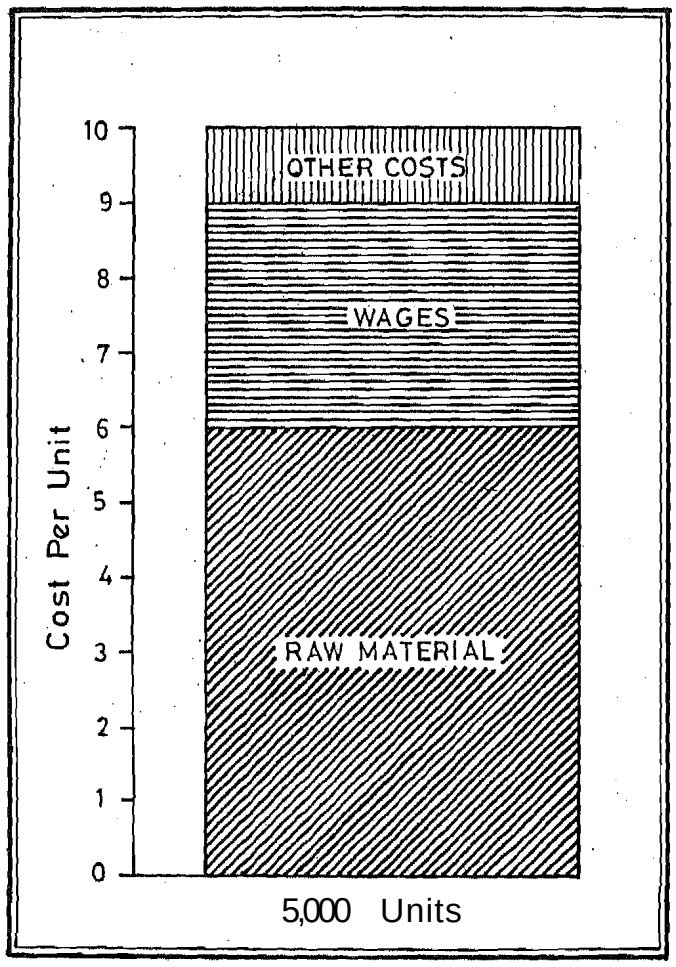
Computation of Cumulative Values

(Units produced = 5,000)

Items	Total Cost Rs.	Cost per Unit Rs.	Cumulative Cost per unit Rs.
1 Cost of Raw Material	30,000	6	6
2. Wages	15,000	3	9
3 Other Costs	5.000	1	10
Total	**50,000**	**10**	

As we have computed the cumulative values,,now we can proceed to construct the sub-divided rectangle. Look at Diagram 8.8 carefully and study how the sub-divided rectangle is prepared.

Diagram 8.8 : Sub-divided Rectangle Showing the Item-Wise Cost of Production



8.6.3 Squares and Circles

Squares and circles are also two dimensional diagrams as they also represent the area. All sides of the square are the same, hence the length and the width in case of square are the same. While determining any one side of the square, we calculate the square root of the given data and then an appropriate scale is taken to draw the squares. *If* more than one square is drawn in a single diagram, the bases of all the squares lie on the same line.

A circle also represents the area, which is πr^2. For constructing a circle, we determine the radius of that circle. We calculate the square root of the given data to determine the ratio of various circles. When we draw more than one circle, the central points of various circles lie on

the same line. You should remember that by calculating the square roots of the given data, the large values are considerably reduced. Usually a square diagram is used when different values of one variable, without any sub-divisions, are to be shown. Now let us learn practically the preparation of squares and circles.

Illustration 9

The following data relates to the annual plan outlay during **1987-88** on different heads of development. Draw squares and circles for this data.

	(Amount) *(Rs. in Crores)*
1) Agriculture and Allied Activities	2,378
2) Energy	12,999
3) Industry and Minerals	5,635

Solution

First, we have to compute the square roots for the given data. Then we have to adopt appropriate scale to determine the sides of the squares and the radius of the circles.

Calculation of Square. Roots

Head of Development	Amount (Rs.)	Square Root	Side of the Square or Radius of the Circle (in Cms.)
1. Energy	12,999	114.01	4.68
2. Industry & Minerals	6,243	79.01	3 24
3. Agriculture and Allied Activities	2,378	48.76	2.00

The minimum square root value has been taken as the basis for determining the side of a square or the radius of a circle. In the above case, the square root value **48.76** has been taken as equal to 2.00 cms. and the other sides/radius have been determined in proportion to the different square root values. Now squares and circles have been drawn as shown .in Diagrams 8.9A and 8.9B

Diagram 8.9A : Squares Showing Annual Plan Outlay for Various Sectors During 1987-88

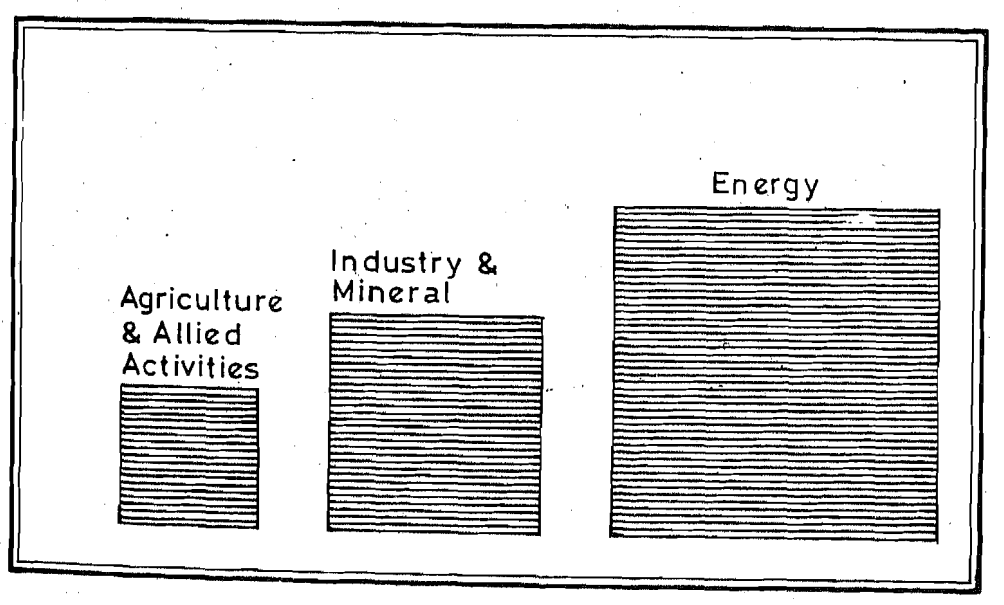
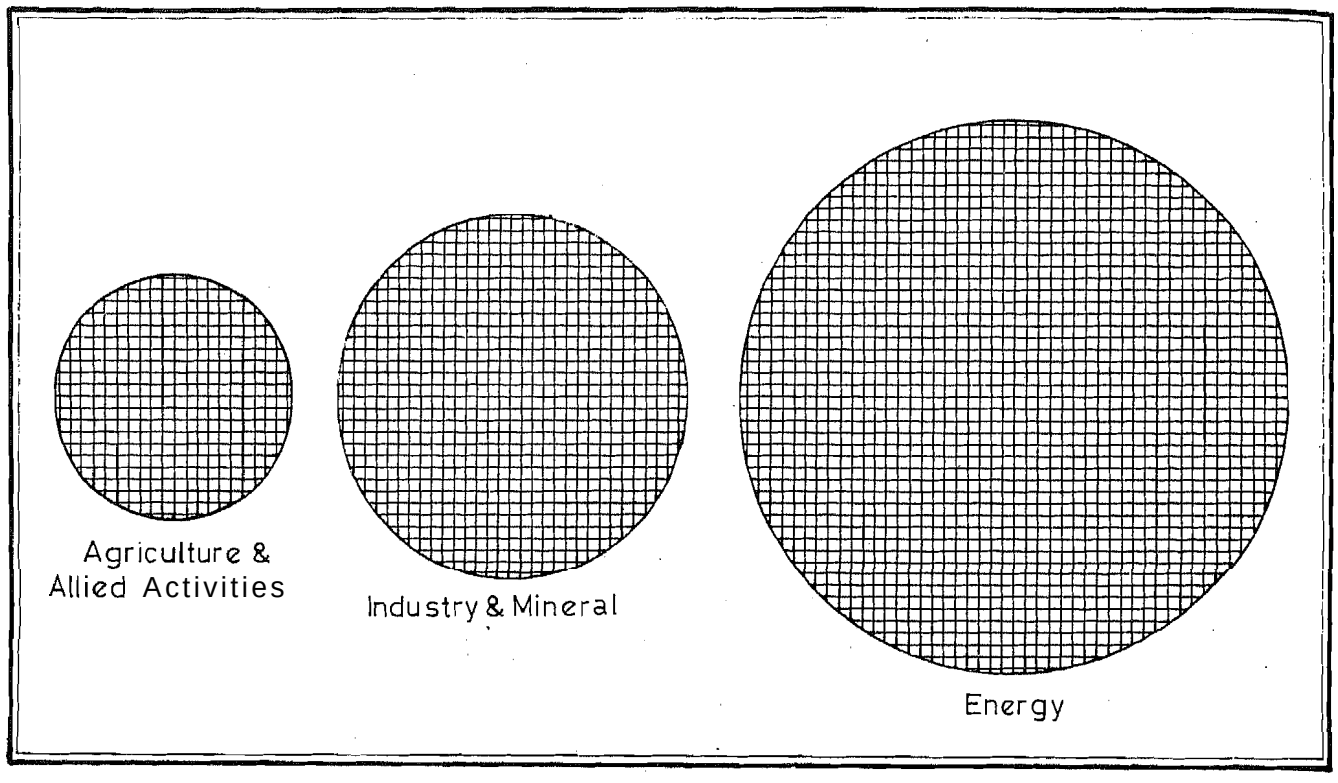


Diagram 8.9B : Circles Showing Annual Plan Outlay for Various Sectors During 1987-88



8.6.4 Pie Diagrams

Pie diagram is a sub-divided circle. A circle is sub-divided to indicate the various components of a given variable. The areas of various sub-divisions in a pie diagram are in the proportion of the data to be represented. The sum total of the different components is taken as 360'' (the total number of degrees around a point) and the degree of the various components are calculated in the proportion of the values of different components to the total. For example, the total cost of production of a dressing table is Rs. 500 of which the cost of raw material is Rs. 200. The total cost of production (i.e., Rs. 500) will be taken as equal to 360° and the share of raw material in it is 40% i.e., 144°. The degree for various other components will be computed in the same manner and the circle will be subdivided on the basis of cumulative degrees. A pie diagram helps us in ascertaining the relationship between the various components very easily. However, the number of sub-divisions in a circle should not be very large. Now let us take up an illustration and prepare a pie diagram.

Illustration 10

Draw a pie diagram to represent the following information regarding the expenditure of a family on various items during a month.

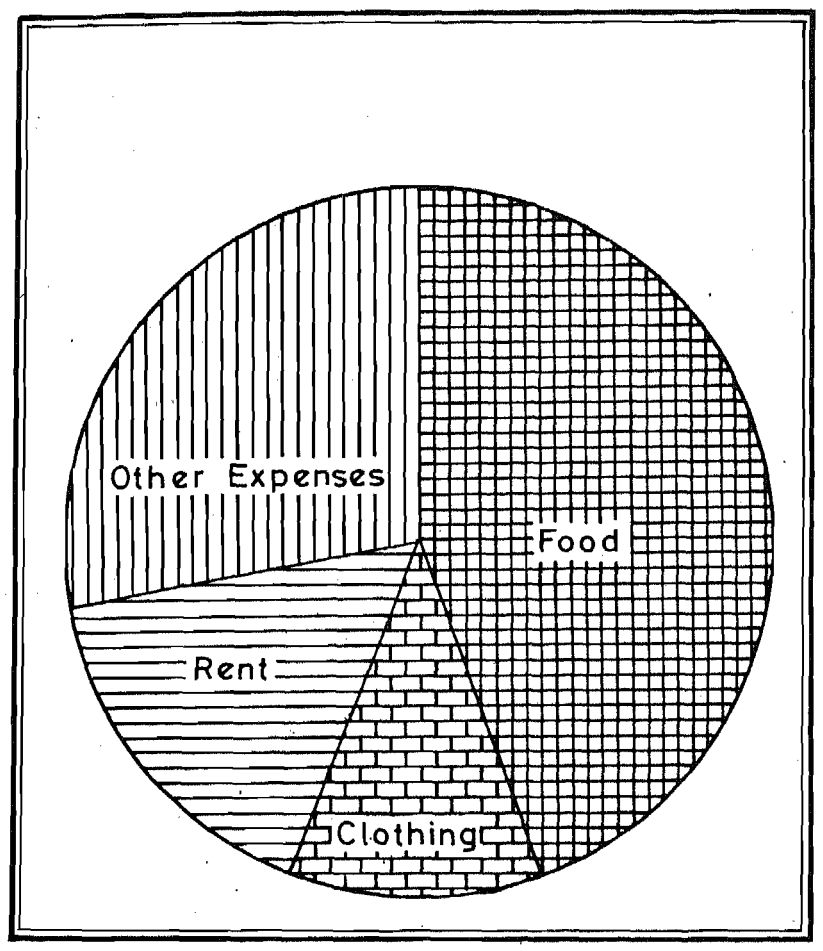
Items	Amount *Rs.*
1) Food	800
2) Clothing	200
3) Rent	300
4) Other Expenses	500

Solution

The total expenditure on various items is Rs. 1,800. This has been taken as equal to 3600. On this basis the number of degrees for various items has been calculated and presented in the table. Then the pie diagram is drawn as shown in Diagram 8.10.

Items	Amount (Rs.)	Degrees	Cumulative Degrees
1. Food	800	160	160
2. Clothing	200	40	200
3. Rent	300	60	260
4. Other Expenses	500	100	360
Total	1800	360	

Diagram 8.10 : Pie Diagram Showing the Monthly Expenditure of a Family on Various Items



Check Your Progress B

1) **Differentiate between a rectangle and a bar.**

.....

.....

.....

.....

.....

2) State the significanceof computing square roots for the given values for the construction of circles.

.....

.....

.....

.....

.....

3) Differentiate between a circle and a pie diagram.

.....

.....

.....

.....

.....

4) State whether the following statements are True or False.

i) Two dimensional diagrams are prepared on the basis of length and height.

ii) Rectangle can also be sub-divided.

iii) In a sub-divided rectangle, the sub-divided portions also indicate area.

iv) There is no need for taking cumulative values for preparing the subdivided rectangle.

v) We calculate the square root of the given data for preparing a circle.

vi) The calculation of square roots of the given values for constructing squares reduce the large values considerably.

vii) While preparing a pie diagram, the degrees of various components are calculated in the proportion of the values of different components to the total.

5) Fill in the blanks with the appropriate word given in the brackets.

i) Two dimensional diagrams indicate . . . . . (volume/area)

ii) We . . . . . square root of the given data for determining the side of a square. (calculate/do not calculate)

iii) If more than one square is to be drawn in a single diagram, then it is . . . . . that the base of all the squares lies on the same line. (necessary/not necessary).

iv) For constructing a circle, we determine its . . . . . (diameter/radius)

v) While preparing a pie diagram, the sum total of various components is taken as . . . . . (360°/300°)

vi) The number of segments in a pie diagram should . . . . . very large. (be/not be)

8.7 LET US SUM UP

Besides presenting in the tabular form, data can also be presented in the form of diagrams and graphs. Visual presentation of data eliminates the dullness of numerical data, makes the comparison of data simpler, helps in locating various statistical measures and establishes trends of past performance.

Diagrams are prepared on the two graphic axes viz., 'X' axis and 'Y' axis. But they need not be prepared on a graph paper. A diagram should have a concise and self-explanatory title.

Colours and shades are used to exhibit various components in a graph. The available space should be optimally utilised to make the diagram significant.

The diagrams may be classified into three categories : 1) one dimensional diagrams, 2) two dimensional diagrams, and 3) three dimensional diagrams. One dimensional diagram is prepared on the basis of length only. the length of various bars is proportionate to the magnitude of the given data. Negative values can also be shown in these diagrams and the bars can either be prepared horizontally or vertically. The one-dimensional diagrams are classified as : 1) simple bar diagrams, 2) multiple bar diagrams, and 3) sub-divided bar diagrams, and 4) percentage sub-divided bar diagrams. A simple bar diagram represents one value whereas the multiple bar diagram represents more than one value. A sub-divided bar diagram represents the different components of a given variable and can also be prepared on percentage basis.

The two dimensional diagrams are prepared on the basis of length and width. These diagrams signify area. Two dimensional diagrams are classified as : 1) rectangles, 2) sub-divided rectangles, 3) squares and circles, and 4) pie diagrams. For preparing a rectangle, its length and width are to be determined. A rectangle can also be sub-divided and the sub-divided rectangles also represent area. For preparing squares and circles, the square root of the given data is to be calculated and then the side (in case of a square) or the radius (in case of a circle) is determined. A pie diagram is segmented circle, where the segments are determined on the basis of 360° around a point. The degrees of the various segments are in the proportion of the values of different components to the total.

8.8 KEY WORDS

Bar : A one dimensional diagram which signifies only length, and width is not significant. Even though it looks like a rectangle, it is not a rectangle because in a rectangle both length and width are significant.

Circle : A two dimensional diagram indicating the area πr^2.

Multiple Bars : A one dimensional diagram with more than one bar indicating either the values of different variables or the values of various components of the same variable.

One Dimensional Diagram : A diagram prepared only on the basis of one dimension i.e., length.

Pie Diagram : A circle divided into sectors showing the relative areas of various components of the same variable.

Rectangle : A two dimensional diagram prepared on the basis of length and width representing variable such that the area represents some variable.

Square : A two dimensional diagram indicating the area. Since all the four sides in a square are identical. one side is determined to prepare it.

Sub-divided Bar Diagram : A bar diagram which is sub-divided on the basis of the values of various components of the same variable.

Sub-divided Rectangle : A rectangle which is sub-divided on the basis of the values of various components of the same variable.

The Dimensional Diagram : A diagram prepared on the basis of two dimension i.e., length and width.

Three Dimensional Diagram : A diagram prepared on the basis of three dimensions i.e., length, width and height.

X-axis : The graphic axis which is drawn horizontally.
Y-axis : The graphic axis which is drawn vertically.

8.9 ANSWERS TO CHECK'YOUR PROGRESS

A) 6) i) True ii) True iii) False iv) True v) True vi) True
vii)True viii) True ix) False x) True xi) True xii)True

7) i) eliminates ii) easier iii) one iv) can v) identical vi) cumulative vii) can viii) can

B) 4) i) False ii) True iii) True vi) False v)True vi) True vii) True

5) i) area ii) calculate . iii) necessary iv) radius v) 360° vi) not be

8.10 TERMINAL QUESTIONS/EXERCISES

Questions

1) What is visual presentation of data? What are the objectives of visual presentation of data?

2) What are the different types of diagrams used in statistics? Discuss various types of one dimensional diagrams.

3) Discuss the principle of diagrammatic presentation of data.

4) Explain with an illustration the method of constructing squares and circles.

5) Discuss the different types of two dimensional diagrams.

Exercises

1) Prepare a simple bar diagram for the following data relating to the production of different oilseeds in 1984-85.

Oilseeds	Production *(kgs./hectare)*
Groundnut	898
Rapeseed and Mustard	711
Soyabean	768
Sesamum	246
Nigerseed	251

2) The following data relate to the production of rice in India in different years.

Year	Production *(lakh tonnes)*
1983-84	601
1984-85	583
1985-86	638
1986-87	604

Prepare a simple bar diagram for the above data.

3) Present the following data in the form of a suitable diagram.

Year	1984	1985	1986	1987	1988
Profit (+) or Loss (-) (in Rs. 000)	+5	+3	-2	+4	-1

4) Present the following data in the form of a multiple bar diagram.

Crops	Production *(million tonnes)*	
	1984-85	1985-86
Wheat	44	47
Rice	58	64
Pulses	12	13
Other Cereals	31	26

5) Represent the following data by means of a subdivided bar diagram.

Results	No. of Students		
	1985-86	198687	1987-88
Fint Division	50	90	80
Second Division	250	300	300
Third Division	100	120	200
Failed	100	90	170
Total	500	600	750

6) Draw percentage sub-divided bar diagram for the following data on cost, proceeds, profit or loss per almirah.

Particulars	1989 January	February	March
Materials	400	550	550
Wages	200	300	350
Other Costs	100	150	200
Sale Proceeds per Almirah	800	1000	1000
Profit (+) or Loss (–)	+100	0	-100

7) Represent the following data relating to the monthly expenditure of two families by means of rectangles.

Head of Expenditure	Family X	Family Y
Food	1,000	1,000
Clothing	300	500
Rent	500	750
Miscellaneous	200	250
Total	**2,000**	**2,500**

8) Prepare squares and circles for the following data :

Country	Yield of Rice (In *lbs.* per *acre*)
India	728
USA	1,469
Italy	2,903

9) Percentage shares of different newspapers sold in New Delhi are given below. Prepare a pie diagram for this data.

Name of the Newspaper	% Share
Times of India	
Indian Express	12
Hindustan Times	28
Others	14
Total	100

10) The following data relates to the demand for electricity in 1984-85 by different consuming sectors. Prepare a pie diagram for this data.

Consuming Sector	Demand (*Billion Kwh*)
Industrial	73.5
Domestic	15.5
Agriculture	21.0
Others	15.0

Note: These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not send your answers to the University for evaluation. These are for your practice only.

UNIT 9 GRAPHIC PRESENTATION

Structure

9.0 OBJECTIVES

After studying this unit, you should be able to :

- state the importance of graphic presentation
- describe the principles of preparing a graph
- list different types of graphs, and
- prepare different types of graphs.

9.1 INTRODUCTION

In Unit 8 you have learnt that the visual presentation of data eliminates the dullness in the presentation of quantitative data and makes it more interesting. Visual presentation also helps in comparison of data or determining the trends of the past performance. You have already studied about one of the techniques of visual presentation of data i.e., diagrammatic presentation. Another important technique of visual presentation of data is the presentation in the form of graphs. In this unit you will learn about the principles of preparing graphs, different types of graphs for time series and frequency distributions and the methods of preparing them.

9.2 IMPORTANCE OF GRAPHIC PRESENTATION

Graphic presentation of data is also pleasing to the eye. It leaves a strong impact on the mind and it is easier to draw trends of data. The graphic presentation of data had the following advantages :

i) Graphic presentation of data renders comparison of data much easier. The direction of curves or straightlines on the graphs makes it very simple. to draw comparisons.

ii) Graphic presentation of data helps in establishing trends of the past performance. The presentation of time series data on a graph makes it possible to interpolate or extrapolate the values. Thus it helps in forecasting.

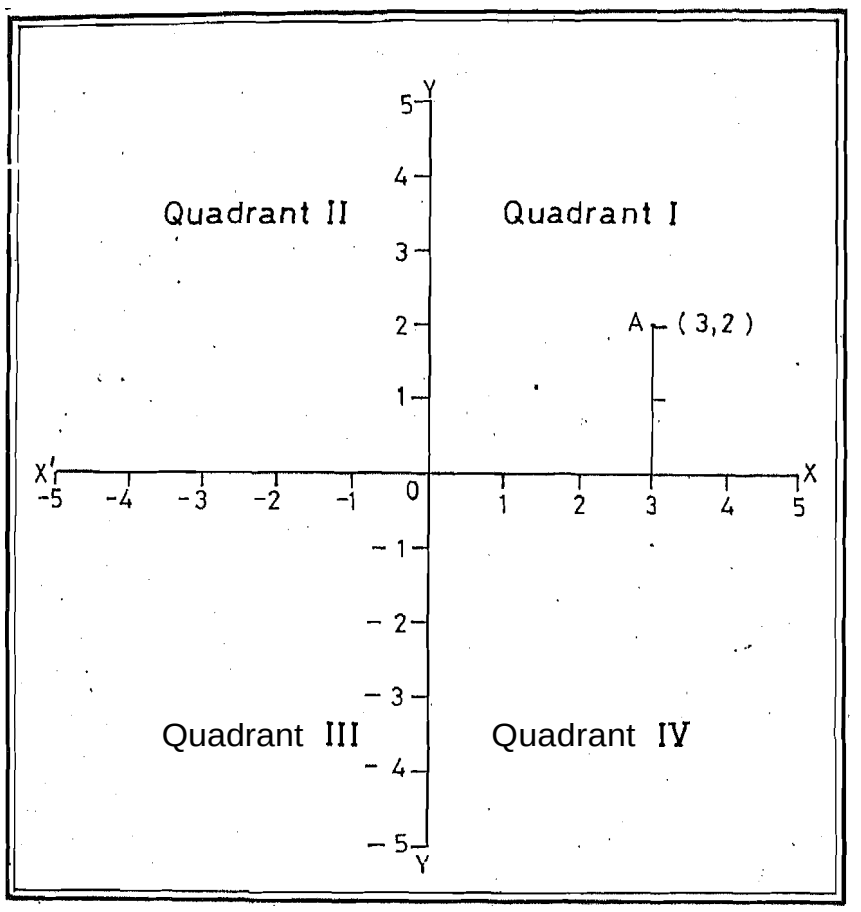
iii) Graphic presentation of data makes it possible to determine the values of the positional averages such as median, quartiles, mode, etc. The graphs of frequency distribution help us to locate these values.

iv) Through graphic presentation it is also possible to establish correlation between two variables. Scatter diagram is a graphic presentation technique to determine the degree of correlation. As the study of correlation is not included in this course, we will not discuss about the scatter diagram method of graphic presentation in this unit. We discuss only graphs of time series and frequency distribution.

9.3 PRINCIPLES OF PREPARING A GRAPH

Graphs are prepared on the basis of the coordinate system of plotting points. When two perpendiculars are drawn on each other, the intersecting point of these perpendiculars is called **the originating point or the origin.** The horizontal line is called X axis and the vertical line Y axis. The intersection of two perpendiculars drawn on each other provides us **four quadrants.** The two perpendicular lines and the four quadrants generated by them have been shown in Figure 9.1.

Graph 9.1 : Parts of a Graph



The positive values of X are taken to the right of origin and negative values on the left of origin. The positive values of Y are taken above the origin and negative values below the origin. The position of a point of the graph is fixed by measuring how much it is away from the origin along with the X axis and along with the Y axis. It is designated by writing the X distance and then Y distance and enclosing both in a bracket. In the graph 9.1 above, point A (3,2) has been plotted. This point **A** is away from the origin by **3** units along X axis and by **2** units along Y axis.

A point with positive values on both axes is plotted in Quadrant I. If, however, there is any negative value, then the point will be plotted in a different quadrant. If X values are negative and Y values are positive, the point on the graph will be in Quadrant II. If X values are positive but Y values are negative, the position will be in Quadrant IV. And if both the X and Y values are negative, the plotting will be in Quadrant III. The graph of statistical data is usually in Quadrant I.

9.4 GRAPHS OF TINE SERIES — HISTORIGRAMS

Broadly, the graphs of statistical data have the following two types :

i) Graphs of Time Series called Historigrams

ii) Graphs of Frequency Distribution

Time series is a series of data which is depicted in a chronological order. The data relating to the sales of an organisation over a period of ten years is an example of time series data. A graph of time series is prepared to show the value of one or more variables over a period of time. A graph of time series data is called Historigram because history is represented graphically.

Principles of Constructing Historigrams

A time series graph or a historigram is constructed on the basis of the following principles :

1) The time is taken as an independent variable and is, therefore, represented on X axis. The value of data is taken as dependent variable and is represented on Y axis For example, in preparing the graph of the data relating to sales of a business concern for the period. 1980-88, sales will be shown on Y axis, whereas the years will be taken on X axis. After plotting the different points corresponding to given data, the points are joined by straight line in the order of time. The graph so formed is called the historigram of the given data. The rise and fall of lines plotted indicates how the data is changing over time. The various levels so plotted, taken together, are also called the curve plotted corresponding to the given data.

2) The Y axis (also called vertical axis) normally starts with zero. However, when there is a wide difference between the lowest value of the given data and zero, the Y axis can also be broken and a false base line can be taken. You will study in detail about the false base line later in this unit.

3) The scales on two axesshould be taken in a manner so that the values of the data are depicted significantly.

4) A graph must have a concise and self-explanatory title.

5) The data relating to more than one variable can be shown by a historigram. In such a case, there are more than one curve in a historigram, each curve representing a separate variable. Normally different curves are marked differently so that they can easily be distinguished from each other. Sometimes, colours are also used to show the different curves.

6) If the variables are measured in different units, double scale can be taken on Y axis.

7). While constructing a graph, the scales adopted should be clearly indicated.

9.5 TYPES OF HISTORIGRAMS

A historigram may be constructed in two ways : 1) It can be constructed on a natural scale where the graph reflects the changes in absolute values over a period of time. 2) It can be constructed on a ratio scale where the graph reflects the relative changes over a period of time.; In this course, however, we will study only the first method i.e., the natural scale graph. In a natural scale graph, the values of the dependent variable are taken on Y axis by marking the scale in such a way that equal distance on Y axis represents equal addition of values. This is the usual method of marking the scale and in the same as shown in Graph 9.1 in Section 9.3. Historigrams can be classified into four types as below :

1) One Dependent Variable Historigram

2) More than one Dependent Variable Historigram

3) Mixed Graph

4) Range Graph

These different types of historigrams and the methods of preparing them are discussed below]

9.5.1 One Dependent Variable Historigram

In this type, there is only one dependent variable. As stated earier, the values of dependen variable are taken on Y axis, whereas the time is taken on X axis. For instance, the dat

relating to sales over a period of time is an example of one dependent variable historigram. In this case, the data on sales will be plotted on Y axis and time will be taken on X axis. This is the simplest type of graph. Study Illustration 1 carefully and understand the method of constructing one dependent variable historigrams.

Illustration 1

The following data relates to the training of officers by an institute during different years. Prepare a suitable historigram for the same.

Years	:	1980	1981	1982	1983	1984	1985
No. of Officers Trained.	:	100	217	36	90	56	70

Solution

Since the data relates to only one dependent variable, we have to construct a one dependent variable historigram. The different years will be taken on X axis and the number of officers trained on Y axis. As all the values of dependent variable are positive, the graph will be in the first quadrant. Therefore, we can mark X axis at the bottom of the graph paper and the Y axis towards the extreme left hand side. Look at Graph 9.2 carefully. On X axis there is a space of 10 big squares. So let us take the starting year of the data at the origin itself and then let two big squares represent one year. So, the scale on X axis should be marked accordingly and the word 'years' will be written under it. On Y axis we have a space of about 12 big squares. We have to show a maximum value of 217. The scale is normally marked in round numbers. So let one big square represent 20. The scale should be written accordingly, and the words 'No. of Officers'should be written on the side. Now plot various points (1980, 100), (1981, 217), etc., and join them by straight line. We have to write a short title also. Look at Graph 9.2 for the graph drawn for this data.

Graph 9.2 : One Variable Historigram Showing the Number of Officers Trained by an Institute

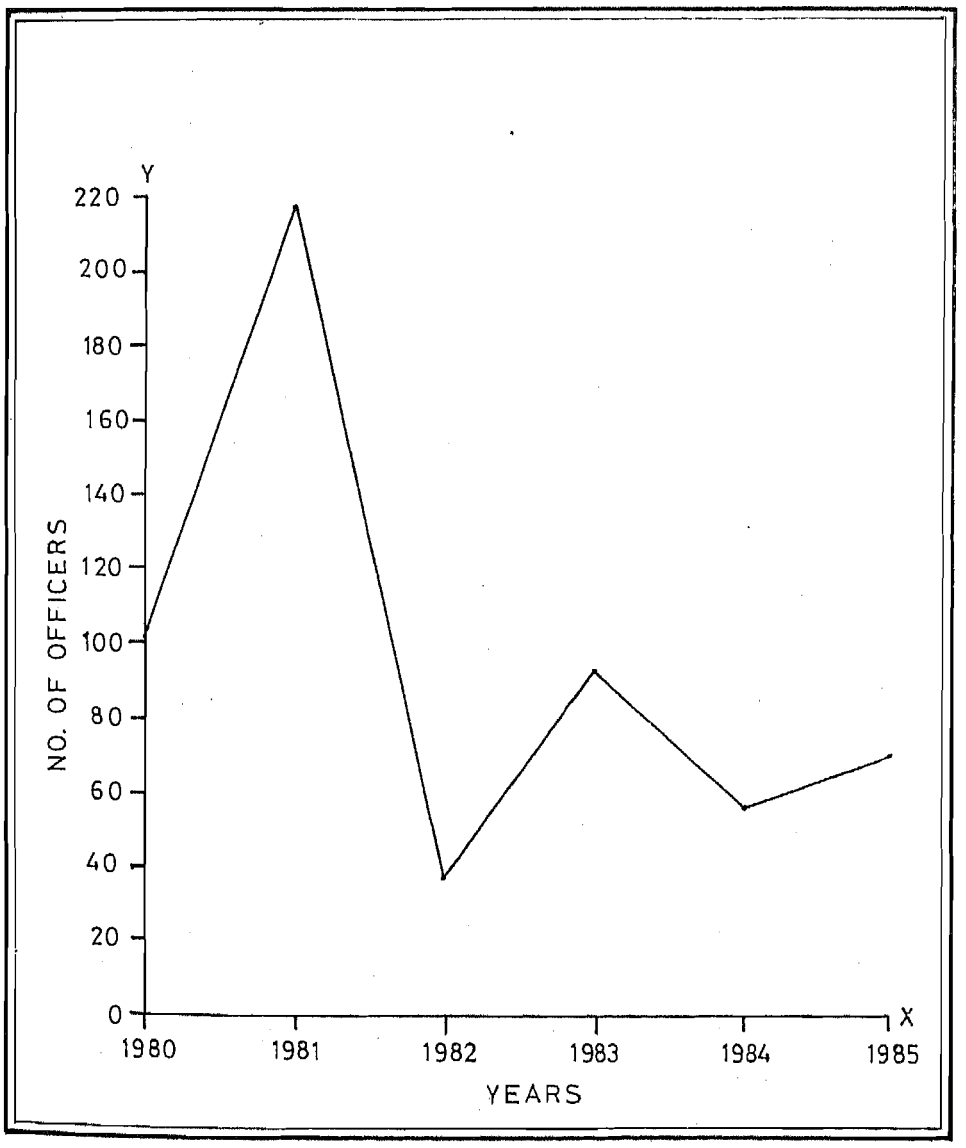


Illustration 2

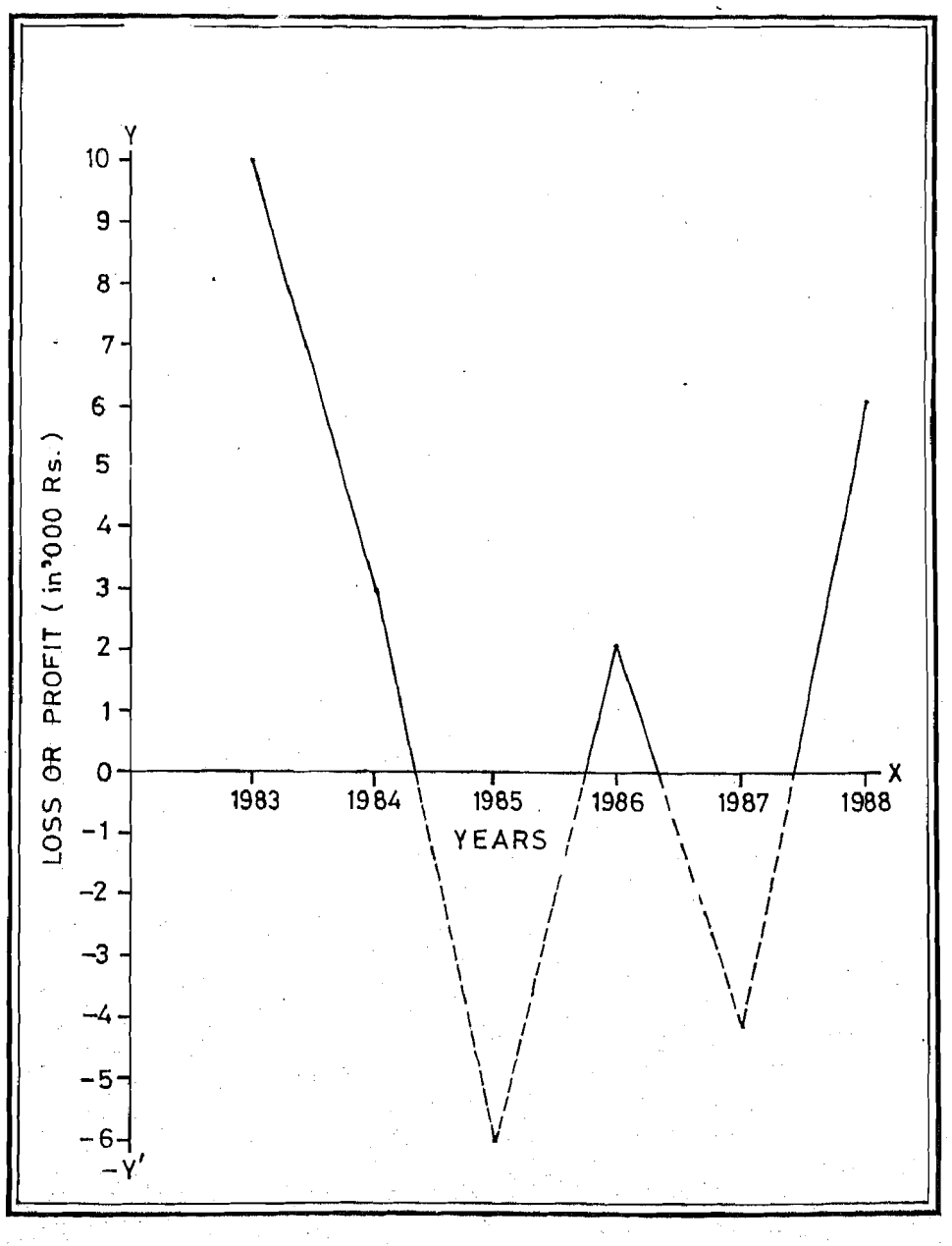
The following data relates to profit/loss of a business concern. Prepare a single variable graph for this data.

Years	:	1983	1984	1985	1986	1987	1988
Profit (+) or Loss (-) ('000 Rs.)	:	+10	+3	-5	+2	-3	6

Solution

In this illustration, both the profit and loss are given. If profit is shown by positive values, loss should be shown by negative values. The positive values should be plotted on the Y axis above the origin, whereas the negative values should be plotted on the Y axis below the origin. So in this case X axis will not be taken towards the bottom of the graph but it will be somewhere in the middle, depending on the magnitude of the negative values to be shown. Look at Graph **9.3** carefully and examine how the X axis is drawn. In the present case, depending on the space available, we have taken two big squares to represent one year and one big square to represent profit/loss of rupees one thousand. Study Graph **9.3** carefully and understand how both profit and loss are plotted.

Graph 9.3 : One Variable Graph Showing the Profit and Loss of the Firm during 1983-88



False Base Line

Wh'le studying the principles of constructing .istorigram, we have discussed that the vertical axis sh uld usually start with zero. If there is a wide differe ce between zero and the lowest value of the given data, the vertical axis is b oken and a **'false base line'** is taken (which is drawn in a zig-zag manner). Th. basic purpose of drawing a 'false base line' is to show the data more significantly. A value equal to the lowest value in the data, or a round fi; re less than the lowest given value is marked on Y axis as the 'false base'. Let us understa d this practically through an illus ration.

Illustration 3

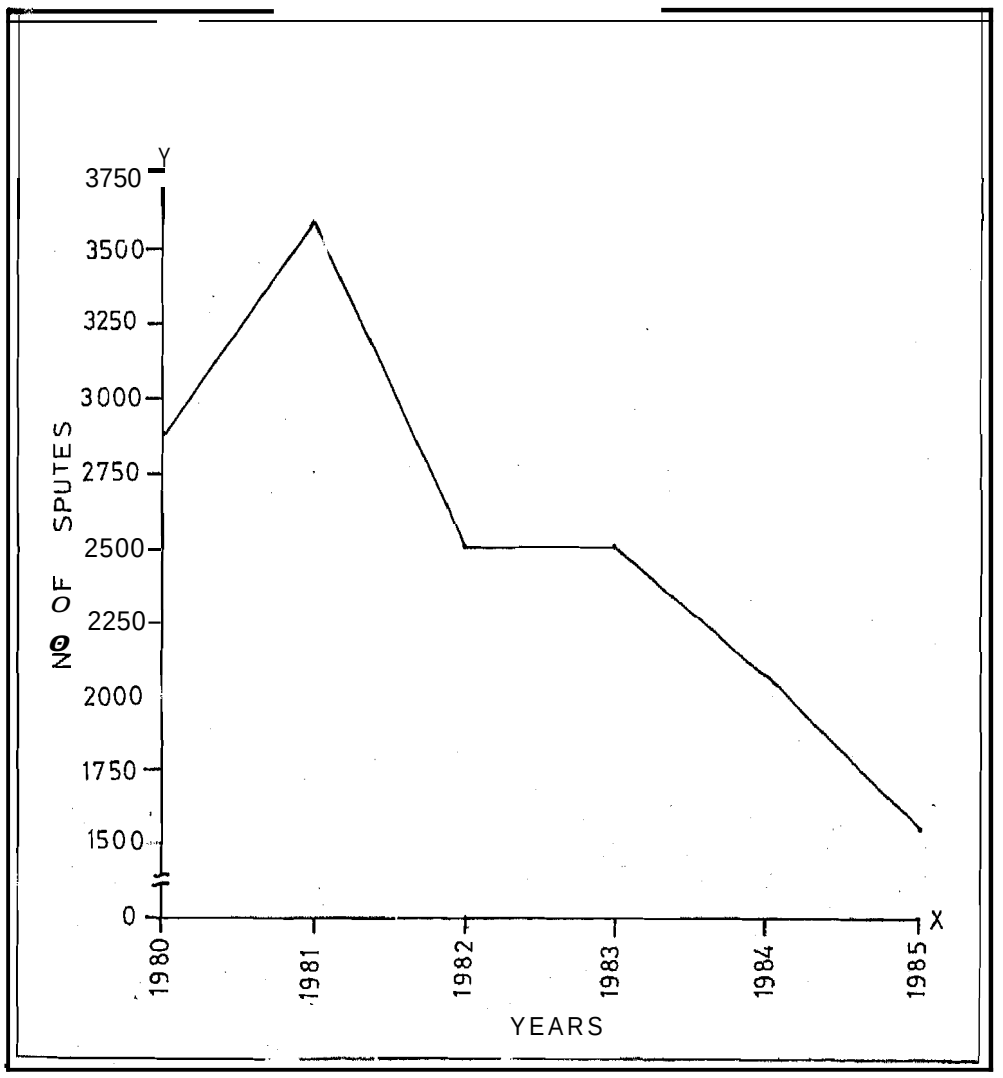
The following data relates to the number of industrial disputes in India. Prepare a historigram for this data.

Years	:	1980	1981	1982	1983	1984	1985
No, of Disputes	:	2,856	3,589	2,483	2,488	2,061	1,522

Solution

The lowest number of industrial disputes is 1,522 which is very much away from '0'. Hence, the Y axis has been broken near the X axis and th fal *e* base line has been drawn. The first value for marking the scale on the false base line may be taken as 1,500, a value which is less than lowest number of disputes i.e., 1,522. Depending on the space available on Y axis, . ow one big square is taken to represent 250. Now study Graph 9.4 carefull nd understand how the false base line is drawn and the data is plotted.

Graph 9.4 : One Variable Graph with a False Base Line



Note : If the false base is not taken, space corresponding to the scale f om 0 to 1,500 on Y axis (which will be equal to six big squares on the scale selected) will go waste. This also condense the graph drawi in the upper portion of the paper. The broken line indicates that a portion of the graph paper, having no values plotted in, has not been shown.

9.5.2 More than One Dependent Variable Historigram

Sometimes, the data in a historigram may relate to more than one dependent variable. For example, the data may relate to the production of both wheat and rice over a period of time. The histogram, which will be prepared to show the two curves (one for wheat production and the other for rice production), will be known as **more** than **one** dependent **variable** graph. These more than one variable graphs are prepared exactly in the same manner as we prepare one dependent variable graph. The 'false base line' can also be taken in this case. Now study Illustration 4 carefully and understand the practical procedure involved in the preparation of this type of graphs.

Illustration 4

The following data relates to birth rate, death rate and growth rate in India. Draw a historigram with more than one dependent variable for this data.

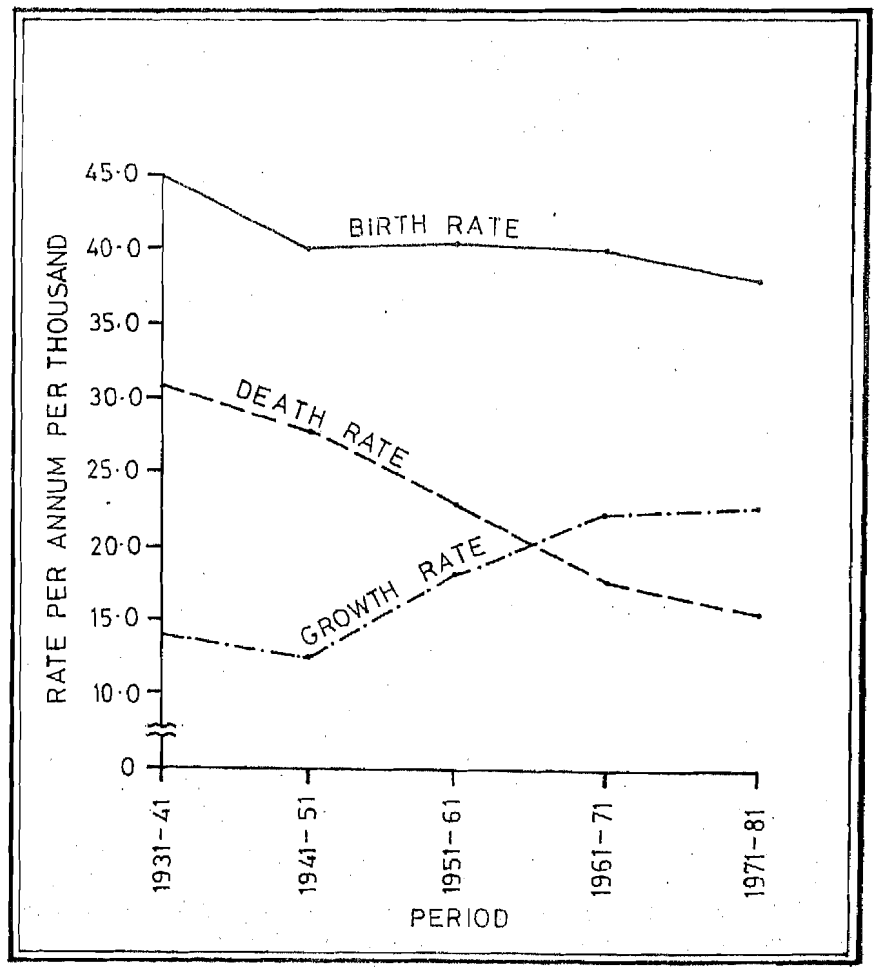
(Rate per nnnum per thousands ***population)***

Period	:	**1931-41**	**1941-51**	**1951-61**	**1961-71**	**1971-81**
Birth Rate	:	**45.2**	*39.9*	**40.0**	40.0	**37.9**
Drath **Rare**	:	**31.2**	**27.4**	22.8	**17.8**	**15.4**
Growth Rate	:	14.0	**12.5**	18.1	22.2	22.5

Solution

A multiple variable historigrarn has been prepared for this data and presented in Graph 9.5. The false base line has also been taken. In this graph we have done one curve for each variable viz., 'birth rate', 'death rate', and 'growth rate'. For easy identification, each curve is marked with a different method of joining. The graph clearly shows that with passage of time, **the** fall in death rate is higher than the fall in birth rate and hence the growth rate is rising with tune after 1941.

Graph 9.5 : *More* than One Variable Historigram Showing Birth, Death and *Growth* Rates in India



9.5.3 Mixed Graph

Mixed graph is a type of historigram prepared for two dependent variables where the units of measurement in respect of these two variables are not the same. The values of these two dependent variables are represented by two different scales—one on the usual Y axis and the other Y axis taken on the right of the horizontal axis. The following points should be kept in mind while preparing a mixed graph :

i) The mid-points of the scales taken on two vertical axis should be on the same line.

ii) Two different curves will be drawn—one on the usual Y axis and another on the second Y axis taken on the right of X axis.

iii) False base line may also be taken in this case, if necessary.

Now let us learn the preparation of a mixed graph by taking up an illustration.

Illustration 5

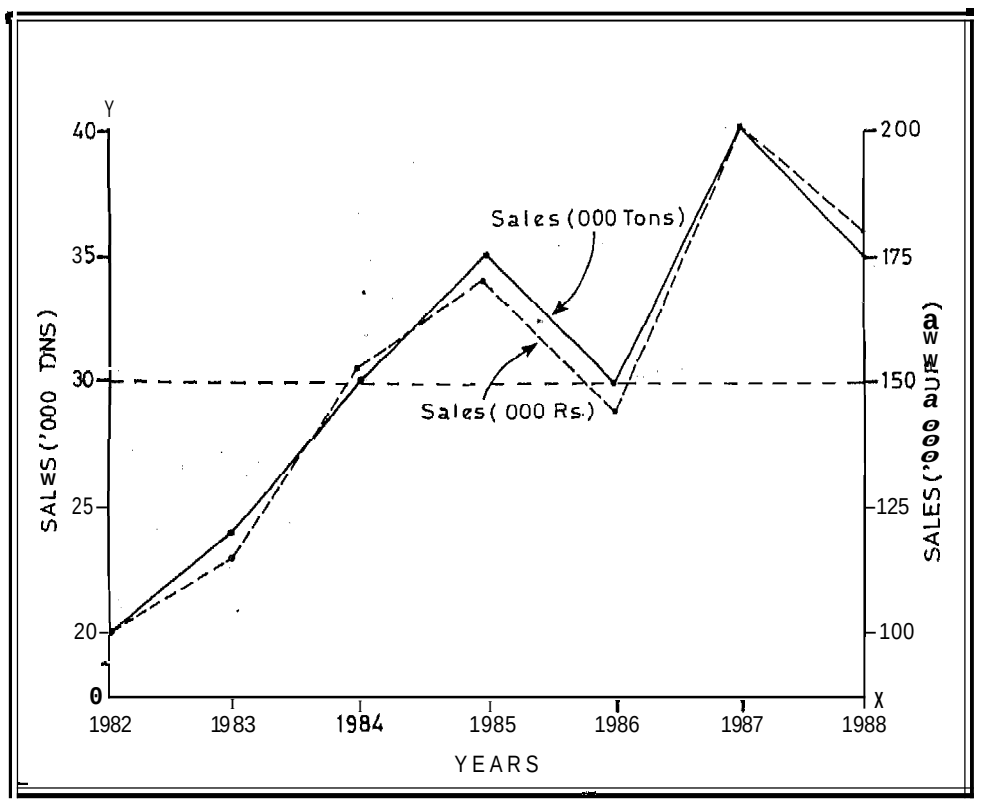
The following data relates to sales of a business concern. Show this data with the help of a mixed graph.

Years	1982	1983	1984	1985	1986	1987	1988
Sales in Quantity ('000 tons)	20	24	30	35	30	40	35
Sales in Value ('000 Rs.)	100	115	155	170	145	200	180

Solution

Now look at the mixed graph presented in Graph 9.6 carefully. Sales in quantity have been shown on the usual Y axis and sales in value have been shown on Y axis taken on the right of X axis. The range for sales in quantity is 20 to 40 thousand tons and that of sales in value 100 to 200 thousand rupees. To show the two curves prominently (not to waste graph space) let us take false base line and mark the starting point 20 on the tons side and 100 on the rupees side. The mid value **30** of tons and 150 of rupees will also be taken on one line. This is shown as dotted line in the graph. On the Y axis relating to quantity, two big squares represent 5 thousand tons. Similarly, on the Y axis relating to sales value, two big squares represent Rs. 25 thousands. To make a distinction between the two curves, we have represented sales in tons by continuous line and sales in rupees by dotted line.

Graph 9.6 : Mixed Graph Showing Quantity and Value of Sales



9.5.4 Range Graph

Sometimes, for the dependent variable, two extreme values (i.e., the maximum and minimum values) are given. Maximum and minimum temperature on a particular day is an example. The graph showing these two extreme values is called a range graph. This type of graph shows two curves, one curve for maximum value and another for the minimum value, for different time periods. This graph is called range graph because it shows the range in the values of the given data (range is the difference between the two extreme values of the data at different points of time).

Illustration 6

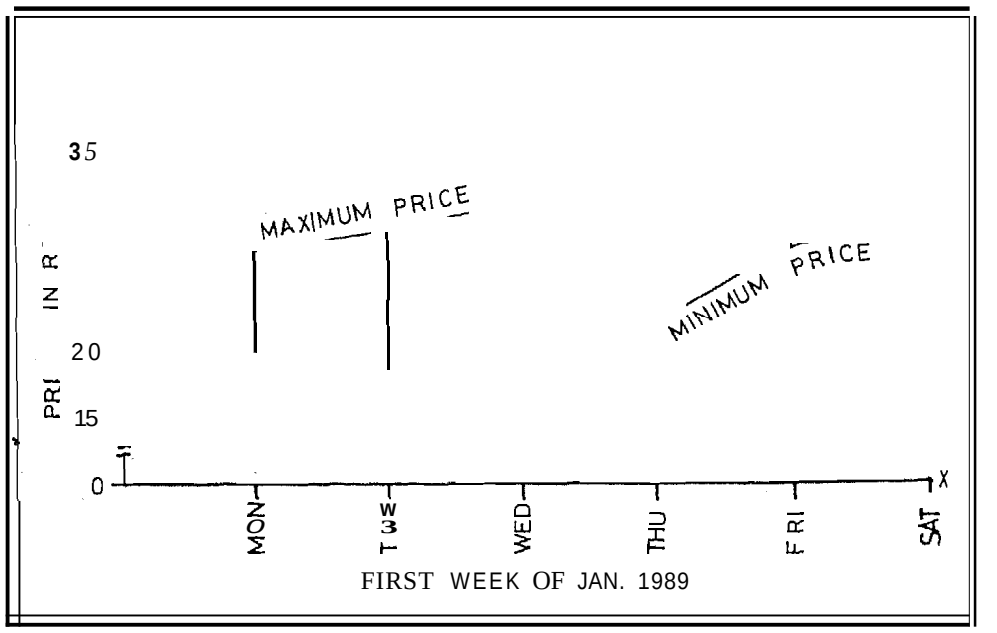
Draw a range graph for the following data on minimum and maximum prices of the share of XYZ Company during the first week of January 1989.

Day	Minimum Price (Rs.)	Maximum Price (Rs.)
Monday	20	27
Tuesday	18	29
Wednesday	27	31
Thursday	22	25
Friday	28	30
Saturday	30	34

Solution

The range graph has been constructed and presented in Graph 9.7. It presents two curves, one for maximum prices and the other for minimum prices.

Graph 9.7 : Range Graph Showing the Maximum and Minimum Prices of the Shares of XYZ Company



The vertical lines on different days show the range of the prices on that day. The rise and fall of two horizontal lines show the rise and fall of maximum and minimum prices over time. Such a graph, therefore, brings to light these two kinds of information.

Check Your Progress **A**

1) What is a graph of time series?

.....

.....

.....

.....

2) Distinguish between one dependent variable historigram and more than one dependent variable historigram.

.....

.....

.....

.....

.....

3) What is a false base line?

.....

.....

.....

.....

.....

4) Distinguish between a mixed graph and a range graph.

.....

.....

.....

.....

.....

.....

5) State whether the following statements are True or False.

i) Graphic presentation of data renders comparison of data easy.

ii) The trends of past performance cannot be established with the help of graphs.

iii) If the values are positive, the graphs are generally prepared in Quadrant I.

iv) In the historigram, the Y axis is never broken.

v) A historigram cannot show data pertaining to more than one dependent variable.

vi) The basic objective of taking a false base line is to show data more significantly.

vii) In case of more than one dependent variable historigram, the dependent variables are taken on X axis.

viii) False base line cannot be taken in case of a mixed graph.

6) Fill in'the blanks with the appropriate word given in the bracket.

i) Graphic presentation of data ..... in determining the values of positional averages. (helps/does not help)

ii) If the value of dependent variable is negative, the graph will be prepared in quadrant ..... (III/IV)

iii) In a historigram, the time is taken on ..... axis. (X/Y)

iv) A graph .....be prepared on a ratio scale. (can/cannot)

v) False base line ..... be taken in case of more than one dependent variable historigram. (can/cannot)

vi) Mixed graph is prepared to show the values in respect of ..... dependent variables having different units of measurement.

vii) In case of a mixed graph, the mid points of the two scales should..... on the same line. (be/not be)

viii) Range graph shows the ..... of the two extreme values of a variable. (difference/addition)

9.6 GRAPHS OF FREQUENCY DISTRIBUTION

You have studied in Unit 7 that the frequency distributions can be presented in the form of tables. In fact, frequency distribution can also be presented in the form of graphs. Compared to tabular presentation, graphs of frequency distribution are helpful in identifying characteristics and relationships. Graphs of frequency distribution are also useful in locating the positional averages such as mode, median, quartiles, etc. Let us study the basic principles to be followed while preparing the graphs of frequency distribution.

Principles of Constructing Graphs of Frequency Distribution

Graphs of frequency distribution are constructed on the basis of the following principles:

1) The values of the variable are shown on X axis. For example, in **a** frequency distribution showing age (in years) of the students, the age will be shown on X axis.

2) The values derived from the frequency of the items/classes are shown on Y axis.

3) It is not necessary that horizontal axis should start from zero. But vertical axis is not broken in these graphs. False base line cannot be taken.

4) The scales on the two axes be taken such that they reflect the data significantly. These scales should be clearly specified.

5) The graph should have a concise and self-explanatory title.

9.7 TYPES OF FREQUENCY DISTRIBUTION GRAPHS

The graphs of frequency distribution can be classified as : 1) Histogram, 2) Frequency Polygon, **3)** Frequency Curve, and 4) Ogive or Cumulative Frequency Graph. Let us now understand the procedure involved in the preparation of these four types of graphs.

9.7.1 Histogram

A histogram is a series of rectangles each proportionate in width to the magnitude of a class interval and proportionate in area to the number of frequencies concerning the class intervals. In this graph the items/class intervals are taken on X axis and values derived from frequency are taken on Y axis as height of the rectangle. To derive the height, smallest class interval will be taken as one unit width. The width of all other class intervals will be determined in terms of this. The frequency will then be divided by this number. The value so obtained is called as frequency density or adjusted frequency. This represents frequency per unit class interval. When class intervals are all equal, every class interval will have a width of one unit. So dividing frequency (which is represented by the area) by one, we get the height (i.e., frequency density) equal in number to frequency. So in case of equal class intervals, height may be taken proportionate to frequency. But actually in concept it is area which is proportionate to the frequency.

In constructing histogram there should not be any gap between two successive rectangles. If the data is given in inclusive class interval or in discrete series, it should be first converted to continuous data with exclusive type class intervals. **A** histogram helps in determining the mode of the data, about which you will learn later in this course.

Illustration 7

The following distribution relates to the marks secured by the students of a college in statistics. Present it graphically.

Marks	:	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No. of Students	:	5	15	25	50	40	30	10	5

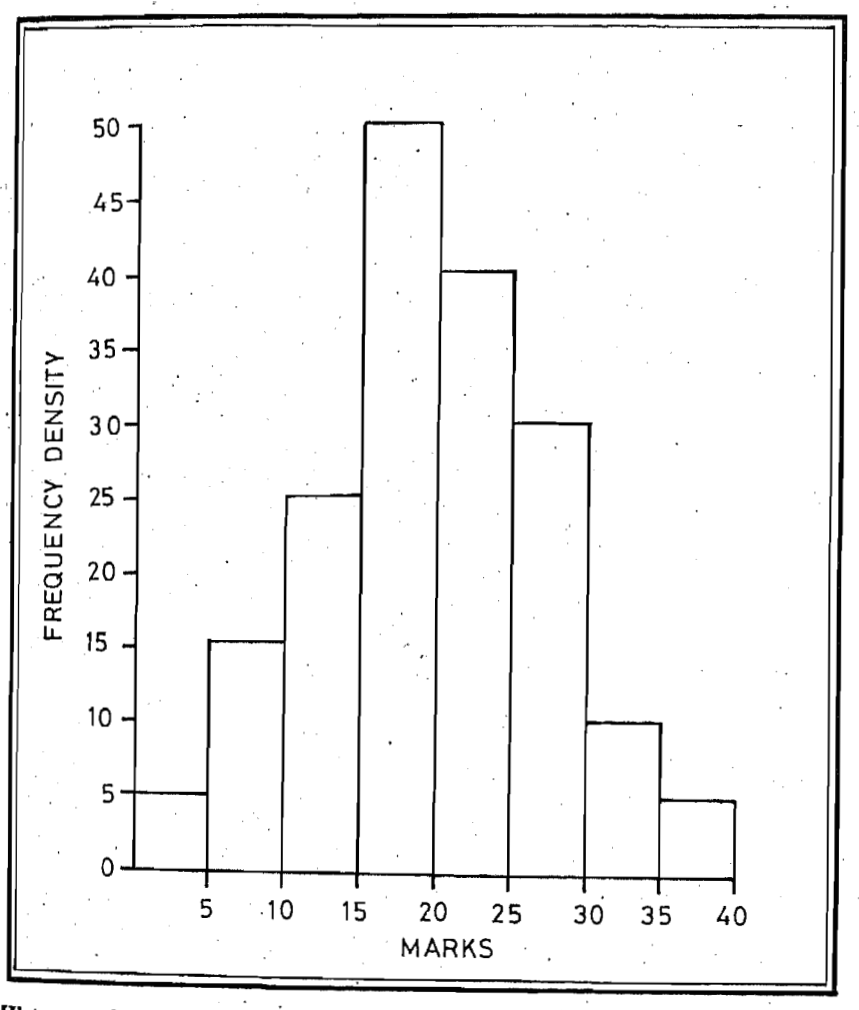
Solution

The data in this illustration has been given in the form of continuous frequency distribution with exclusive type. So the given class intervals, as it is, will be taken on X axis. As all the class intervals are of the same size, by denoting their length by one unit, each class interval will be of one unit width. Dividing frequency (i.e., area of rectangle) by this width, the height of the rectangles will be equal to their frequencies.

Marks	No. of Students (Area)	Width Units (Base)	Frequency Density (Height)
0-5	5	1	5/1 = 5
5-10	15	1	15/1 = 15
10-15	25	1	25/1 = 25
15-20	50	1	50/1 = 50
20-25	40	1	40/1 = 40
25-30	30	1	30/1 = 30
30-35	10	1	10/1 = 10
35-40	5	1	5/1 = 5

The height will be taken on Y axis. Now on both X axis and Y axis mark the scale in a suitable manner. To construct the histogram. we will construct rectangles on the base bar as class intervals, with height equal to the value derived as above. The histogram, is drawn and presented in Graph 9.8.

Graph : 9.8 : Histogram Showing the Marks of Students in Statistics (Class Intervals with Equal Width)



Histogram in case of Class Intervals with Unequal Width

You must have observed in Illustration 7, that the width of the class intervals is equal, and therefore, frequency and frequency density are numerically same. However, it is likely that the frequency distribution has class intervals with unequal width. Under such circumstance, the frequencies of the class intervals and frequency density will differ. As pointed out above, the class interval with the minimum width is taken as the basis for making the necessary adjustments in other class intervals. The adjusted frequencies or the frequency density of the other class intervals can also be calculated as follows :

$$\text{Adjusted Frequency of any Class} = \frac{\text{Width of the Basic Class}}{\text{Width of the Given Class}} \times \text{Frequency of the Given Class}$$

This method of adjustment will give the same value of height of rectangle as by the method pointed out in the previous example. Under this method also the area of the rectangle represents the frequency. Now let us take up an illustration and understand this practically.

Illustration 8

Draw a histogram for the following distribution relating to the marks secured by the students of a class in accountancy.

Marks	:	0-5	5-10	10-15	15-20	20-25	25-30	30-40	40-60
No. of Students	:	5	15	25	50	40	30	20	16

Solution

In this illustration the width of the various classes are not equal. Hence, necessary adjustment should be made in the frequencies of some of the class intervals. Taking 5 (the smallest width) as the normal width, the frequencies of two classes i.e., 30-40 and 40-60 will be adjusted as follows :

$$\text{Adjusted Frequency of any Class} = \frac{\text{Width of the Basic Class}}{\text{Width of the Given Class}} \times \text{Frequency of the Given Class}$$

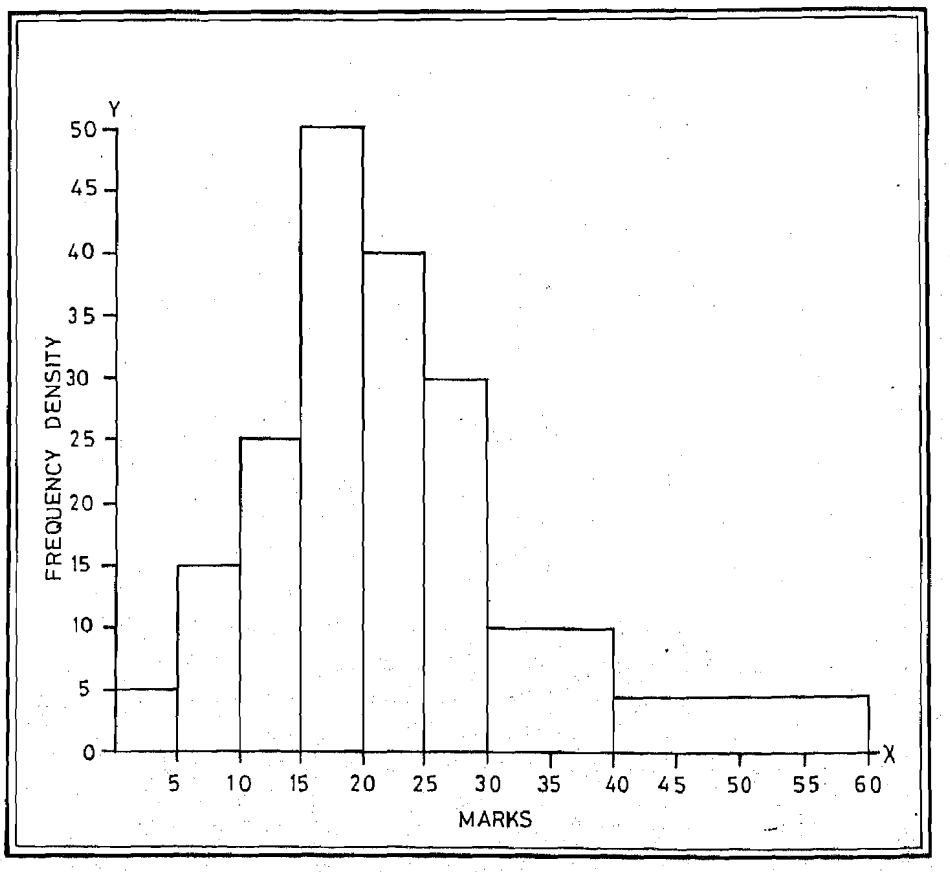
For Class 30-40, the adjusted frequency or frequency density will be = 5/10 × 20 = 10.

For Class 40-60, the adjusted frequency or frequency density will be = 5/20 × 16 = 4.

Let us also calculate the frequency density by the method pointed out in the Illustration 7 and study whether the two methods give the same value for height or not.

The histogram has been constructed and presented in Graph 9.9

Graph 99 : Histogram Showing the Marks of Students in Accountancy (Class Intervals with Unequal Width).



Marks	No. of Students (Area)	Width Units (Base)	Frequency Density (Height)
0-5	5	1	5/1 = 5
5-10	15	1	15/1 = 15
10-15	25	1	25/1 = 25
15-20	50	1	50/1 = 50
20-25	40	1	40/1 = 40
25-30	30	1	30/1 = 30
30-40	20	2	20/2 = 10
40-60	16	4	16/4 = 4

Illustration 9

Draw a histogram for the following data relating to the age (in 'years on the nearest birthday) for members of a club.

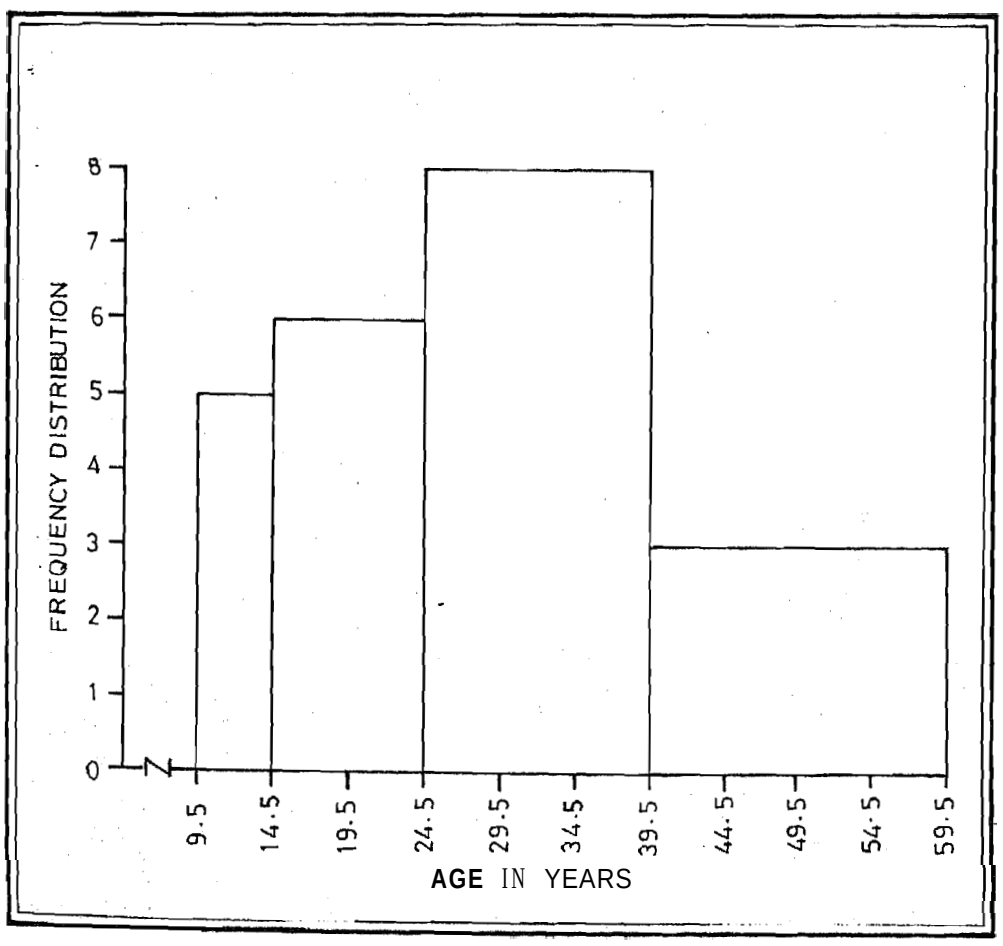
Age in Years	:	10-14	15-24	25-39	40-59
No. of Members	:	5	12	24	12

Solution

In this illustration the class intervals are of unequal size, and are given in inclusive form. Converting them in 'exclusive' form to have no gaps in between and taking the smallest class interval size (i.e., 5) as one unit width, the data is as follows :

Age in Years	:	9.5-14.5	14.5-24.5	24.5-39.5	39.5-59.5
No. of Members (Area)	:	5	12	24	12
Width Units (Base)		1	2	3	4
Frequency Density (Height)	:	5	6	8	3

Graph 9.10 : Histogram Showing the Age of Club Members



Look at Graph 9.10 carefully the histogram for the data is presented. In this graph one big square is taken to represent 5 years. The points 9.5, etc., are marked on a thick line to make the plotting easy. A break is shown on X axis near the origin to indicate that this part of the X axis is not on the same scale as the rest of the axis.

9.7.2 Frequency Polygon

Frequency polygon is another way of showing frequency distribution graphically, There are two ways of constructing a frequency polygon. In the first method, along with the given class intervals, two more class intervals with zero Frequencies are taken one at the beginning of the given classes and the other at the end. Then the frequency densities are plotted against the mid points of all the class and they are joined by straight line. Another way of preparing a frequency polygon is to construct a histogram in a normal way. Here put a dot against the mid points of the tops and the first and last vertical lines of the histogram. Then join the different dots by straight line segments to get the frequency polygon. These methods point out that the frequency polygon originates at a point lying to the left of the lower limit of the first class interval equal to one-half of the width of the class interval. Similarly, it is extended at the point lying at half—the class interval width away to the right of the upper boundary of the last class interval.

It should be noted that the area under the two histograms and frequency polygon is always the same and they represent the total frequency.

Illustration 10

Draw a histogram and a frequency polygon for the following data :

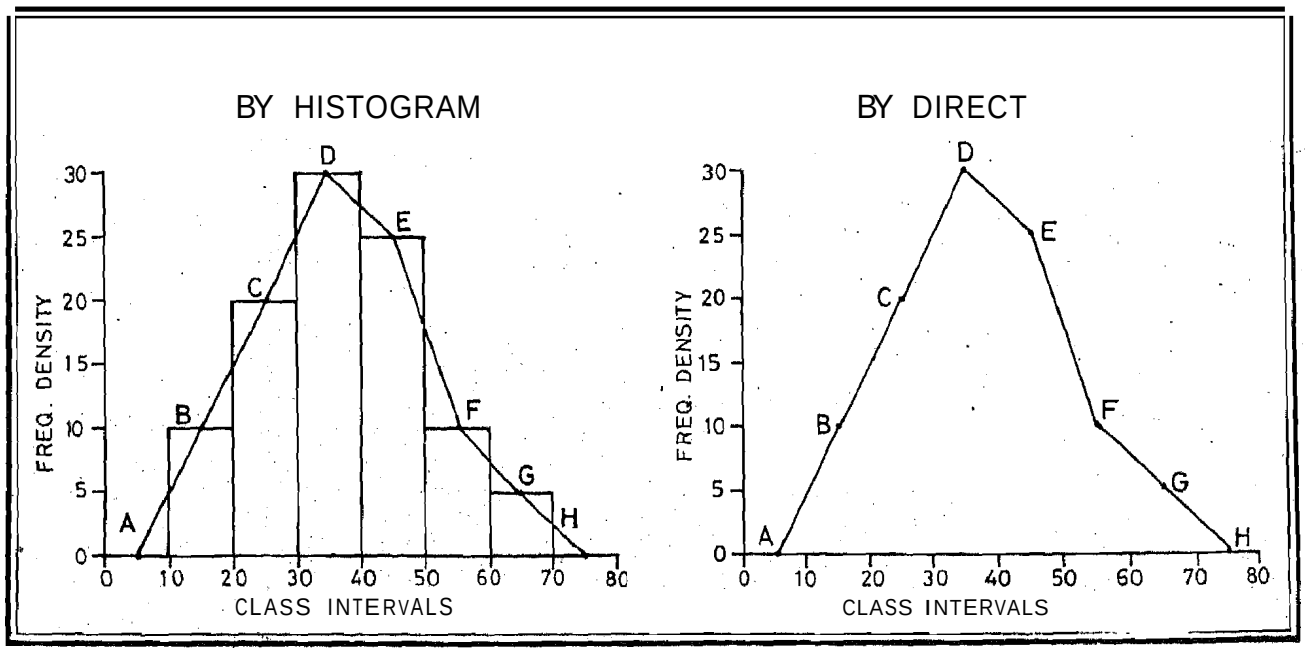
Class Interval	:	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	:	10	20	30	25	10	5

Solution

The calculations for constructing histogram and frequency polygons are as follows :

Class Interval	Frequency (Area)	Width Units (Base)	Frequency Density (Height)	Mid Points
10-20	10	1	10	15
20-30	20	1	20	25
30-40	30	1	30	35
40-50	25	1	25	45
50-60	10	1	10	55
60-70	5	1	5	65

Graph 9.11 : Frequency Polygon by Two Methods



The two class intervals, one before and one after the given class intervals, will be 0-10 and 70-80, since there are no items for these two class intervals. The frequencies for them will be taken as zero. Now the frequency polygon is constructed by two methods : 1) By using histogram and joining the mid points of the top (i.e., A, B, C, D, E, F, G and H). 2) By direct method i.e., using mid points of given class intervals and two extra class intervals, and joining them (i.e., A, B, C, D, E, F, G and H).

Study Graph 9.11 carefully. You will note the following points :

1) In cases of histogram, height at various points within a class interval is same throughout the class interval, while in frequency polygon it is changing. This means that in a histogram frequencies are uniformly distributed within a class interval, while in a frequency polygon frequency density is different at different points.

2) The frequency polygon shows a gradual rate of rise and fall of frequency density, while in histogram it suddenly jumps at the end points of the class'intervals.

Thus, histogram is used when rise and fall of frequencies from one group to another are to be shown permanently, and when gradual rise and fall are to be shown then frequency polygon is used.

Illustration 11

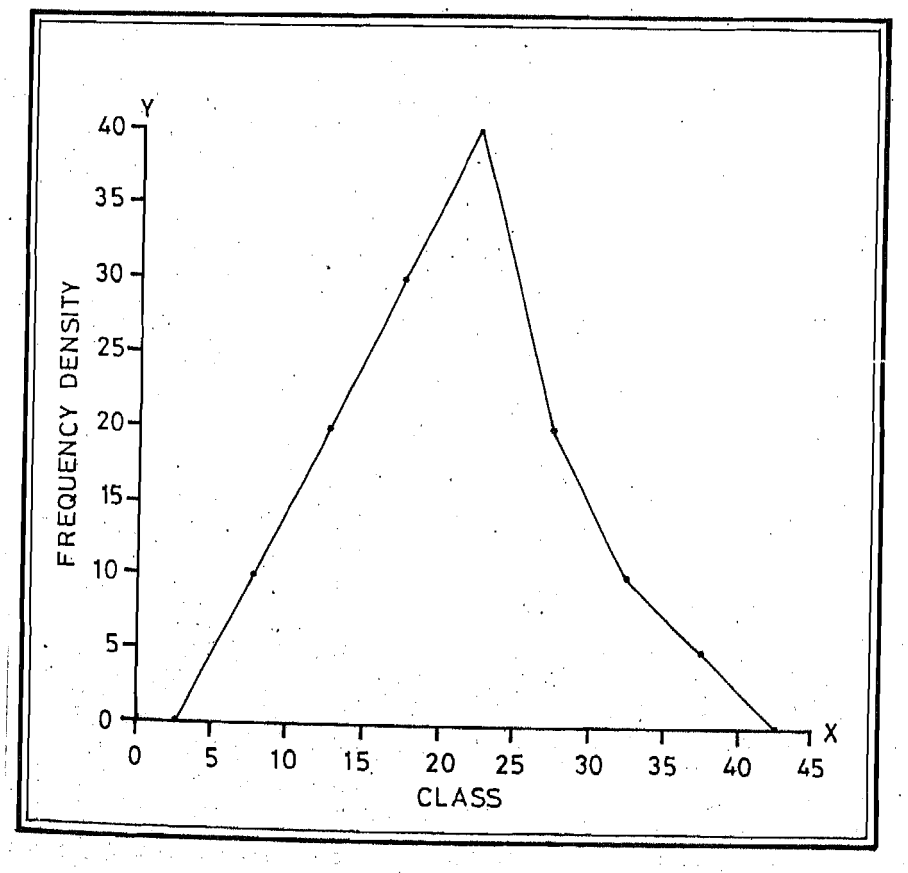
Draw a Frequency polygon for the following distribution.

Mid-Points	:	7.5	12.5	17.5	22.5	27.5	32.5	37.5
Frequency	:	10	20	30	40	20	10	5

Solution

In this illustration, the frequency polygon can easily be constructed on the basis of the mid points. The length of the first class interval is the difference between two successive mid points i.e. 12.5 – 7.5 = 5. Similarly, the length of class interval is same throughout the data. So the mid points of two class intervals, one before and one after the given data, will be 2.5 and 42.5, and the frequency density will be numerically the same as the given frequencies. Study carefully the frequency polygon presented in Graph 9.12.

Graph 9.12 : Frequency Polygon



9.7.3 Frequency Curve

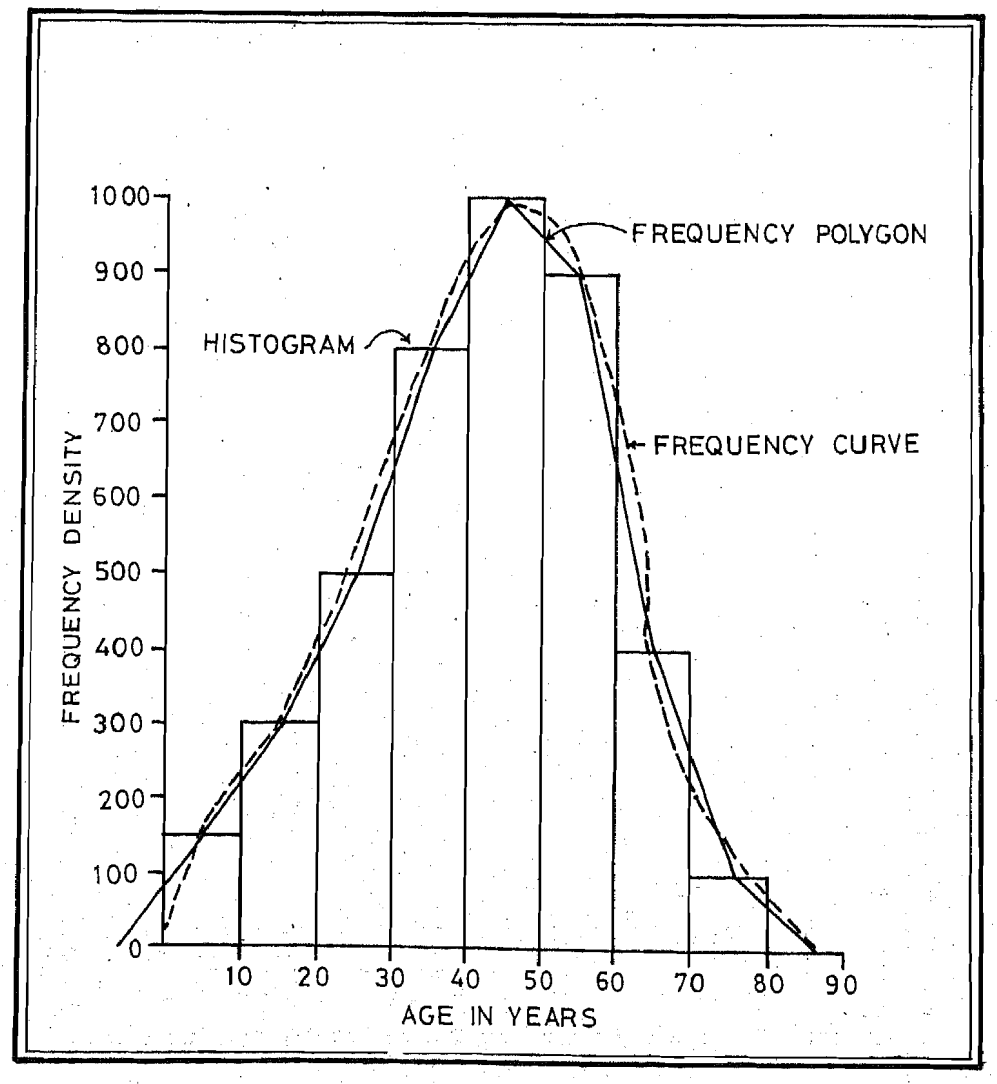
A frequency curve is Constructed to smoothen the frequency polygon. It is prepared by drawing a free hand curve by the side of the frequency polygon in such a manner that the area under polygon and the area under curve drawn is the same. The total area of the frequency curve also represents the total frequency. Frequency curve represents the generalisation of frequency polygon. This means that if the data given is a small sample out of a very large group, then frequency curve gives the general tendency as exhibited by the big group.

Illustration 12
Draw a histogram, a frequency polygon and a frequency curve for the following data.

Age (in Years) :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No, of Residents :	150	300	500	800	1000	900	400	100

Solution
In this case all the class intervals are equal. We can construct histogram in the same manner as described earlier. Frequency polygon can now be drawn with the help of histogram. Now study Graph 9.13 carefully. You will notice that the first vertical line of histogram is the Y axis itself. So when we join the mid point of this first vertical line with the mid point of the top of the first rectangle, the polygon line will go to negative side of X axis. Normally, this implies that there are negative items in the data. In fact, there are no negative values in the data. Actually, frequency polygon is allowed to go on negative side so that areas of histogram and polygon are the same and represent the total frequency. To draw the frequency curve we have to follow the general pattern of the frequency polygon and draw a smooth line which is sometimes below the polygon and sometimes above it so that the areas of polygon and the curve are equal. In a case like the present one, adjustment for negative side is done by starting the frequency curve from '0'.

Graph 9.13 : Histogram, Frequency Polygon and Frequency Curve for Age of Residents



9.7.4 Ogive or Cumulative Frequency Graph

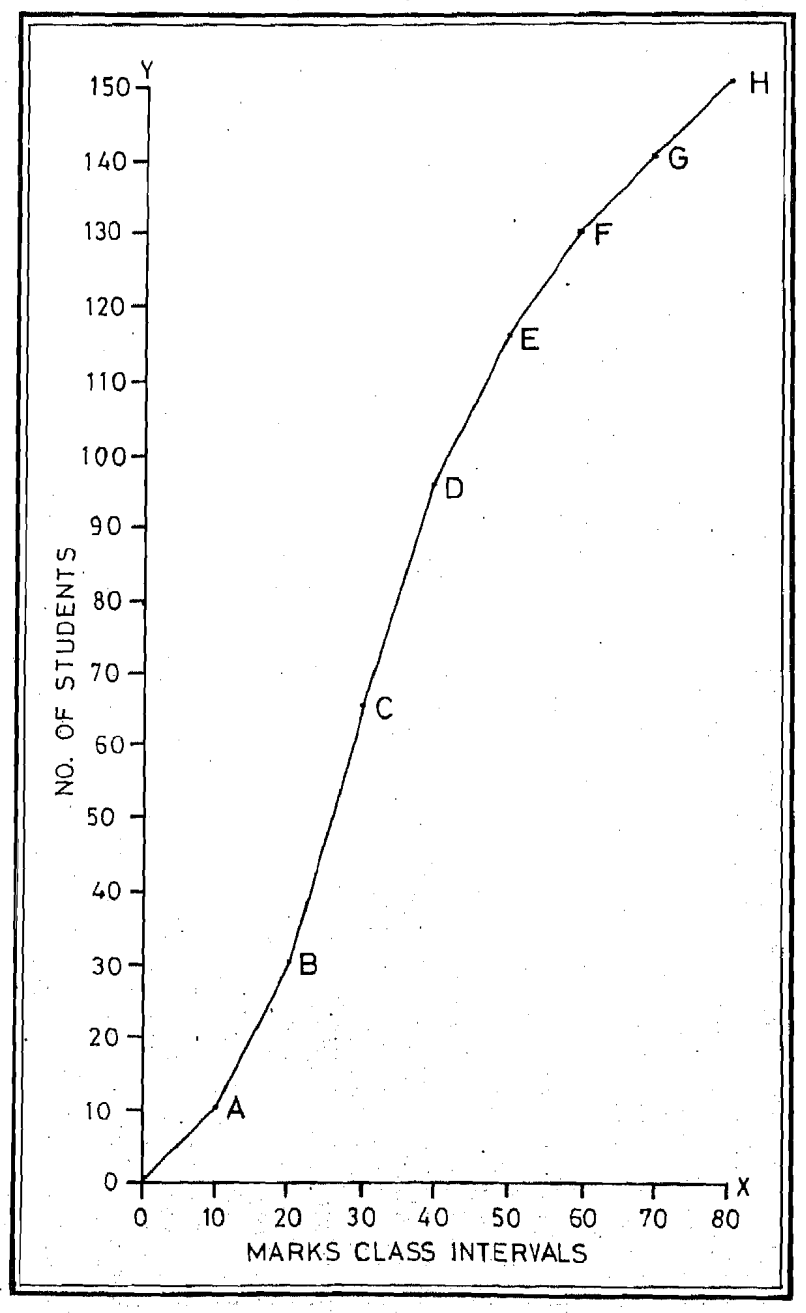
An ogive or a cumulative frequency graph is constructed on the basis of cumulative frequencies. We have already learnt that the cumulative frequency distributions can be either in the ascending order or in the descending order. Similarly, an ogive can also be either a "less than ogive" or a "more than ogive". **A "less than ogive" is** constructed on the basis of cumulative frequencies which are in the ascending order. Similarly, a **"more than ogive" is constructed** on the basis of cumulative frequencies which are in the descending order. In a "less than" ogive, the less than type cumulative frequencies of a class are plotted against its upper limit. In case of a "more than" ogive, the more than type cumulative frequencies of a class are plotted against its lower limit. Ogive helps us to determine median, quartiles, percentages, etc.

Illustration 13

The following data relates to the marks secured by the students of a class in Accountancy. **Draw** "less than" and "more than" ogives in separate graphs. Also show the two ogives in the same graph.

Marks	:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Students	:	10	20	35	30	20	15	10	10

Graph 9.14 : Less than Ogive Showing *the* Marks *of* Students in Accountancy



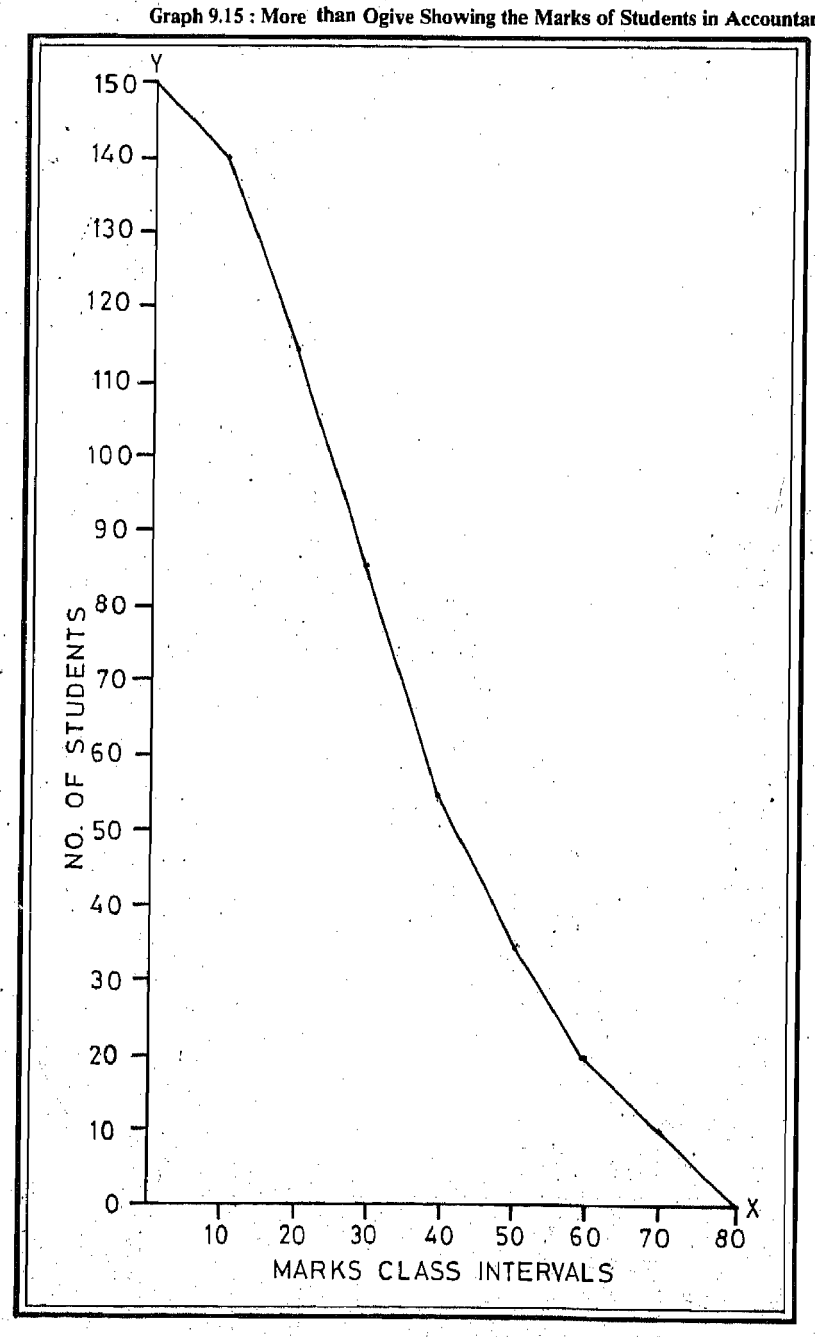
Solution

First, we have to convert the above distribution into "less than" and "more than" cumulative frequency distributions. The two cumulative frequency distributions have been constructed below :

Construction of Cumulative Frequency Distribution

Marks	Frequency	"Less Then" Cumulative Frequencies	"More Than" Cumulative Frequencies
0-10	10	10	150
10-20	20	30	140
20-30	35	65	115
30-40	30	95	85
40-50	20	115	55
50-60	15	130	35
60-70	10	140	20
70-80	10	150	10

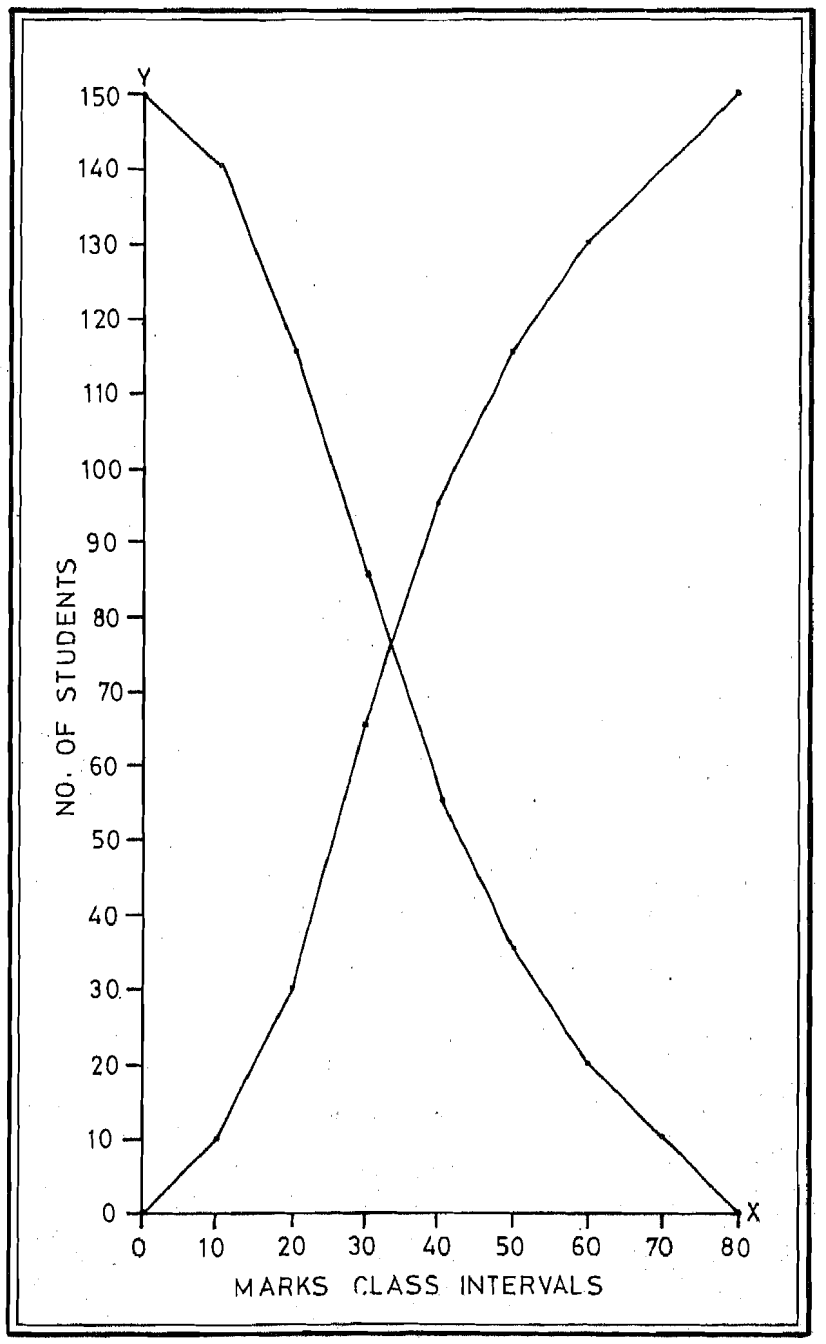
Graph 9.15 : More than Ogive Showing the Marks of Students in Accountancy



Now study Graph 9.14 carefully. To draw less than ogive, less than cumulative frequencies are plotted at upper limits of the various class intervals. So the points (10,10), (20,30), (30,65), (40,95), (50,115), (60,130), (70,140), and (80,150) are plotted as points A, B, C, D, E, F, G, and H. Every plotted point represents the number of students having marks below a particular upper limit of the class interval. If you look at the limits of class intervals, you will find that the points have been plotted for all the values except the lower limit of the first class interval (i.e., '0'). There are no students having marks below 0. So to plot points for all the given class limits, the point (0,0) is also plotted as point P. Now all the points are joined by straight lines.

Study Graph 9.15 carefully. The more than ogive is drawn by plotting points corresponding to more than cumulative frequencies against the lower limits of the various class intervals and the (i.e., 80). Look at Graph 9.16 which shows both the "less than" and "more than" ogives in the same graph.

Graph 9.16 : Less than Ogive and More than Ogive Showing the Marks of Students in Accountancy



You will notice from Graph 9.16 that the "more than ogive" and "less than ogive" intersect at a point which corresponds to 50%of the frequency. In this case that point is 150/2 or 75 frequency.This will be always the case.

Check Your Progress B

1) Distinguish between a histogram and a historigram.

.....

.....

.....

.....

2) What is a frequency polygon?

.....

.....

.....

.....

3) How do you adjust the frequencies for constructing a histogram, when the class intervals have unequal width?

.....

.....

.....

.....

4) What is a frequency curve?

.....

.....

.....

5) Distinguish between a "less than" and "more than" ogives?

.....

.....

.....

.....

6) State whether the following statements are True or False.

i) In a frequency distribution graph, value of variable is shown on X axis,

ii) False base line can also be taken in the graphs of frequency distribution.

iii) A histogram is a series of rectangles, each proportionate in width to the magnitude of class interval and proportionate in area to the number of frequencies.

iv) A histogram in case of class intervals with unequal width can be prepared without making any adjustment in the frequencies.

v) In a frequency polygon, the frequencies are plotted against the upper limit of the class interval.

vi) A frequency polygon is extended on both sides on the left to a class before the first class interval and on the right to a class after the last class interval.

vii) A frequency curve is constructed to smoothen the frequency polygon.

viii) A frequency curve is prepared by free hand.

7) Fill in the blanks with the appropriate words given in the brackets,

i) The frequencies of a distribution are shown on ..... in case of a graph of frequency distribution. (X axis/Y axis)

ii) A histogram helps is determining the value of ..... (mode/median)

iii) The frequencies in case of a histogram are represented by ..... (volume/area)

iv) The area under the histogram and the frequency polygon .....the same. (is/is not)

v) An ogive is constructed on the basis of ..... frequencies. (given/cumulative)

vi) A "less than" ogive is constructed on the basis of cumulative frequencies being in ..... order. (ascending/descending)

vii) An ogive helps in locating the value of ..... (median/mode)

9.8 LET US SUM UP

Graphic presentation renders comparison of data easier, helps in establishing trends of past performance and makes it possible to determine positional averages. Graphs are prepared on the basis of coordinated system of plotting points and joining them by lines. Graphs are of two types : 1) graphs of time series, and 2) graphs of frequency distribution.

Graphs of time series, also called Historigram, depicts chronological data. Time, being the independent variable, is always taken on X axis and the dependent variable is taken on Y axis. The Y axis or vertical axis normally starts with zero but can also be broken and a false base line can be taken.

This type of graph can either be prepared for one dependent variable or more than one dependent variable. A historigram can also be prepared for negative values. In case two dependent variables arc given with two different units of measurement, the data can be represented by means of mixed graphs. Two vertical axes are taken in this case and two curves are drawn on the bases of these axes for the two variables. Another type of historigram is a range graph which is prepared to show the range of data. This graph is prepared to show how the extreme values (i.e., maximum and minimum values) of a dependent variable are changing with time.

Graphs of frequency distribution represent the data of frequency distribution. In this type of graph the value of a variable is taken on X axis and the values derived from frequencies on Y axis. Vertical axis is not broken in this type of graph. Graphs of frequency distribution are classified as : 1) histograms, 2) frequency polygon, 3) frequency curve, and 4) ogive.

A histogram is a series of rectangles each proportionate in width to the magnitude of class interval and proportionate in area to the frequency pertaining to that class interval. When class intervals are all equal, the height of the rectangles will also be numerically proportionate to the frequency. In case the class intervals are unequal the height will be proportional to frequency density.

A frequency polygon is constructed by plotting the frequencies density against the mid points of classes. It can also be prepared first by preparing a histogram and then by putting dots against the mid points and then joining these dots by straight lines. It is extended on both the sides of histogram. The area under a histogram and a frequency polygon is always the same. A frequency curve is prepared by free hand to smoothen the frequency polygon. It is generalisation of frequency polygon.

An ogive or a cumulative frequency graph depicts cumulative frequencies. We can either prepare a "less than" ogive when cumulative frequencies are in ascending order or a "more than" ogive, when cumulative frequencies are in descending order. An ogive helps us to locate median and other partition values.

9.9 KEY WORDS

Adjusted Frequency or Frequency Density : Represents frequency per unit class interval.

False Base Line : Js taken by breaking y-axis in case of a historigram.

Frequency Curve : A curve constructed to smoothen the frequency polygon.

Frequency Polygon : A graph of frequency distribution constructed by plotting the frequencies density against the mid-points of the class.

Histogram : A graph of frequency distribution where rectangles are awn with area proportionate to the frequency of a class interval and the class interval as the base.

Historigram : A graph of time series.

Mixed Graph : A graph constructed to show the two dependent variables with two different units of measurement.

Ogive : A graph of frequency distribution depicting cumulative frequencies.

Originating Point : The point of Inter-section of X **axis** and Y axis.

Range Graph : A graph showing the range of data between two extreme values of a variable at different points of time.

X **axis** : The horizontal axis for plotting points.

Y **axis** : The vertical axis for plotting points.

9.10 ANSWERS TO CHECK YOUR PROGRESS

A) 5) i) True ii) False iii) True iv) False v) False
vi) True vii) False viii) False

6) i) helps ii) IV iii) X iv) can v) can vi) two vii) be viii) difference

B) 6) i) True ii) False iii) True iv) False. v) False
vi) True vii) True

7) i) Y-axis ii) mode iii) area iv) is v) cumulative vi) ascending vii) median

9.11 TERMINAL QUESTIONS/EXERCISES

Questions

1) Describe the rules of graphic presentation of data.

2) Discuss the importance of graphic presentation of data.

3) What is a graph of time series? Discuss the principles of constructing a graph of time series.

4) What is a graph of a frequency distribution? Discuss the principles of construction of graphs of frequency distribution.

5) Describe the different types of graphs based on frequency distribution.

6) Explain different types of graphs of time series.

Exercises

I) The following data relates to the number of man-days lost in a workshop on account of power failure. Show it by means of a suitable graph.

Years	:	1980	1981	1982	1983	1984	1985
Man-days Lost	:	1900	1588	1469	1461	1927	1011

2) Draw a historigram for the following data relating to the financial outlay on education during the Various Five Year Plans.

Plans	:	I	II	III	IV	V	VI	VII
Percentage Outlay		7.6	5.9	6.9	4.9	3.3	2.6	3.6

3) The following data relates to the population of India :

Years	1931	1941	1951	1961	1971	1981
Population(In crores)	27.9	31.9	36.1	43.9	54.8	68.5

Draw a suitable graph for the above data.

4) Draw a suitable graph for the following data relating to the balance of trade of a country :

Years	:	1981	1982	1983	1984	1985	1986	1987	1988
Balance of Trade ('000 Rs)		+3	+5	+1	-4	-2	+3		+6

5) The following data relates to the occupational distribution of working population in India:

Year	Primary (%)	Secondary (%)	Tertiary (%)	Total Population
1941	76.0	10.5	13.5	100
1951	72.1	10.7	17.2	100
1961	—	—	—	—
1971	72.1	11.2	16.7	100
1981	70.6	12.9	16.5	100

Draw a more than one dependent variable graph for the above data.

6) Draw a mixed graph for the following data relating to the production of a commodity in terms of quantity and value.

Years	1982	1983	1984	1985	1986	1987	1988
Quantity ('000) :	20	24	24	26	25	28	30
Value (lakh Rs.) :	1000	1100	1200	'1350	1350	1500	1800

7) The following data relates to the temperature in Delhi. Draw a range graph for this data.

Date	(Celsius) Maximum Temperature	Minimum Temperature
1-11-1989	32.5	16.5
2-11-1989	31.6	16.0
3-11-1989	30.2	16.0
4-11-1989	29.5	15.5
5-11-1989	30.5	15.0
6-11-1989	31.6	14.5
7-11-1989	30.0	14.5

8) Prepare a histogram for the data given below :

No, of Goals Scored	0	1	2	3	4	5	6
No. of Matches	8	12	20	10	6	4	2

9) Prepare a histogram for the following data :

Size	1	2	3	4	5	6	7	8
Frequency	5	15	20	30	25	20	10	10

10) Prepare a histogram for the data given below :

Age (in Years)	4-6	6-8	8-10	10-12	12-14	14-16	16-18
No. of Children	SO	80	140	200	300	180	40

11) The following data relates to the marks scored by the students of a college in economics. Draw a mixed graph for this data.

Marks	0-10	10-20	20-40	40-50	50-60	60-70	70-100
No. of Students	10	15	40	30	50	30	15

12) Draw a histogram, a frequency polygon and a frequency curve for the following data :

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Frequency	5	20	40	50	30	10	5

13) Prepare a frequency polygon for the following data :

Mid Point	5	15	25	35	45	55	65
Frequency	20	40	50	80	60	30	10

14) Draw separately a "less than" and a "more than" ogive for the following data.

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Frequency	5	20	40	50	30	10	5

15) Prepare a "less than" and a "more than" ogive for the following data in one graph :

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-100
No. of students :	20	40	80	150	250	100	60	10

Note : These questions and expressions will help you to understand the unit better. Try to write answers for them. But do not send your answers to the University for evaluation. These are for your practice only.

SOME USEFUL BOOKS

Elhance, D.N., and Veena Elhance, 1988, *Fundamentals of Statistics*, Kitab Mahal : Allahabad. (Chapters 3-6)

Gupta, C.B., 1983, *An Introduction to Statistical Methods*, Vikas Publishing House : New Delhi. (Chapters 4-8)

Gupta, S.P., 1989, *Elementary Statistical Methods*, Sultan Chand & Sons : New Delhi. (Chapters 3-6)

Sancheti, D.C., and Kapoor, V.K., 1989. *Statistics Theory, Methods and Applications*, Sultan Chand & Sons : New Delhi. (Chapters 2-4)

Shenoy, G.V. Srivastava V.K., and Sharma, S.C., 1989, *Business Statistics*, Wiley Eastern : New Delhi. (Chapters 2-3)

Simpson, G., and Kafka, F. *Basic Statistics*, Oxford & IBH Publishing : New Delhi (Chapters 3, 6-9)

ECO-07 ELEMENTS OF STATISTICS
Course Components

BLOCK	UNIT NO.	PRINT MATERIAL
1		Basic Statistical Concepts
	1	Meaning and Scope of Statistics
	2	Organising a Statistical Survey
	3	Accuracy, Approximation and Erros
	4	Ratios, Percentages and Rates
2		Collection, Classification and Presentation of D
	5	Collection of Data
	6	Classification of Data
	7	Tabular Presentation
	8	DiagrammaticPresentation
	9	Graphic Presentation
3		Measures of Central Tendency
	10	Concept of Central Tendency and Mean
	11	Median
	12	Mode
	13	Geometric, Harmonic and Moving Averages
4		Measures of Dispersion and Skewness
	14	Measures of Dispersion-1
	15	Measures of Dispersion-I1
	16	Measures of Skewness

UNIT 10 CONCEPT OF CENTRAL TENDENCY AND MEAN

Structure

10.0 OBJECTIVES

After studying this unit, you should be able to :

- describe what is central tendency
- appreciate the purpose of calculating averages
- enumerate the qualities of an ideal average
- define and compute the arithmetic mean and the weighted arithmetic mean for different types of data
- explain the properties and merits of mean
- state the limitations and uses of mean.

10.1 INTRODUCTION

You have studied in detail how the data is to be classified and presented in the form of tables, diagrams and graphs. If the characteristics of the data are to be properly understood, it is necessary to summarise and analyse the data further. The first step in that direction is the computation of Average or Central Tendency, which gives a bird's-eye view of the entire data.

In this unit you will study the purpose of calculating averages and the essentials of an ideal average, and identify different measures of averages. You will further learn in detail the calculations, merits, and limitations of two measures of averages, viz. Arithmetic Mean and Weighted Arithmetic Mean.

10.2 CONCEPT OF CENTRAL TENDENCY

For a proper appreciation of various statistical measures used in analysing a frequency distribution, it is necessary to note that most of the statistical distributions have some common features. If we move from lowest value to the highest value of a variable, the number of items at each successive stage increases till we reach a maximum value, and then as we proceed further they decrease. The statistical data which follow this general pattern

may differ from one variable to another in the following three ways:

1) They may differ in the values of the variables around which most of the items cluster (i.e., Average)
2) They may differ in the extent to which items are dispersed (i.e., Dispersion).
3) They may differ in the extent of departure from some standard distributions called normal distribution (i.e., Skewness and Kurtosis).

Accordingly, there are three sets of statistical measures to study these three kinds of characteristics. At present, however, we are confined to the first set of measures which are called Averages or Measures of Central Tendency or Measures of Location. We discuss about the other two sets of measures (i.e., measuresof dispersion and skewness) in Block 4 in this course.

In the general pattern of distribution, in the data we may identify a value around which many other items of the data congregate. This is a value which is somewhere in the central part of the range of all values. When this typical item of the data is towards the central part of the data, it is known as Central Tendency. As it indicates the location of the clustering of items, it is also called a measure of location. Just as the title of an essay gives the central theme of the essay, the central tendency of the numerical data gives the central idea of the entire data. Look at Figure 10.1 carefully. It shows the central locations of three different curves A, B and C. You must have noticed that the central locations of curve A and curve C are equal. The central location of curve B lies to the right of those of curves A and C.

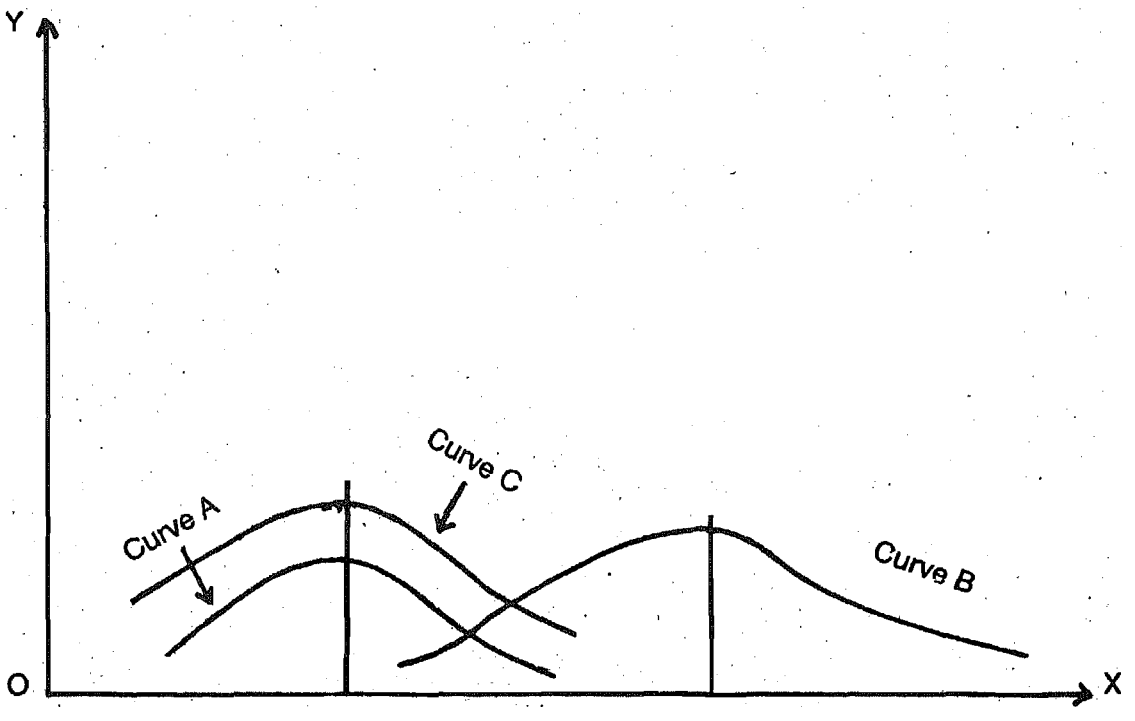


Figure 10.1 Central Location of Different Curves

10.3 ESSENTIALS OF AN IDEAL AVERAGE

As suggested by the eminent statisticians Yule and Kendall, an ideal average should possess the following characteristics:

1) Easy to understand and simple to compute: It should be easy to make out an average and its computation should also be simple.

2) Rigidly defined: An average should be rigidly defined by a mathematical formula so that the same answer is derived by different persons who try to compute it. It should not depend on the personal prejudice or bias of a person computing it.

3) Based on all items in the data: For calculating an average, each and every item of the data set should be included. Not a single item should be dropped, otherwise the value of the average may change.

4) **Not to be unduly affected by extreme items: A** single extreme value i.e., a maximum value or a minimum value, can unduly affect the average. A too small item can reduce the value of an average, and a too big item can inflate its value to a large extent. If the average is changing with the inclusion or exclusion of an extreme item, then it is not a truly representative value of the data set.

5) **Capable of further algebraic treatment :** An average should be amenable to further algebraic treatment. That should add to its utility. For example, if we are given the averages of three data sets of similar type, it should be possible to obtain the combined average of all those three data sets.

6) **Sampling stability:** The average should have the same 'sampling stability'. This means that if we take different samples from the aggregate, the average of any sample should approximately turn out to be the same as those of other samples.

10.4 OBJECTIVES OF AVERAGES

You have studied the features of an ideal average. Now let us discuss the major objectives of computing averages. The following are the main objectives of averages.

1) **To supply one single value that describes the characteristics of the entire data :** An average reduces the complex mass of data into a single representative value which enables us to grasp the salient features of data, without getting lost in its details. Thousands or lakhs of values can be, thus, represented by a single value. For example, it is almost impossible to remember monthly salary of each and every worker of a big factory. But if the average salary is obtained by dividing the total pay bill of all the workers by the number of workers, it enables us to know, on an average, how much the worker is getting.

2) **To facilitate comparison:** It is not easy to compare the two sets of huge raw data. But the two different data sets could be easily compared by working out their averages. Comparison can be made either at a point of time or over a period of time. For example, the current year sales of two business firms A and B can be compared by comparing their average sales: The current year sale of a unit can be compared with its own sales in the previous year by working out the average sales during the previous year and the current year's average. Moreaver, the same measure of average should be used for comparing the average of two data sets, the same method of computation should be followed. For example, comparing the mean income of the people of one locality with the median income of the people of another locality is not reasonable. We will discuss in detail about mean later in this unit and about median in Unit 11.

3) **To facilitate statistical inference:** To draw inferences about the unknown measures or 'parameters' of the population, we depend on values calculated from sample. This process is known as **statistical inference.** An average obtained from a sample is helpful in estimating the average of the population.

4) **To help the decision-making process:** The averages are computed to help the managers in decision-making. The managers are often interested in knowing normal output of aplant, representative sales volume, overall productivity index, price index, etc. These all are the connotations of an average.

10.5 DIFFERENT MEASURES OF CENTRAL TENDENCY

Following are the various measures of averages or central tendency:

1) Mathematical Averages
   i) Arithmetic Mean
   ii) Geometric Mean
   iii) Harmonic Mean

All these measures can be either simple or weighted. You will study in detail about Arithmetic Mean later in this unit. Geometric and Harmonic Means are discussed in Unit 13.

2) Averages of Position
   i) Median
   ii) Mode

We discuss in detail about Median in Unit 11 and Mode in Unit 12.

3) Special Averages
   i) Moving Average
   ii) Progressive Average

   These special averages are commonly used in the analysis of the time series data relating to business. In Unit 13 we study in detail about the moving averages.

10.6 WHAT IS ARITHMETIC MEAN?

The arithmetic mean is commonly known as mean. It is a measure of central tendency because other figures of the data congregate around it. Arithmetic mean is obtainkd by dividing the sum of the values of all observations in the given data set by the number of observations in that set. It is the most commonly used statistical average in the disciplines such as commerce, management, economics, finance, prodaction, etc. The arithmetic mean is also called as simple Arithmetic Mean.

10.7 COMPUTATION OF ARITHMETIC MEAN

As you know, the collected data is classified by arranging into different classes or groups on the basis of their similarities and resemblances. Arithmetic mean can be computed for the unclassified or ungrouped 'data (raw data) as well as classified or grouped data. But the methods of computation are different. Now let us understand the methods of computing the arithmetic mean for unclassified data and classified data. Normally, arithmetic mean is denoted **by** $\bar{x}$ which **is read** as **'X bar'**

10.7.1 Ungrouped .Data

Method **1 :** Computation of arithmetic mean is very simple when the data is ungrouped, **i.e.** when frequency distribution is not done. Just add all the values of the observations and divide it by the number of observations. This can be explained and expressed in the form of a formula as follows:

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n}$$

Where $\bar{x}$ (X bar) is the arithmetic mean of the variable x
$x_1, x_2, \ldots, x_n$ are the various values of the variable x
n is the number of observations

This formula can be simplified as follows:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Where the Σ (read it as sigma) is the Greek symbol denoting the summation over all values of x.

Illustration **1**
The grocery store sells five different products. The profit per unit on the sales of each of these products is given below. Find out the average profit.
Product 1 - Rs. 4
Product **2** - Rs. 9
Product **3** - Rs. 6
Product 4 - Rs. 2
Product 5 - Rs. 9

Solution
Average profit can be computed as follows:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

$$= \frac{4+9+6+2+9}{5}$$

$$= \frac{30}{5}$$

= Rs. **6.00**

Method 2: When the values of the observations in the given data are too large or they are in fractions, this method may be followed. This method is based on the fact that the algebraic sum of the deviations of a series of individual observations from their mean is always equal to zero. For example, the arithmetic mean of **8, 14, 16, 12** and **20** is **14.** The difference of each of these items from the mean would be **−6, 0, +2, − 2, +6** and their total is zero. This is true always. To compute arithmetic mean under this method, the following steps are to be followed.

1) Assume any arbitrary mean (A) to find out the deviations of items from their assumed mean.

2) Compute the deviation (d) of each individual value (x) from the assumed mean **i.e., d = x − A.**

3) Obtain the sum of all deviations (I d called sigma d)

4) Compute the arithmetic mean by using the following formula:

$$\bar{x} = A + \frac{\Sigma d}{n}$$

where $\bar{x}$ is the arithmetic mean of the variable x
A is the assumed mean
Id is the sum total of the deviations of each individual value from the assumed mean
n is the number of observations

Illustration 2

Monthly sales of scooters of **10** dealers is presented below. Calculate the average sales per month:

Dealer :	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sales :	**23**	8	14	**'31**	**6**	**28**	**11**	**27**	**32**	**46**

Solution

Calculation of Arithmetic Mean

Dealer	Sales (x)	d = x − A
1	23	−2
2	8	−17
3	14	−11
4	31	6
5	6	−19
6	28	3
7	11	−14
8	27	2
9	32	7
10	46	21
n = 10		Σd = −24

Assumed mean A = **25**
Σd = −24
n = 10

$$\bar{x} = A + \frac{\Sigma d}{n}$$

$$= 25 + \frac{-24}{10}$$

$$= 25 - 2.4$$

$$\bar{x} = 22.6.$$

Average scooters sold = 22.6

10.7.2 Grouped Data

As studied in Unit 6, variables can be categorised as discrete variables and continuous variables. The frequency distribution prepared for discrete variable is called discrete distribution and the frequency distribution prepared for continuous variable is called continuous distribution. Methods of computing arithmetic mean for these two types of distributions are different. Now let us study these methods.

Arithmetic Mean for Discrete Series:

Method 1: Under this method the mean for grouped data can be obtained by using the following formula:

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{f_1 + f_2 + ... + f_n} = \frac{\Sigma fx}{\Sigma f}$$

where, x_1, x_2, x_3 etc., refer to the values of the variable in classes 1, 2, 3 etc., respectively. Similarly, f_1, f_2, f_3 etc., refer to the frequency of classes 1, 2, 3 etc., respectively. Here $f_1 x_1$ indicates the multiplication of the frequency of the first class (f_1) by the value of the variable in that class (x_1). f_2x_2, f_3x_3 .... f_nx_n indicate the same meaning. Similarly, Σf is the sum total of f, to f_n.

Method 2: When the number of classes in the given frequency distribution is large, this method is preferred. The procedure followed in this method is almost the same as it is for ungrouped data. Steps to be followed in this method are as follows:

a) Take an assumed mean **A.**

b) Find the deviations of the variable x from the assumed mean and denote it by $d = x - A$, Any value can be taken as an assumed mean, but the value of variable x in centrally located class of the given distribution should be chosen.

c) Obtain Σfd by multiplying deviations (d) with their respective class frequencies (f) and summing it.

d) Take a ratio of Σfd to Σf, that is, $\frac{\Sigma fd}{\Sigma f}$. If is also called a correction factor.

e) Add this correction factor to the assumed mean to obtain $\bar{x}$

The formula used in computing the arithmetic mean under this method is as follows:

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \quad \text{or} \quad \bar{x} = A + \frac{\Sigma fd}{n}$$

Where A is the assumed mean

Σf denotes the total number of items which can also be denoted by 'n'.

Σfd is the sum total of the deviations ($d = x - A$) multiplied with their respective class frequencies.

Now let us take an illustration and study how arithmetic mean is computed under these two methods.

Illustration 3

Calculate the arithmetic mean for the following data by using the two methods:

Marks :	10	20	30	40	50	60	70	80
No of Students :	8	21	23	17	15	9	5	2

Solution

Calculation of Arithmetic Mean

Marks (x)	No. of students (f)	d = x – 40	fd	fx
10	8	–30	–240	80
20	21	–20	420	420
30	23	–10	–230	690

	40	17	0	0	680
	50	15	10	150	750
	60	9	20	180	540
	70	5	30	150	350
	80	2	40	80	160
	Total	$\Sigma f = 100$		$\Sigma fd = -330$	$\Sigma fx = 3.670$

In this case assumed mean (A) is 40.

Method 1:

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{3,670}{100}$$

$$= 36.70$$

Method 2:

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f}$$

$$= 40 + \frac{-330}{100}$$

$$= 40 - 3.30$$

$$= 36.70$$

Arithmetic Mean for Continuous Series

For continuous series (i.e. when the data is classified according to class intervals), arithmetic mean can be calculated by the following methods:

Method 1: As you know, to find out the arithmetic mean you need the total values of all the items. When data is classified according to class intervals, you do not know the values of all the items. What you know is that the items belonging to various groups are spread out in the respective class intervals. Therefore, for calculating the total value, you assume that all the items of a class interval are uniformly spread out in that group. This means, for calculation purposes you can assume that the values of items belonging to a group are equal to the mid-point of that group. In the case of continuous series, the mid-points of the various class intervals are computed to replace the class intervals. Once it is done, there is no difference between a continuous series and discrete series. After this stage, the method of computing arithmetic mean is same as the method used in the case of discrete series. The two methods followed in the case of discrete series can be used here as well. The methods, however, would be the same for both inclusive class intervals as well as exclusive class intervals. Under this method the arithmetic mean is obtained by using the following formula:

$$\bar{x} = \frac{\Sigma fm}{\Sigma f} \text{ or } \bar{x} = \frac{\Sigma fm}{n}$$

where 'm' is the mid-value of a class: First obtain the product of mid-value; of each class and its corresponding frequencies and then add those products to get Σ fm. Divide it by total of frequency (Σ f).

Method 2: The same formula as used for discrete series can be used here also with a slight change in obtaining 'd'. Here deviations of mid-values from assumed mean are obtained (i.e., d = m–A).

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \text{ or } A + \frac{\Sigma fd}{n}$$

Step-Deviation Method: If the deviations from assumed mean have some common factor, a further reduction in the size of deviations is possible by dividing deviations by the common factor 'c' and denoting these step deviations by d' i.e., d' = (in – A) /c. The mean is then worked out as:

$$\bar{x} = A + \frac{\Sigma fd'}{\Sigma f} \times c \text{ or } \bar{x} = A + \frac{\Sigma fd'}{n} \times c$$

Note: If all class intervals are equal, the class interval will be the common factor.

Illustration 4

Weekly sales of **50** salesmen of a company are given below. Calculate the arithmetic mean by following the step deviation method.

Total Sales (Rs. '000)	: 0-5	5-10	10-25	25-50
No of Salesmen	: 3'	6	25	10

solution

Calculation of Arithmetic Mean

Sales per Week Rs.'000s	No of Salesmen (f)	Mid-point P' (m)	Deviations (m − 17.5)	Step Deviations $d' = \frac{m-17.5}{5}$	fd'
0 - 5	3	2.5	−15	−3	−9
5 - 10	12	7.5	−10	−2	−24
10 - 25	25	17.5	0	0	0
25 - 50	10	37.5	20	4	40
Total	Σf = 50				Σfd' = 7

It is apparent from deviation column that here assumed Mean (A) is **17.5** and the common factor 'c' is **5.**

Now $\bar{x} = A + \frac{\Sigma fd'}{n} \times c$

$= 17.5 + \frac{7}{50} \times 5$

$= 17.5 + 0.7$

$= 18.2$

Average mean of sales is Rs. 18.2 thousands per week.

Illustration 5

Find the average number of hours worked by the employees of the **Yamto** Machine Co. from the data given below:

Houn worked	No. of employees
36.0 - 37.b	6
37.8 - 39.6	7
39.6 - 41A	24
41.4 - 43.2	7
43.2 - 45.0	2
45.0 - 46.8	4
Total	50

Solution .

First obtain the mid-values **(m) of** all the classes **and** take deviations from **assumed mean** 'A' **(i.e.** 42.3). The common factor 'C' is **1.8 which is** equal to the class interval of different groups.

Calculation of Arithmetic Mean

Hours worked	m	f	m–A (m–42.3)	$d' = (m-A)/C \left(d' = \frac{m-42.3}{1.8}\right)$	fd'
36.0 - 37.8	36.9	6	−5.4	−3	−18
37.8 - 39.6	38.7	7	−3.6	−2	− 14
39.6 - 41.4	40.5	24	−1.8	−1	−24
41.4 - 43.2	42.3	7	0	0	0
43.2 - 45.0	44.1	2	+1.8	1	2
45.0 - 46.8	45.9	4	+3.6	2	8
	m = Σf = 50				Σfd' = − 46

$$\bar{x} = A + \frac{\Sigma fd'}{n} \times C$$

$$= 42.3 + \frac{(-46)}{50} \times 1.8$$

$$= 42.3 + (-0.92) \times 1.8$$

$$= 42.3 - 1.656$$

$$= 40.644$$

Arithmetic mean of the hours worked is 40.6 hours.

You may notice when class intervals are all equal, d' values will be 1, 2, 3, ... and –1, –2, –3, ... etc. But when class intervals are not equal, the d' values need not be in numbers in order. In such a case it is necessary to make the column m–A, and then divide it by 'C'. However, when class intervals are all equal, writing of the Column m–A may be avoided and the values of d' may be written directly.

Check Your Progress A

1) Fill in the blanks with appropriate words given in the brackets.

i) An average gives a ..... of the entire data (bird's-eye view/picture)

ii) An average summarises main characteristics of the data and therefore it is also known as a ..... measure. (central tendency/summary)

iii) An ideal average should not be unduly affected by ..... items. (middle/extreme)

iv) An average ..... comparison. (facilitates/does not help)

v) The process of estimating the unknown parameters of the aggregateon the basis of sample values is known as statistical. ..... (study/inference).

2) State whether the following statements are True or False.

i) When whole data is available there is no need for computing central tendency as it will not give any thing more than what is contained in the data.

ii) . Arithmetic mean is a positional average,

iii) In step deviation method of finding arithmetic mean, 'C' always stands for class intervals.

iv) Values of all the items are taken into acdount while calculating arithmetic mean.

v) For a given data, if mean is calculated by different methods they can give different results.

3) i) If the sum of the deviations of 6 items taken from an assumed mean 12 is – 6, find their mean.

.....

.....

.....

.....

ii) Write the formulas for the methods used in computing the arithmetic mean of the grouped data of continuous series.

.....

.....

.....

.....

iii) Wherever possible, step-deviation method should be preferred, why?.

.....

.....

.....

.....

iv) For the given data set if $\bar{x} = 33$, $\Sigma fd' = -20$, $\Sigma f = 100$ and c = 10; find the assumed mean A.

.....

.....

.....

.....

v) What is the major assumption we make while computing a mean from grouped data?

.....

.....

.....

.....

4) The monthly income of twelve families in a town is given below. Calculate the arithmetic mean.

Family	: 1	2	3	4	5	6	7	8	9	10	11	12
Monthly Income Rs.	: 280	180	96'	98	104	75'	80	84	100	75	600	200

.....

.....

.....

.....

.....

.....

.....

.....

5) In 12 consecutive months the number of rejected pieces produced by the operator of a machine was 82, 74, 65, 67, 62, 73, 68, 63, 65, 62, 69, and 66.

i) What was the average number of rejects?

.....

.....

.....

ii) What is the sum of the deviations from this average?

.....

.....

.....

6) Calculate arithmetic average of the following data by using alternative methods:

Weekly wages of workers (Rs.)	No of workers
100 - 105	200
105 - 110	210
110 - 115	230
115 - 120	320
120 - 125	350
125 - 130	320
130 - 135	410
135 - 140	320
140 - 145	280
145 - 150	210
150 - 155	160
155 - 160	90

7) Find the mean from the following distribution by step deviation method.

Class Interval :	15-25	25-35	35-45	45-55	55-65	65-75
Frequency :	4	11	19	14	0	2

10.8 WEIGHTED ARITHMETIC MEAN

You have studied various methods of computing arithmetic mean for different types of data sets. In all these methods we presume that all the items of the given data set have equal importance. But it is not necessarily true in all situations. In practical situations some items are of greater importance than the others. For example, while constructing the cost of living index for a particular class, the commodities they consume have varying importance. The simple arithmetic mean of the prices of such commodities will not depict a true picture of their living pattern. Different commodities are to be assigned weights and a weighted arithmetic mean is to be worked out in such situations. In a factory where unit cost of manufacturing is to be worked out, a weighted average is more appropriate.

10.8.1 Computation of Weighted Arithmetic Mean

To compute weighted arithmetic mean, different values of the variable x (viz. $x_1, x_2, ..., x_n$) are assigned different weights (viz. $w_1, w_2, ..., w_n$) respectively. These values are multiplied by their respective weights. The products so arrived are added and a total Cwx is obtained. It is then divided by the total of weights (Cw) and the resulting figure is the weighted arithmetic mean.

The main difficulty in the computation of weighted arithmetic mean is with regard to selection of weights. These weights may be either actual or estimated. If actual weights are available, they must be used. If they are not available, some arbitrary weights may be assigned depending upon the situation.

Illustration6

Prices of three commodities viz., A, B & C rised by 40 %, 60 % and 90 % respectively. Commodity A is six times more important than C, and B is three times more important than C. What is the mean rise in price of these three commodities?

Solution

As the mean rise in price is to be determined, the figures of rise in price will be denoted as x. The relative importance of A: B:C is 6:3:1. So these figures will be taken as weights 'w'.

Commodity	Percentage rise in prices (x)	Weights (w)	wx
A	40	6	240
B	60	3	180
C	90	1	90
Total	–	$\Sigma w = 10$	$\Sigma wx = 510$

$$\text{Weighted Arithmetic Mean} = \frac{\Sigma wx}{\Sigma w}$$

$$= \frac{510}{10}$$

$$= 51\ \%$$

Mean rise in the prices is 51 %.

It may be noted that for computation purpose, weights of items are treated in the same way as the frequencies of the items. In fact weights are not frequencies. Frequency means number of times an item is repeated in the data, whereas weights only give the relative importance of various items. The items actually occur only once in the data.

Weighted arithmetic mean is also called ***Weighted Average.*** The word 'Average' in statistics, as pointed out earlier, is also used for other measures of central tendency viz., geometric mean, harmonic mean, etc. So, in broader sense, weighted average also includes weighted geometric mean and weighted harmonic mean (about these two you will learn in detail in Unit 13).

10.8.2 Comparison with Simple Arithmetic Mean

Weighted arithmetic mean differs from simple arithmetic mean because we use weights in the former case. Inter-relationship between weighted mean and simple mean is as follows:

1) If all items are given equal importance, weighted mean will be equal to simple mean.

2) If large items are given large weights and small items given small weights, then weighted mean is greater than simple mean.

3) If large items are given small weights and small items given large weights, then weighted mean is less than simple mean.

Illustration 7

To understand this inter-relationship clearly, let us take up some illustrations. Let us take Illustration 6 once again and find out mean rise in price by taking the following two sets of weights.

A : B : C as 1 : 3 : 6 set w_1

A : B : C as 10 : 10 : 10 set w_2

Solution

Calculation of Weighted Arithmetic Mean

Commodity	%rise	Set 1		Set 2	
	x	w_1	xw_1	w_2	xw_2
A	40	1	40	10	400
B	60	3	180	10	600
C	90	6	540	10	900
Total	$\Sigma x = 190$	$\Sigma w_1 = 10$	$\Sigma xw_1 = 760$	$\Sigma w_2 = 30$	$\Sigma xw_2 = 1900$

1) Weighted Mean for Set 1 $= \frac{\Sigma xw_1}{\Sigma w_1} = \frac{760}{10} = 76\%$

2) Weighted Mean for Set 2 $= \frac{\Sigma xw_2}{\Sigma w_2} = \frac{1900}{30} = 63.3\%$

3) Simple Mean = $\frac{\Sigma x}{n} = \frac{190}{3} = 63.3$ %

If we compare the results carefully, we can notice the following points:

i) Under weights Set 2, all commodities are given equal weights. Here weighted mean (63.3) is equal to simple mean (63.3).

ii) Under weights Set 1, large value 90 is given a large weight 6 and small item 40 is given small weight 1. Here weighted mean (76) is greater than simple mean (63.3).

iii) Under the original set of weights (look at Illustration 6) large value 90 was given a small weight 1 and small value 40 was given a large weight 6. In that case weighted mean (51) was less than simple mean (63.3).

These three properties of weighted average (as they are true for all kinds of weighted averages) point out the following important fact. The weighted mean is not only the mean of items, but also it gives the average of two things : (i) average of items, and (ii) how items are affected by the pattern of weighting. Thus, when items are of unequal importance, calculation of weighted average is a must for finding out proper average.

10.8.3 Uses of Weighted Arithmetic Mean

Weighted arithmetic mean is mainly useful under the following situations:

1) When the given items are of unequal importance

2) When averaging percentages which have been computed by taking different number of items in the denominator

3) When statistical measures such as mean of several groups are to be combined

To be more specific, weighted arithmetic mean is used in the following cases :

1) Construction of Index Numbers.

2) Computation of standardised birth and death rates.

3) Finding out an average output per machine, where machines are of varying capacities.

4) Determining the average wages of skilled, semi-skilled and unskilled workers of a 'factory.

10.9 PROPERTIES OF ARITHMETIC MEAN

You have studied the meaning and methods of computing the arithmetic mean. You have also studied how a weighted arithmetic mean is different from simple arithmetic mean. Now let us study the main properties of arithmetic mean.

1) The sum of the deviations of the iqdividual items from the arithmetic mean is always zero i.e., $\Sigma (x - \bar{x}) = 0$. This is explained in the following illustration.

x	$(x-\bar{x})$
5	−1
6	0
7	1
9	3
3	−3
30	$\Sigma(x-\bar{x}) = 0$

$\bar{x} = \Sigma x/n$
$= 30/5$
$= 6$

In this illustration you should note that the sum of positive deviations from the mean is equal to the sum of negative deviations. Precisely, therefore, mean is also known as the centre of gravity. This is true for all kinds of data with class intervals or without class intervals.

2) The sum of the square of deviations from the arithmetic mean is minimum **i.e.,** it is always less than the sum of squares of deviations of the items taken from any other value. In other words, $\Sigma(x-\bar{x})^2$ is always minimum. We can verify this for the illustration discussed above.

Squared Deviations taken horn mean $(\bar{x}=6)$			Squared deviations taken from any other values say 5		
x	$(x-\bar{x})$	$(x-\bar{x})^2$	x	$(x-5)$	$(x-5)^2$
5	–1	1	5	0	0
6	0	0	6	1	1
7	1	1	7	2	4
9	3	9	9	4	16
3	–3	9	3	–2	4
		20			25

It is clear that $\Sigma(x-\bar{x})^2 < \Sigma(x-5)^2$

3) If the number of items and mean are known, the total of the items can be obtained by multiplying the mean by the number of items, **i.e.,** $\Sigma x = nx$, where 'n' is the number of items.

This property has a great practical significance. For example, if we know the number of workers in a factory, say 100, and average monthly wage is Rs. 400, we can easily obtain the total monthly wage bill as Rs. 400×100 = Rs. 40,000.

4) If we add or delete an observation which is equal to mean, the arithmetic mean remains unaffected.

5) If each of the values of a variable **'x'** is increased or decreased by some constant C, the arithmetic mean also increases or decreases by C. Similarly, when the values of a variable 'x' are multiplied by a constant, say k, the arithmetic mean is also multiplied by the same quantity k.

For example, take the previous illustration, and add 2 to each observation and multiply each of them by **3,** the new mean will be: (original mean + 2) $\times$ **3** = (6 + 2) x 3 = 24. Let us verify it.

x	x + 2	3 (x + 2)
5	7	21
6	8	24
7	9	27
9	11	33
3	5	15
30	40	120

Mean of $x = 3015 = 6$
Mean of $x + 2 = 40/5 = 8 = 6 + 2$ i.e., old mean + 2
Mean of 3 $(x + 2) = 12015 = 24$ or 8 x 3 or $(6 + 2) \times 3$ i.e., (old mean + 2) x 3.

6) If we have the arithmetic mean and number of items of two or more related groups, we can have a combined mean of these groups as follows :

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_1\bar{x}_2}{n_1 + n_2}$$

where x_1 and x_2 are the arithmetic means of group 1 and group 2 respectively, and n, and n_2 are the number of items in group 1 and group 2 respectively.

For example, arithmetic mean of the production of a commodity during the period January to August is 400 tonnes per month, and the arithmetic mean for the period September to December is 430 tonnes per month. Now we can compute the mean production for the whole year as follows:

$\bar{x}_1 = 400$
$\bar{x}_2 = 430$
$n_1 = 8$ (January to August - 8 months)

n_2 = 4 (September to December - 4 months)

The average for the whole year $\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

$$= \frac{8 \times 400 + 4 \times 430}{8 + 4}$$

$$= \frac{4920}{12}$$

= 410 tonnes per month.

The logic behind the formula is as follows:
$n_1\bar{x}_1$ is the total value of all the items belonging to the first group and $n_2\bar{x}_2$ is the total for the second group. Thus, $n_1\bar{x}_1 + n_2\bar{x}_2$ is the total of all the items in both the groups. In other words, the combined mean is the weighted average of the means of different groups, weights being the number of items in each group.

Check Your Progress B

1) Distinguish between weighted arithmetic mean and simple arithmetic mean.

.....

.....

.....

.....

.....

2) Calculate the simple mean and weighted mean of price from the following and state the reasons for the difference between the two.

Price per tonne (Rs.)	: 45.60	40.70	42.45
Tonnes purchased	: 135.00	40.00	25.00

.....

.....

.....

.....

.....

.....

.....

3) From the results of two colleges A and B, state which of them is better.

Name of the Exam.	College A		College B	
	Appeared	Passed	Appeared	Passed
M.A.	30	25	100	80
M.Com.	50	45	120	95
B.A.	200	150	100	70
B.Com.	120	75	80	50
Total	400	295	400	295

.....

.....

.....

.....

.....

.....

.....

.....

.....

4) The marks of a student in written and oral tests in subjects A, B and C are as follows :

Subject	A	B	C
Written (out of 75 marks)	43	32	29
Oral (out of 25 marks)	15	12	18

.....

.....

.....

.....

.....

.....

Find out the mean marks in written examinations taking the percentage of marks in oral as weights.

5) State whether the following statements are True or False.

i) If large items are given large weights, simple mean will be larger than weighted mean.

ii) Under certain conditions, simple mean can also be taken as weighted mean.

iii) Weighted average, if exists, is always a better measure than simple average.

iv) Assigning 'O' weight to an item means that, that item is excluded from calculating weighted average.

v) Weighted average can be found by picking up the most important item of the data.

6) Fill in the blanks with the appropriate words given in the brackets.

i) In the construction of Index Numbers ..... mean is specially used. (weighted/unweighted)

ii) If the weights are not given ..... weights may be used. (no/arbitrary)

iii) The total wage bill to a certain number of workers is Rs. 5,000. Each worker on an average gets Rs. 250, then the number of workers is ..... (30/20)

iv) The sum of deviations of a set of 10 items measured from 30 is zero. Hence their mean is ..... (30/10)

v) The sum of squares of deviations of a set of 15 items from a number 35 is 890. The sum of the squares of deviations of these items from their arithmetic mean 30 must be ..... than 890. (more/less)

vi) The mean wage of 60 labourers working in the dayshift is Rs. 40 and the mean wage of 40 labourers in the night-shift is Rs. 35. The mean wage of those two groups together is ..... (38/40)

vii) The mean of the data set of 5 items is 10. From each item number 3 is subtracted and then each item is multiplied by 2. The new mean will be..... (14/15)

10.10 MERITS AND LIMITATIONS OF ARITHMETIC MEAN

The arithmetic mean has the following merits and limitations:

Merits

1) It is easy to understand and simple to compute. It is the widely used summary measure.
2) It is rigidly defined.
3) It acts as a single representative figure of the whole data set.
4) It is based on all items of the data. It does not depend on its position in the series.
5) It leads itself to further mathematical treatment.
6) It is useful in further statistical analysis. It is used in the computation of other statistical measures like standard deviation, coefficient of variation, co-efficient of skewness, etc. (you will learn about them in Block 4).

7) It is characterised as a centre of gravity—a point of balance.
8) For various sampling methods, the simple mean is an unbiased estimate of the population mean.

Limitations

1) It is unduly affected by extreme values. Very small or very big values in the data unduly affect the value of mean. Therefore, for the distribution where concentration is on small or big values, the mean will not be a proper average to yield a representative figure.

2) For the open-ended distribution, mean cannot be computed with accuracy. For example, in an income distribution starting with the class 'below 500' and ending with the class 'above 5,000' mean cannot be computed without making assumptions regarding the values of two extremes. As a result, error may creep in.

3) Mean is not useful for studying the qualitative phenomena e.g., beauty, honesty, intelligence, etc.

4) For the reasonably normal (bell shaped) distribution, mean can act as a good measure of central tendency. But for a U-shaped distribution (which has high frequency in the beginning, low in the middle and again high towards the end) it hardly succeeds to be a point of location around which other individual values congregate.

5) Mean does not lead a life of its own. For example, the statement that the average number of children in Indian family is 4.8 does not imply that there is even a single family having 4.8 children. Nor was a duck ever killed by the average of two shots—one a yard in front of it and one a yard behind it.

6) For non-homogeneous data, average may give misleading conclusion. For example, sales (in lakh rupees) of two business units A and B during the last five years are as follows:

A :	30	25	20	15	10
B :	10	15	20	25	30

Here it is clear that the average sales of both the units are exactly the same and yet unit B is thriving whereas unit A is flickering.

10.11 SOME ILLUSTRATIONS

Illustration 8

Weekly wages (in rupees) of 30 workers are given below :

140	139	126	114	100	88	62	77	99	103
108	129	144	148	134	63	69	148	132	118
142	116	123	104	95	80	85	106	123	133

The firm gave bonus of Rs. 10, 15, 20, 25, 30 and 35 for individuals in the respective salary groups of Rs. 61-75, Rs. 76-90, Rs. 91-105, Rs. 106-120, Rs. 121-135 and Rs. 136-150. Find the average bonus paid to all the workers.

Solution

To find out, how many persons were paid bonus of Rs. 10, Rs. 15, 20, etc., you have to find out the number of workers in the salary classes of Rs. 61-75, Rs. 76-90, etc. Using tally bar method for this, the average bonus is calculated as follows:

Calculation of Average Bonus

Weekly Wages Rs	Tally Bars	Frequency f	Bonus Paid x	fx
61 - 75	III	3	10	30
76 - 90	IIII	4	15	60
91 - 105	𝍸	5	20	100
106 - 120	𝍸	5	25	125
121 - 135	𝍸 II	7	30	210
136 - 150	𝍸 I	6	35	210
		$n = \Sigma f = 30$		$\Sigma fx = 735$

$$\text{Arithmetic Mean of Bonus Paid} = \frac{\Sigma fx}{n} = \frac{735}{30} = \text{Rs. } 24.50$$

Illustration 9

The average salary paid to all workers of a company is Rs. 500. Average salaries paid to skilled and unskilled workers are Rs. 520 and Rs. 420 respectively. Determine the percentage of skilled and unskilled workers.

Solution

Let the percentage of skilled workers be n_1. Then the percentage of unskilled workers will be $100 - n_1$. Let the mean of skilled and unskilled workers be denoted as $\bar{x}_1$ and $\bar{x}_2$ respectively.

$$\text{Combined mean } \bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Now substitute the given values of $\bar{x}_1$, $\bar{x}_2$, n_1 and n_2 in the formula.

$$500 = \frac{520 \times n_1 + 420\,(100 - n_1)}{n_1 + (100 - n_1)}$$

$$500 = \frac{520n_1 + 420{,}000 - 420n_1}{100}$$

$$50{,}000 = 100n_1 + 42{,}000$$
$$8{,}000 = 100n_1$$

$$n_1 = \frac{8{,}000}{100}$$
$$= 80$$
$$n_2 = 100 - n_1$$
$$= 100 - 80$$
$$= 20$$

Percentage of skilled worker is 80% and unskilled worker is 20%.

Illustration 10

Arithmetic mean of 100 items was found to be 50.8. It was later discovered, one item 47 was wrongly taken as 67. Find the correct mean.

Solution

Calculated total of all the items $(n\bar{x}) = 100 \times 50.8 = 5{,}080$. By subtracting the wrong entry and adding the correct entry, we can find the correct total.

Correct total = $5{,}080 - 67 + 47 = 5{,}060$

$$\text{Correct Arithmetic Mean} = \frac{5{,}060}{100} = 50.6$$

Illustration 11

Find the missing frequency from the following data :

Marks	:	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	:	5	15	20	-	20	10

Arithmetic Mean of Marks = 34.

Solution

Let the missing frequency be denoted as 'F'.

Calculation of Mean

Marks	frequency f	Mid-points m	fm
0 - 10	5	5	25
10 - 20	15	15	225
20 - 30	20	25	500
30 - 40	F	35	35F
40 - 50	20	45	900
50 - 60	10	55	550
	n = 70 + F		Σ fm = 2,200 + 35F

Now $\bar{x} = \frac{\Sigma fx}{n}$

$$34 = \frac{2,200 + 35F}{70 + F}$$

$34 (70 + F) = 2,200 + 35F$

$2,380 + 34F = 2,200 + 35F$

$2,380 - 2,200 = 35F - 34F$

$180 = F$

$\therefore$ Missing frequency = 180.

Illustration 12

Following data gives the number of students appearing for the examination and the pass percentage of different courses of two Universities. Calculate the pass percentage by simple arithmetic mean and weighted arithmetic mean and comment which University has higher average pass percentage.

Courses	University A		University B	
	Pass %	Students	Pass %	Students
M.A.	71	400	82	200
M.Com.	83	500	75	300
B.A.	72	300	74	600
B.Com.	74	200	73	700

Solution

Calculation of Simple and Weighted Means

Course	University A			University B		
	Pass% X_A	Students W_A	$X_A W_A$	Pass% X_B	Students W_B	$X_B W_B$
M.A.	71	400	28,400	82	200	16,400
M.Com.	83	500	41,500	75	300	22,500
B.A.	72	300	21,600	74	600	44,400
B.Com.	74	200	14,800	73	700	51,100
Total	300	1,400	1,06,300	304	1,800	1,34,400

University A

$$\text{Simple Arithmetic Mean} = \frac{\Sigma X_A}{n} = \frac{300}{4} = 75\%$$

$$\text{Weighted Mean} = \frac{\Sigma X_A W_A}{W_A} = \frac{106,300}{1,400} = 75.9\%,$$

University B

$$\text{Simple Arithmetic Mean} = \frac{\Sigma X_B}{n} = \frac{304}{4} = 76\%$$

$$\text{Weighted Mean} = \frac{\Sigma X_B W_B}{W_B} = \frac{134,400}{1,800} = 74.7\%$$

Simple Arithmetic Mean for University B is higher and weighted mean of University A is higher. So from the point of view of simple mean, University B is better but from the point of view of weighted mean University A is better. In fact in this problem the number of students appearing in various courses is different from one another. Thus, the basis of calculations of pass percentage for various courses are different from one another. In such a situation weighted mean would give the correct result. Thus, University A has the higher average pass percentage. In other words, in University A, the actual number of students passing out of total 1,400 is 1,063. So the pass percentage is $(1063/1400) \times 100 = 75.9$ % which is same as weighted average. Similarly, 1,344 students pass out of 1,800 in University B, giving a pass percentage of $(1344/1800) \times 100 = 74.7$ % which is equal to weighted average. So comparison of weighted average only can point out which University is better.

10.12 LET US SUM UP

The main characteristics of the data are represented by a single figure known as 'an average' or 'a mean'. It is the point of location around which individual values cluster. An ideal average must satisfy certain properties such as ease of calculation, rigidity in its definition, should be based on all items, should remain unaffected by extreme items, should be capable of further algebraic treatment and should have sampling stability. An average gives a bird's-eye view of the entire data, facilitates comparison and becomes useful in statistical inference.

There are easy formulas for obtaining simple mean for ungrouped and grouped data. When values in the data set are of unequal importance, a weighted arithmetic mean will be a truly representative average. The weighted mean gives the summary of two things: (i) the items, and (ii) how weights affect the items. So depending on weighting pattern, simple mean can be equal to or greater than or less than weighted mean.

Mean has a few important properties: (a) Sum of deviations from the mean is always zero. (b) The sum of square of deviation of items from mean is minimum. (c) If mean and the number of items are known, we can estimate the total Σx. (d) If a constant 'C' is added to or subtracted from every item of a data set, the mean is also increased or decreased by the quantity 'C'. Similarly, if every item is multiplied by a constant 'k', the mean is also multiplied by 'k'. (e) A combined mean of two or more groups can be obtained.
Mean is a very useful measure. It is a point of balance and it forms the basis of advance analysis. It also suffers from several limitations. It is affected by the extreme items. For an open-end distribution it can only be estimated with certain assumptions. Moreover, it gives misleading results for a non-homogeneous data.

10.13 KEY WORDS AND LIST OF SYMBOLS

Key Words

Central Tendency : A single value that has a tendency to be somewhere at the centre and within the range of all values.

Extreme Values : The items that are too big or too small in comparison with the other items of data. They unduly influence the mean.

Mean : The value obtained by dividing the sum of values of all observations in the given data set by the number of observations.

Measure of Location : A measure which is a point of location around which other individual values of data set congregate.

Sampling Stability : The averages obtained from different samples drawn from the same aggregate should be approximately the same.

Weighted Arithmetic Mean : An average whose component items are assigned weights according to their relative importance.

List of Symbols
The symbols used in writing different formulas in your study material are the one which are widely used. In the list given below, they are given as the first symbols under various words. Many textbooks have used symbols different from these. Most of the commonly used symbols are also given in the list below. Whenever you read a book it is important to clearly understand the meanings of the various symbols used there. For writing different formulas you may use any set of symbols you like, but along with them it is always desirable to explain their meanings.

Arithmetic Mean	$\bar{x}$ A.M., $\bar{x}$ mean of the population is generally denoted by μ
Assumed Mean	A , a , x_o , $\bar{x}_d$

Class Interval	i , c , h , w
Combined Mean of Groups	$\bar{x}_c$, $\bar{\bar{x}}$, $\bar{x}_{12}$
Common Factor	c , i , h
Deviation of Items from Mean	$(x - \bar{x})$, x, D
Deviation of Mid-points from Assumed Mean	d, x', x, dx
Frequency	f , F
Items	X, x
Mid-point of a Class Interval	m , x
Step Deviation	d', d , x', u , U, dx
Total Number of Items	n , N , Σf. When data is from a sample, generally 'n' and when it relates to population 'N'.
Weights	w, W, wt.
Weighted Mean	$\bar{x}_w$, $\bar{x}_w$, wt. Mean

10.14 ANSWERS TO CHECK YOUR PROGRESS

A) 1) i) bird's-eye view; ii) summary; iii) extreme; iv) facilitates ; v) inference

2) i) False; ii) False; iii) False; iv) True; v) False

3) i) 11; ii) $\bar{x} = \frac{\Sigma fm}{\Sigma f}$, $\bar{x} = A + \frac{\Sigma fd}{n}$, $\bar{x} = A + \frac{\Sigma fd'}{n} \times c$

iii) minimises calculations; iv) A = 35

v) Every value in a class is equal to the mid-point of that class.

4) Rs. 164.33

5) i) 68; ii) 0

6) Rs. 128.33 by both methods

7) 40.2

B) 2) Simple Average = 42.92; Weighted Average = 44.23

3) Both are equal to 73.7 %

4) 34.47

5) i) False; ii) True; iii) True; iv) True; v) False

6) i) Weighted; ii) arbitrary; iii) 20; iv) 30; v) less; vi) 38; vii) 14

10.15 TERMINAL QUESTIONS/EXERCISES

Questions

1) Explain the qualities of a good measure of Central Tendency.

2) Give the properties and limitations of Arithmetic Mean.

3) What is weighted average? Under what conditions weighted average is preferable to a simple average?

Exercises

1) Number of skilled and unskilled labourers and their average hourly wages in two cities

are given below. Determine the average hourly wage for each city.

	Bombay		Calcutta	
Labour	Number	Wage per hr. Rs.	Number	Wage per hr. Rs.
Skilled	150	1.80	350	1.75
Unskilled	850	1.30	650	1.25

(Ans: Rs. 1.38 and Rs. 1.43)

2) An investor buys Rs. 120 worth of shares in a company every month. During the first 5 months he bought the stock at a price of Rs. 10, 12, 15, 20, and 24 per share. After 5 months what is the average price paid for the share in his portfolio?

(Ans: Rs. 14.63)

3) A factory which is running in two shifts has a total of 100 workers. Average wage paid to the workers is Rs. 38 per day. In the first shift 60 persons are working and their average wages is Rs. 40 per day. What is the average wage paid to the remaining 40 workers who are working in the second shift?
(Ans: Rs. 35)

4) Arithmetic mean of 50 items was found as 28.5. It was later found that item 39 was taken extra. Find the correct mean of 49 items.

(Ans: 28.3)

5) The following table shows the number of workers in various trade categories who worked from Monday to Friday in a week for varying number of hours each day. The hourly pay for categories I, II, III, IV and V workers is Rs. 0.97, Rs. 0.77, Rs. 1.01, Rs. 0.67, and Rs. 0.75 respectively. Calculate the average wage per hour per worker for the whole week for all categories together.

Categories	Number of Workers				
	Monday (7 hrs)	Tuesday (6 hrs)	Wednesday (5 hrs)	Thursday (4 hrs)	Friday (5 hrs)
I	30	20	25	15	30
II	25	25	30	20	20
III	30	25	30	25	20
IV	20	20	20	20	25
V	25	20	25	15	25

(**Hint:** Find total hours under each category and take it as weight)

(Ans: Rs. 0.84 per hour)

6) A state authority has estimated the age of households in two districts as given below. Calculate the mean age for:
i) Area 'A'
ii) Area 'B' and
iii) Two areas taken together.

Estimated Age (in years)	Percentage of Houses	
	Area 'A'	Area 'B'
0 - 20	16	13
20 - 40	37	35
40 - 80	35	46
80 - 100	12	6

(Ans: Area A = 58.45, Area B = 58.48 combined Area = 58.47)

7) State whether the following statements are True or False. Also give reasons.
i) A man claims that his average bank balance during the year was Rs. 370. The bank claims that he overdrew his account at least 10 times during that year. Both are right.
ii) The sum of deviations of a set of 10 observations measured from 28 is zero. Hence the mean of observations is zero.

iii) The sum of squares of deviations of a set of 20 observations from a number of 42 is 750 and the sum of squares of deviations of these observations from the arithmetic mean 34 is 800 .

iv) The mean of a certain number of items is 42. If one more item '64' is added to the data, the mean becomes 44. Therefore, there should be 10 items in the original data.

v) If we replace each item in the series by the mean value of that series, sum of these substitutions will be equal to the sum of individual items.

(Ans: i True, ii False, iii False, iv True, v True)

Note: These questions/exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.

UNIT 11 MEDIAN

Structure

11.0 OBJECTIVES

After studying this unit, you should be able to :

- define median
- compute median for different types of data
- enumerate the properties of median
- define different kinds of partition values and compute them graphically locate the median and other partition values
- state the uses and limitation of median as a measure of central tendency.

11.1 INTRODUCTION

You have studied in Unit 10 that there are several measures of central tendency. You have also studied in detail about arithmetic mean which is one of the measures of central tendency. As you know, the arithmetic mean is very much affected by extreme items. Many times we may like to find a measure of average which is not affected by the extreme items. Median is one such measure. There are some other measures called partition values, which are not averages, but similar to median in concept. In this unit you will learn the meaning, computation, properties, limitations and uses of median and other partition values.

11.2 WHAT IS MEDIAN?-

The median is also a measure of central tendency. Unlike arithmetic mean, this median is based on the position of a given observation in a series arranged in an ascending or descending order. Therefore, it is called a **positional average.** It has nothing to do with the magnitude of all the observations, as in the case of arithmetic mean. Simply, median refers to the middlemost value of the variable when they are arranged in order of magnitude. The position of the median in a series is such that an equal number of items lie on either side of it. **Median of a given series is the value of the variable that divides the series into two equal parts. It is the most central point of a series where half of the items lie above this value and the remaining half lie below this value.** In the case of a frequency curve the median is that value of the variable which splits the area into two equal parts. **The median is usually denoted by 'M_d'**

11.3 COMPUTATION OF MEDIAN

Median can be computed for both ungrouped and grouped data. But the methods are different. Now let us study the methods of computing median for grouped and ungrouped data separately. .

11.3.1 Ungrouped Data

Having arranged the data in ascending order or descending order, the median is calculated as $\frac{n+1}{2}$ th item, N being the total number of items.

1) **When N is Odd :** When the number of observations is an odd number, the formula to compute median (M_d) is $\frac{n+1}{2}$ th item, where 'N' is the number of observations..

For example take the series 6, 7, 4, 8, 11, 5, 3, 9, 10. In this case the number of observations is nine which is an odd number. Now the median is $\frac{n+1}{2}$ th item = $\frac{9+1}{2}$ th item = 5th item. It means that when the given series is arranged in an ascending order, the fifth item will be the median. Now we can arrange the data in ascending order and identify the fifth item. The arranged series is 3, 4, 5, 6, 7, 8, 9, 10, 11, and the 5th item is 7. Therefore, median (M_d) is 7.

2) **When N is Even:** When the number of observations (N) is an even number, $\frac{n+1}{2}$ will involve a fraction. In such cases the median is taken as arithmetic mean of two middle values.

For example, take the series 8, 11, 3, 16, 20, 32, 41, 36. In this series the number of observations is eight which is an even number. So the median (M_d) is $\frac{n+1}{2}$ th item = $\frac{8+1}{2}$ = 4.5th item. This involvesa fraction 0.5. You should note that there is po item with the serial number 4.5. Hence, you have to take the average of the items 4th and 5th as median. This happens with all the series when 'N' is an even number. Now we arrange the series in ascending order as shown here : 3, 8, 11, 16, 20, 32, 36, 41. The Median (M_d) is the arithmetic mean of items 4th and 5th in this arranged series. The values of items 4th and 5th in this series are 16 and 20 respectively. Therefore, M_d is 18 (i.e. $\frac{16+20}{2}$)

Even when N is an even number, median can be taken as $\frac{n+1}{2}$ th item. But for this purpose you havk to give a special meaning to interpret the fraction 0.5 in the value of $\frac{n+1}{2}$

In the illustration given above, 4.5th item is to be found out. By convention 4.5th item will be taken as 4th item plus half of the difference between the 4th and 5th items. In the given data arranged in ascending order, 4th item is 16 and 5th item is 20. Thus; Median (M_d) is 18 (i.e. $16 + \frac{1}{2}(20 - 16)$. This value is same as obtained earlier. Hence, we can define median for ungrouped data as $\frac{n+1}{2}$ th item whether N is an odd number or an even number.

You should note that when N is an even number, it is easy to find median as arithmetic mean of two middle items. But the meaning given to fraction size of the item as indicated above is very much useful in calculations of other partition values about which you will learn later in this unit. Moreover, this formula helps us in giving a general definition to median for ungrouped data.

113.2 Grouped Data

As you know, when the data is in the form of frequency distribution, it can be either in the form of discrete series or continuous series. The method of computing median is different for these two types of frequency distributions. Now let us study them separately.

Discrete Series

In this case, first arrange the data in ascending or descending order. Then find out the cumulative frequencies. As median being N t h item, locate the value corresponding to

$\frac{n+1}{2}$ or next higher than that in the column of cumulated frequencies. Thus, having determined a median class, the corresponding value of the variable in that median class is the value of median. Let us understand it by an illustration.

Illustration 1

Calculate the median marks for the following data :

Marks :	40	15	25	5	30	35	10	50	45	20
No. of Students :	9	75	72	20	45	39	43	6	8	76

Solution

First rearrange the data in the ascending order of magnitude of marks, and then prepare the cumulative frequency as shown below:

Marks :	5	10	15	20	25	30	35	40	45	50
No. of Students :	20	43	75	76	72	45	39	9	8	6

Calculation of Cumulative Frequency

Marks	No. of Students	Cumulative Frequency
5	20	20
10	43	63
15	75	138
20	76	214
25	72	286
30	45	331
35	39	370
40	9	379
45	8	387
50	6	393

Here N = 393.

Median = $\frac{n+1}{2}$ th item = $\frac{393+1}{2}$ th item = 197th item

The 197th item falls in the class with cumulated frequency **214.** The value of the variable in that class is **20.** Therefore, median marks are **20.**

Continuous Series

In the case of frequency distribution of continuous series, exact values of various items are not known. So the size of a particular item cannot be found. What can be done is, to find out a value which has half the items below or the above it. Thus, in order to locate median class N/2 is taken in place of $\frac{n+1}{2}$ and the rest of the procedure is the same as the procedure followed in the case of discrete series. Having located the median class, the exact value of the variable can be interpolated from that class by any of the following three methods :

Method 1:

$$M_d = l + \frac{\frac{N}{2} - C}{f} \times i$$

Where l = lower limit of the median class

C = cumulative frequency of a class preceding the median class

f = simple frequency of the median class

i = the class-interval of the median class.

Method 2: The assumption in the formula used in the first method is that cumulated frequencies are calculated from lower values side. In case cumulated frequencies are calculated from higher values side, the above formula can be slightly modified as

$$M_d = U - \frac{\frac{N}{2} - C}{f} \times i$$

where u = upper limit of median class

C = cumulated frequency of a class next to the median class

f = simple frequency of the median class

i = the class interval of the median class

Method 3 : Median can also be calculated by using the following formula :

$$M_d = l + \frac{u - i}{f}(m - C)$$

where l = lower limit of the median class
u = upper limit of the median class
f = simple frequency of the median class
C = cumulated frequency of the class preceding the median class
m = N/2

All these three methods produce exactly the same result. The assumptions and the logic for interpolating median by all these three methods are almost the same. Now let us explain the assumptions for the formula under Method 1.

If items are counted from the lower values side, 'C' items will be completed upto the lower limit 'l' of the median class. But to reach the median point, N/2 items must be covered. Therefore, $\frac{N}{2} - C$ items are to be covered in the median class. There are 'f' items spread over a class interval 'i' of this median class. It is now assumed that all these 'f' items are uniformly distributed over the range 'i'. Thus, to cover N/2 − C items in the median class, a distance of $\frac{i}{f} \times \left(\frac{N}{2} - C\right)$ has to be travelled from 'l' limit (i.e., the lower limit) onwards.

Therefore, median $M_d = l + \frac{i}{f} \times \left(\frac{N}{2} - C\right)$

You should note the difference in the assumptions behind the median and the mean. In case of median the assumption is that items are uniformly spread out in a class interval, whereas. in the case of arithmetic mean it is assumed that the values of all items of a class interval are equal to the mid-point of that class interval,

Illustration 2

The manager of a departmental store compiled information on 200 accounts receivable which were delinquent. For each account he has noted the number of days passed after the due date. He then grouped the data as shown in the following frequency distribution. Determine the median.

No. of Days Passed After Due Date	No. of Accoupts
30 - 44	40
45 - 59	45
60 - 74	40
75 - 89	25
90 - 104	25
105 - 119	20
120 - 134	5

Solution

Calculation of Median

No. of Days Passed After Due Date	No. of Accounts (f)	Cumulative Frequency (Less than)	Cumulative Frequency (More than)
30 - 44	40	40	200
45 - 59	45	85	160
60 - 74	40	125	115
75 - 89	25	150	75
90 - 104	25	175	50
105 - 119	20	195	25
120 - 134	5	200	5

Here N/2 = 200/2 = 100. This implies that there are 100 items below median. Therefore, 60-74 is the class where the median lies, The real limits of this class is 59.5-74.5. Now compute the median using the first method.

$$M_d = l + \frac{\frac{N}{2} - C}{f} \times i$$

where l = 59.5
c = 85
f = 40
i = 15

$N = 200$

$$M_d = 59.5 + \frac{100 - 85}{40} \times 15$$

$= 59.5 + (15/40) \times 15$

$= 59.5 + 225140$

$= 59,5 + 5.625$

$= 65.125$

Median = 65.1 days.

Now let us compute the median by using the second method.

$$M_d = U - \frac{\frac{N}{2} - C'}{f} \times i$$

where u = 74.5

f = 40

c' = 75

i = 15

N = 200

$$\therefore M_d = 74.5 - \frac{\frac{200}{2} - 75}{40} \times 15$$

$= 74.5 - (25/40) \times 15$

$= 74.5 - 375/40$

$= 74.5 - 9.375$

$= 65.125$

Median is 65.1 days. You can obtain the median by using the third method :

$$M_d = l + \frac{U - 1}{f}(m - C)$$

where l = 59.5

u = 74.5

f = 40

m = N/2 = 200/2 = 100

c = 85

$$M_d = 59.5 + \frac{74.5 - 59.5}{40}(100 - 85)$$

$= 59.5 + (15/40) \times 15$

$= 65.125$

Median is 65.1 days. You should note that all the three methods produced the same result.

Illustration 3

Find the median income from the following income distribution :

Monthly Income (Rs.)	No. of Families
Below 100	50
100- 200	500
200-300	555
300 - 500	100
500 - 800	3
800 and above	2

Solution

Monthly Income (Rs.)	No, of Families	Cumulative Frequency
Below 100	50	50
100-200	500	550
200 - 300	555	1,105
300 - 500	100	1,205
500-800	3	1,208
809 and above	2	1,210

Median has N/2 items below it which means 1,21012= 605 items below it. Therefore, the median lies in the 200-30q class. Now applying the formula of interpolation

$$M_d = 1 + \frac{\frac{N}{2} - C}{f} \times i$$

where 1 = 200
c = 550
f = 555
i = 100
N = 1,210

$$M_d = 200 + \frac{605 - 550}{555} \times 100$$

$$= 200 + (55/555) \times 100$$
$$= 200 + 9.91$$
$$= 209.91$$

Median Monthly Income is Rs. 209. 91

You may note that the class intervals in this illustration are unequal and the data is open-ended. This does not affect the calculation of the median. The length of the class interval ('i') in the formula corresponds only to the median class.

Illustration 4

Determine the median wage from .ne following data :

Wages More Than (Rs.)	No. of Workers
20	58
40	54
60	4'8
80	38
100	22
120	10
140	3
160	0

Solution

Wages More Than (Rs.)	No; of Workers (Cumulative Fre.)	Simple Frequency
20	58	58-54 = 4
40	54	54–48 = '6
60	48	48–38 = 10
80	38	38–22 = 16.'
100	22	22–10 = 12
120	10	'10 – 3 = 7
140	3	3 – 0 = 3
1 6 0	0	0

Cumulative frequency is given in this illustration. So, we have calculated simple frequency. Now median has N/2 items i.e., 5812 = 29, items above it. Therefore, median lies in the 'more than 80' class i.e., 80-100 class. We can interpolate median by using the following formula :

$$M_d = U - \frac{\frac{N}{2} - C'}{f} \times i$$

whereu = 100
C' = 22
f = 16
i = 20

$$M_d = 100 - \frac{29 - 22}{16} \times 20$$

$$= 100 - (7/16) \times 20$$
$$= 100 - 8.75$$
$$= 91.25$$

Median wage is Rs. 91.25.

Illustration 5

You are given the following incomplete frequency distribution. It is known that total frequency is 1,000 and that the median is 413.11. Estimate the missing frequencies.

Values	Frequency
300-325	5
325-350	17
350-375	80
375-400	
400-425	326
425-450	
450-475	88
475-500	9

Solution

Let us assume that the frequency of the class 375-400 is F. Now the frequency of the class 425-450 becomes 1,000 − (525 − F) = 475 − F (525 being the total of given frequencies).

Values	Ftequency	c.f.
300-325	5	5
325-350	17	22
350-375	80	102
375-400	F	102+F
400-425	326	428+F
425-450	475 -F	903
450-475	88	991
475-500	9	1000

Since the median is given as 413.11, the median must be in 400-425 class.

Now $M_d = 1 + \frac{\frac{N}{2} - C}{f} \times i$

where $1 = 400$

$f = 326$

$C = 102+F$

$i = 25$

$M_d = 413.11$

$$413.11 = 400 + \frac{500 - (102 + F)}{326} \times 25$$

$$413.11 - 400 = \frac{500 - 102 - F}{326} \times 25$$

$$13.11 = \frac{398 - F}{326} \times 25$$

$$13.11 \times 326 = (398 - F) \times 25$$

$$4{,}273.86 = 9{,}950 - 25F$$

$$25F = 5{,}676.14$$

$$F = 227.04$$

As frequency should be an integral value F = 227. Therefore, frequency for the class 375-400 is 227 and the frequency for the class 425-450 is 475-227 = 248.

11.4 PROPERTIES OF MEDIAN

You have studied the methods of 'computing median. Now let us discuss the properties of median.

1) An important property of the median is that the sum of the absolute deviations (i.e., deviations ignoring signs) from the median is minimum i.e. $\Sigma|x - M_d|$ is the minimum. This property entails the use of median in various practical situations. For example, take the items 5, 7, 8, 9, 21. In this case the median $\frac{(N+1)}{2}$ is 8. Let us calculate absolute deviations from (i) median, (ii) any other value say 7, and (iii) from arithmetic mean. (i.e., $\frac{5+7+8+9+21}{5}$)

Item X X	$\lvert x - M_d \rvert$ $\lvert X - 8 \rvert$	$\lvert X - 7 \rvert$	$\lvert X - \bar{X} \rvert$ $\lvert X - 10 \rvert$
5	3	2	5
7	1	0	3
8	0	1	2
9	1	2	1
21	13	14	11
Total	8	19	22

If you study the above table carefully, you will notice that the least total is 18, which is the sum of absolute deviations from median.

2) It is not affected by the extreme items. It is of course affected by the number of items.

3) For an open-ended distribution, median is the more suitable average. For example, since the income distribution is an open-ended distribution, median income would be a more representativefigure.

4) For the qualitative information, median is probably the only suitable measure of central tendency. For example, a respondent may be asked to rate his evaluation of the corporate image, in the order of importance, as dynamic, prestigious, cooperative (business-wise), successful and withdrawn. Suppose he ranks them exactly as given here, the third adjective viz. cooperative (business-wise) is the median of his five ratings.

5) The median can be located graphically (you will study this later in this unit)

6) It is easy to compute and lucid to understand. In some cases it is obtained even by an inspection.

11.5 MERITS AND LIMITATIONS OF MEDIAN

You have studied the meaning, methods of computation and properties of median. Now let us discuss the merits and limitations of median.

Merits

1) For an open-ended distribution, such as income distribution, the median gives a more representative value.

2) Since median is not distorted by the extreme items, in some cases it is preferred over mean as the latter is likely to be distorted by extreme values.

3) For dealing the qualitative phenomena, median is the most suitable average.

4) Since median minimises the total absolute deviations, median is preferred in the situations wherein the total geographical distance is to be minimised. For example, there is a conference of five top executives from five different cities of India lying almost in a straight line. The city located at a median distance would be a more proper place for the conference.

5) While taking a decision to buy a particular brand of tyre, when only one or two tyres are to be bought, the brand with greater median run will be preferred. Similarly, in buying a washing machine, the machine with greater median life will be preferred, rather than one with a greater mean life.

Limitations

1) Median is not capable of algebraic treatment. That means we cannot have a combined median of two or more groups, unless all the items of the groups are known.

2) It is described, sometimes, as an insensitive measure as it is not based on all items of the series.

3) It is affected more by sampling fluctuations than the value of mean.

4) The computational formula of a median is in a way an interpolation under the assumption that the items in the median class are uniformly distributed, which is not very true.

5) The impression created by median in some cases may be illusory and deceptive because its value is determined strictly by the value of middle observation(s). For example, in lotteries the median value of the prize won by a ticket is always zero when all tickets are considered (more than 50% of the tickets will not get any prize). This median value of prize will, not help in analysing the prizes offered by lotteries as the matter of interest may be the first prize out of a number of prizes offered.

Check Your Progress A

i) Find the median for the following data sets :

a) 1, 2, 4, 8, 16, 32, 64, 128,256

b) 1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10

2) What is the formula for computing median for continuous data, when cumulated frequencies are calculated from higher values side?

3) In a given frequency distribution, if the class intervals are of unequal width, which class interval would you use for computing median?

4) Heights (in inches) of a group of students are given below. Calculate the median.
61, 62, 62, 63, 61, 63, 64, 64, 60, 65, 63, 64, 65, 66, 64

Now suppose, another group of students whose heights are 60, 66, 59, 68, 67, and 70 inches is added to the previous group. Find the median of the combined group.

5) Calculate the median from the following frequency distribution of marks in Economics:

Marks :	5	10	15	20	25	30	35	40	45	50
No of Students:	20	43	75	76	72	45	39	9	8	6

6) The following information is about the life (in hours) of 100 new-type light bulbs. Find the median life.

Life (in hours)	Number failed
1 - 50	2
51 - 100	8
101 - 150	15
151 - 200	20
201 - 250	25
251 - 300	20
301 - 350	10

.....

.....

7) State whether the following statements are True or False.

i) The data must be arranged before determining the median.

ii) Median cannot be computed in open-ended distribution.

iii) Class interval of a distribution may or may not be uniform in calculating the median;

iv) Sum of the deviations from median is the least.

v) For a data set with 16 items, the median is 8.

11.6 PARTITION VALUES

As you know median is the middle value of the variable when the items are arranged in the order of magnitude. Thus, median splits the series into two equal parts, Hence, it is called positional average. In fact there are other positional measures that partition the series into still more number of equal parts, say four equal parts or 10 equal parts or 100 equal parts. Such measures are generally known as **Partition Values.** There are three partition values: 1) Quartiles, 2) Deciles and 3) Percentiles, which are in much use. They are, of course, the measures of non-central location. Now let us study about them one by one.

11.6.1 Quartiles

The values of a variate that divide the series or the distribution into 4 equal parts are known as Quartiles. Since three points are required to divide the data into 4 equal parts, we have three quartiles Q_1, Q_2, and Q_3

The first quartile (Q_1), known as a lower quartile, is the value of a variate below which there are 25% of the observations and above which there are 75% of the observations.

The second quartile (Q_2) is the value of a variate which divides the distribution into two equal parts. It means, there are 50% observations above it and 50% below it. Therefore, Q_2 is the same as median.

The third quartile (Q_3), known as an upper quartile, is the value of a variate below which there are 75% observations and above which there are 25% observations.

It is clear that $Q_1 < Q_2 < Q_3$

Computation of Quartiles

i) **Discrete Series** (i.e. Individual Values Known). When the data is arranged in the ascending order:

Q_1 = Size of $\frac{N+1}{4}$ th item

Q_2 = Size of $\frac{2(N+1)}{4}$ th item

Q_3 = Size of $\frac{3(N+1)}{4}$ th item

ii) **Continuous Series** (i.e. Data with Class Intervals)

$$Q_J = l + \frac{\frac{JN}{4} - C}{f} \times i \qquad J = 1, 2, 3$$

where l = Lower limit of quartile class
c = Cumulated frequency preceding the quartile class
f = Simple frequency in the quartile class
i = Class-interval of quartile class

11.6.2 Deciles

The values of a variate that divide the series or the distribution into 10 equal parts are called

Deciles. Each part contains 10% of total observations. Obviously there should be nine such values denoted as $D_1, D_2 \ldots\ldots D_9$. They are called first decile, second decile, etc. The 5th decile (D_5) is the median.

Computation of Deciles

i) Discrete Series (i.e. Individual Values Known).

D_j = Size of $j\frac{(N+1)}{10}$ th item. J = 1 to 9

ii) Continuous Series (i.e. Data with Class Intervals)

$$D_j = l + \frac{\frac{JN}{10} - C}{f} \times i. \quad J = 1 \text{ to } 9$$

where C is the cumulated frequency preceding the jth decile class, the other symbols have usual meaning.

11.6.3 Percentiles

The value of a variate which divides a given series or distribution into 100 equal parts are known as percentiles. Each percentile contains 1% of the total number of observations. The percentile P_j is that value of the variate upto which lie exactly j % of the total number of observations. For example:

P_{10} = Value of a variate upto which lies exactly 10% of observations. This is same as D_1

P_{20} = Value of a variate upto which lies exactly 20% of observations.

P_{25} = Value of a variate upto which lies exactly 25% of the total number of observations. This is same as Q_1

P_{50} = Value of a variate upto which lies exactly 50% of the total number of observations. This is the same as D_5 or Q_2, or median.

Similarly, $P_{75} = Q_3$

Computation of Percentiles

i) Discrete Series (i.e. Individual Values Known).

P_j = Size of $\frac{j(N+1)}{100}$ th item

e.g. P_{45} = Size of $\frac{45(N+1)}{100}$ th item

ii) Continuous Series (i.e. Data with Class Intervals)

$$P_j = l + \frac{\frac{JN}{100} - C}{f} \times i \quad J = 1 \text{ to } 99$$

where C is the cumulated frequency preceding the jth percentile class. The remaining symbols have usual meaning. Let us understand the computation of partition values by two illustrations.

Illustration 6

Marks of 16 students in a class test (maximum marks 20) are as follows:

2, 3, 6, 7, 10, 10, 11, 11, 11, 12, 12, 14, 15, 16, 18, 19.

Calculate Q_1, P_{35}, D_9

Solution

Marks are already arranged in ascending order.

Q_1 = Size of $\frac{N+1}{4}$ th item

$= \frac{16+1}{4}$ th item

$= 4\frac{1}{4}$ th item

$\therefore Q_1$ = 4th item $+ \frac{1}{4}$ (5th item − 4th item)

$= 7 + \frac{1}{4}(10 - 7)$

$= 7 + \frac{3}{4}$

$= 7.75$

P_{35} = Size of $\frac{35(N+1)}{100}$ th item

$= \frac{35(16+1)}{100}$ th or $5 \times \frac{95}{100}$ th item

$\therefore P_{35} = 5\text{th item} + \frac{95}{100}(6\text{th item} - 5\text{th item})$

$= 10 + \frac{95}{100}(10 - 10)$

$= 10 + 0$

$= 10$

D_9 = Size of $\frac{9(N+1)}{10}$ th item = $\frac{9(16+1)}{10}$ th item on $15\frac{3}{10}$

$D_9 = 15\text{th item} + \frac{3}{10}(16\text{th item} - 15\text{th item})$

$= 18 + \frac{3}{10}(19 - 18)$

$= 18 + 0.3$

$= 18.3$

You may note that there is no student who has obtained 7.75 or 18.3 marks. When the size of item to be selected involves fraction, such hypothetical values can arise. The interpretation of such values become valid if the given data is a continuous series and not a discrete series.

Illustration 7.

The following table gives the distribution of monthly income of 600 families in Ahmedabad city.

Monthly Income Rs.	Families
Below 75	69
75 - 150	167
150 - 225	207
225 - 300	65
300 - 375	58
375 - 450	24
450 and above	10

a) Find D_2, D_9, P_{25}, P_{75}, Q_3 and Median.

b) Obtain the limits of income of central 50% of observed families.

c) Interpret the results.

Solution

Monthly Income (Rs.)	Families	Cumulative frequency
Below 75	69	69
75 - 150	167	236
150 - 225	207	443
225 - 300	65	508
300 - 375	58	566
375 - 450	24	590
450 and above	10	600

a) D_2 has $2N/10$ items below it. It means $2 \times 600110 = 120$ items below it. Therefore, D_2 falls in the 75-150 class.

Now $D_2 = 1 + \frac{\frac{2N}{10} - C}{f} \times i$

$= 75 + \frac{120 - 69}{167} \times 75$

$= 75 + \frac{51}{167} \times 75$

$= 75 + 22.9$

$= 97.9$

D_2 is Rs. 97.90

D_5 has 5N/10 items below it, which means 5 x 600/10 = 300 items below it. So D, lies in the 150-225 class

$$\text{Now } D_5 = 1 + \frac{\frac{5N}{10} - C}{f} \times i$$

$$= 150 + \frac{300 - 236}{207} \times 75$$

$$= 150 + \frac{64}{207} \times 75$$

$$= 150 + 23.19$$

$$= 173.19$$

D, is Rs. 173.19.

P_{25} has 25N/100 items below it, which means 25 x 600/100 = 150 items below it. SoP, lies. in the 75-150 class;

$$\text{Now } P_{25} = 1 + \frac{\frac{25N}{100} - C}{f} \times i$$

$$= 75 + \frac{150 - 69}{167} \times i$$

$$= 75 + \frac{81}{167} \times 75$$

$$= 75 + 36.38$$

$$= 111.38$$

P_{25} is Rs. 111.38

P_{75} has 75N/100 items below it, which means 75 x 600/100 = 450 items below it. So P_{75} lies in 225-300 class.

$$\text{Now } P_{75} = 1 + \frac{\frac{75N}{100} - C}{f} \times i$$

$$= 225 + \frac{450 - 443}{65} \times 75$$

$$= 225 + 8.077$$

$$= 233.077$$

P_{75} is Rs. 233.08

Q_3 has 3N/4 items below it, which means 3 x 600/4 = 450 items below it. P_{75} also has 450 items below it. So Q_3 must be same as P_{75}.

Q_3 = Rs. 233.08.

Median has N/2 items below it, which means 600/2 = 300 items below it. So it falls in the 150-225 class.

$$\text{Now } M_d = 1 + \frac{\frac{N}{2} - C}{f} \times i$$

$$= 150 + \frac{300 - 236}{207} \times 75$$

$$= 150 + \frac{64 \times 75}{207}$$

$$= 150 + 23.19$$

$$= 173.19 \text{ which is same as } D_5$$

Therefore, Median is Rs. 173.19

b) Central 50% of observations are given by an interval Q, to Q_3 as Q_1 has 25% of items below it and Q_3 has 25% of items above it.

Here Q, = P_{25} = Rs. 111.38 and Q_3 = Rs. 233.08. Required limits of income of central 50% of observed families are Rs. 111.38 to Rs. 233.08

c) Interpretation

D_2 = 20% of the families have monthly income of Rs. 97.90 or less and 80% of the families have monthly income of Rs. 97.90 or more.

D_5 = 50% of the families have the monthly income of Rs. 173.19 or less, and 50% have the monthly income of Rs. 173.19 or more. Median being the same as D, both have same interpretation.

p_{25} = 25% of the families have monthly income of Rs. 111.38 or less and 75% of the families have Rs. 111.38 or more.

P_{75} = 75% of the families have monthly income of Rs. 233.08 or less and 25% of the families have Rs. 233.08 or more. Q_3 and P_{75} being the same, they have the same interpretation.

11.7 GRAPHIC DETERMINATION OF MEDIAN AND OTHER PARTITION VALUES

You have learnt the method of computing median and other partition values. In fact, they can be determined graphically also. The median can be determined graphically by any of the following two methods:

1) "Less than" ogive curve and " more than" ogive curve are drawn. From the point of intersection of these two curves, a perpendicular is drawn on X-axis.

2) Only one ogive curve namely "less than" ogive is drawn. The variable is taken on X-axis and the cuinulated frequency on Y-axis. Then on Y-axis number N/2 is located. A horizontal line is drawn from it on the ogive curve. From the point where it meets the curve, perpendicular is drawn on the X-axis. The point where it meets the X-axis is the median.

Similarly, other partition values also can be determined graphically. For quartile Q_j, number JN/4 is located on Y-axis. For decile D_j, number JN/10 is located on Y-axis. For percentile P_j, the number JN/100 is located on Y-axis. And then the similar procedure is followed as prescribed for median.

Illustration 8

Find graphically median, $D_,$ P_{20}, Q , from the following information:

Profit per Shop, (less than Rs. '000)	No. of Shops
100	15
200	35
300	63
400	95
500	113
600	125
700	130

Solution

Profit per Shop (less than Rs. '000)	No, of Shops (c.f.)
100	15
200	35
300	63
400	95
500	113
600	125
700	130

Now take the values of a variable X on x-axis and the corresponding cumulated frequencies (c.f.) on y-axis, Then plot the 'less than" ogive curve. Look at Figure 11.1 carefully and study how less than ogive curve is drawn.

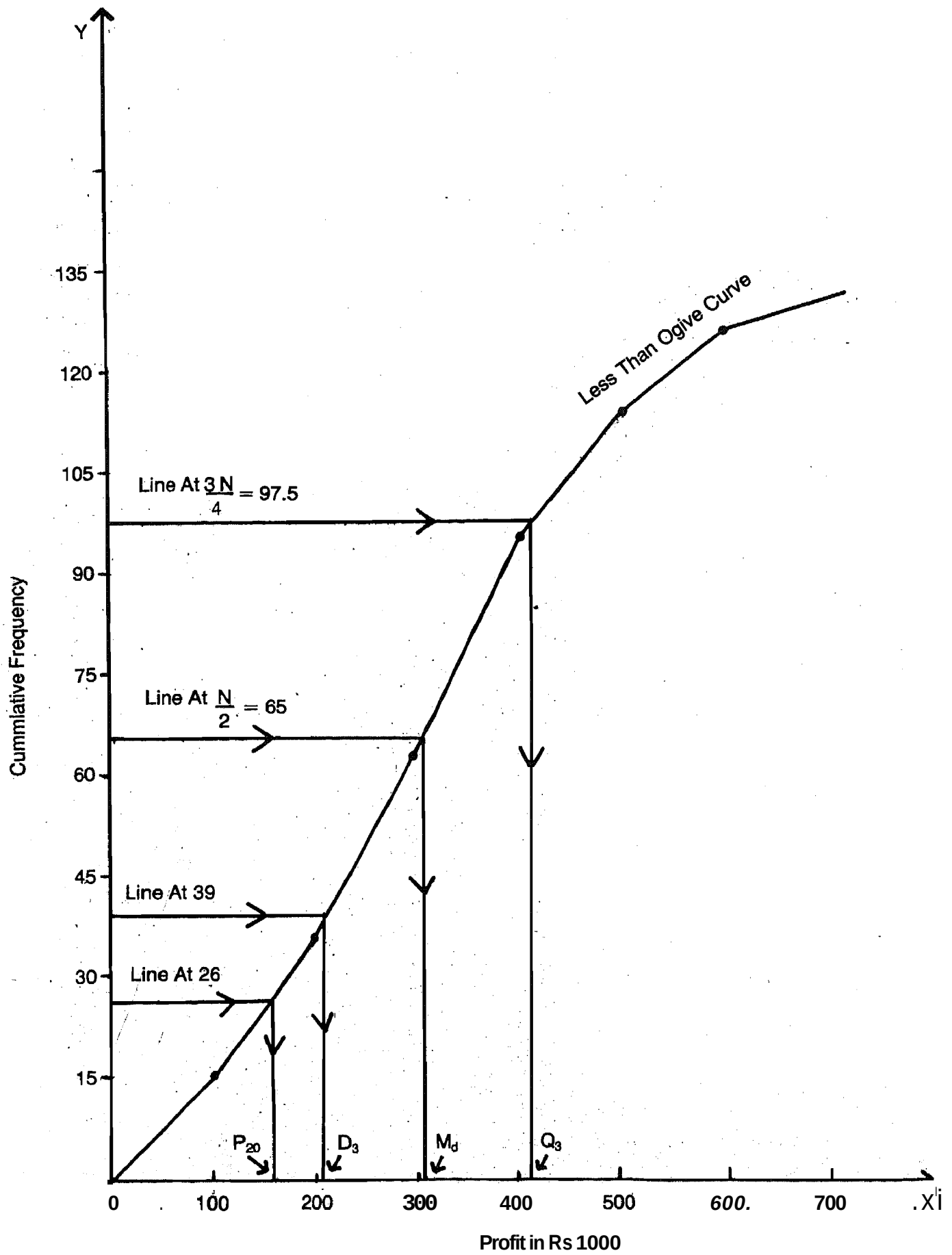


Figure 11.1 Less then Ogive Curve for Profit Per Shop and Location of Median, D_3, P_{20} and

For ***median,*** firstly, N/2 (i.e. 130/2 = 65) is located on y-axis and then from N/2 (i.e., 65) a horizontal line is drawn. From the point where this horizontal line meets the curve, a perpendicular is drawn on x-axis. It meets at the point '307' which is the median. Therefore, median profit is Rs. 307 thousands.

Median

For D_3: Firstly, $3 \times 130110 = 39$ is located on y-axis and corresponding value of D, is read on x-axis which is 213. Therefore, D, of profit is Rs. 213 thousands.

For P_{20}: Firstly $20 \times 130 / 100 = 26$ is located on y-axis. Proceeding the same way like median, you can read the value of P_{20} on x-axis of the graph which is 155. Therefore, P_{20} is Rs. 155 thousands. Similarly, Q_1 = Rs. 414 thousands. Except for small errors in plotting and reading the scale, the graphical values would almost correspond to the values obtained by applying their formulas.

Check Your Progress B

1) Define partition values. Name the partition values used in statistics.

.....

.....

.....

.....

.....

.....

2) Write the formulas for finding different partition values.

.....

.....

.....

.....

.....

3) State whether the following statements are True or False.

i) Graphical and calculation methods cannot give same result for a partition value.

ii) P_{70} means 70% of items have values less than this point.

iii) All deciles are included in percentiles:

iv) All quartiles are contained in deciles.

v) A point which has 30% of items above it will be called D_3.

vi) Median is same as third quartile.

vii) In a given series $Q_1 = 30.5$ and $Q_3 = 25.5$.

viii) 70% of the values of a variate are less than 70,.therefore $D_3 = 70$

ix) Third decile is the 30th percentile.

4) Fill in the blanks.

i) The sum of absolute deviations from median is.....

ii) Median is ..... quartile, ... decile and ..... percentile.

iii) For an income (monthly) distribution of $P_{70} = 125$ it means that ..... % of the families have the monthly income of Rs. 125 or..... .

iv) 60% of the variable are less than 70, therefore, 70 = ...

v) The limits of income of central 50% in a given income distribution refer to the range ..... to ..... .

11.8 LET US SUM UP

The median is a positional average, referring to the middlemost value of the variate above and below which half of the items lie. There are different formulas of computing the median from ungrouped as well as grouped data. Similarly, in grouped data itself methods are different for discrete series and continuous series. Median possesses some important properties: 1) the sum of the absolute deviations from the median is minimum, 2) it is not influenced by the extreme items, 3) it is a more suitable average for an open-ended distribution, 4) the qualitative phenomena can be better dealt with the help of median, and 5) it can be graphically located.

The median is very useful in the following situations: 1) where mean is likely to be distorted by the extreme items, 2) when the study is regarding the qualitative phenomena, 3) where the purpose is to minimise the geographical distance, and 4) where the 'buy' decision is to be made as regards the specific make of type of a household appliance like a washing machine. Median also suffers from certain limitations such as: 1) incapable of further algebraic treatment, 2) insensitive, 3) sampling instability, etc. In some situations, it even turns out to be a deceptive and unrealistic measure.

Like median, there are other positional measures known as partition values which partition the series into still more number of equal parts. They are: 1) quartiles, 2) deciles, and 3) percentiles. Quartiles are the three values of the variate dividing the series into four equal parts, each occupying 25% of the total observations. Deciles are the nine values of the variate dividing the series into 10 equal parts, each occupying 10% of the total observations. Percentiles are the values of the variate that divide the variate into 100 equal parts, each containing 1% of the total observations. Almost similar procedure is followed in the computation of the partition values, as prescribed for median. The median and partition values also can be located from the ogive curves. The less than ogive curve is very widely used for this purpose.

1.9 KEY WORDS AND SYMBOLS

Deciles: The values of the variate that divide the series or distribution into ten equal parts.

Less than Ogive : A cumulative frequency curve that starts from the lowest class boundary on the horizontal axis and gradually rising upward and ends at the highest class boundary corresponding to the total frequency of the distribution.

Median : The value of the variate that divides the skries intb two equal parts.

Partition Values : The values of the variate that divide the distribution into a fixed number of equal parts.

Percentiles: The values of the variate that divide the series or distribution into 100 equal parts.

Positional Average : An average based on the position of a given observation in a series arranged in the order of magnitude.

Quartiles: The values of the variate that divide the series or distribution into four equal parts.

List of Symbols

In addition to the list of symbols given in Unit 10, the following list of symbols is used in connection with median and other partition values. The list is on the same lines as in Unit 10.

Cumulative frequency	C, c, f, F, Σf_1,
Decile - jth	D_j
Deviations of items from median	$X - M_d$, d, D.
.Frequency of the medial group	f, f_1, f_m, f_{md}.

Lower limit of medial group or the group in which any partition value lies	$1, 1_1, L, L_m$
Median	$M_d, M, \bar{m}$
Percentile	P_j
Quartile-lower	Q_1
Quartile-upper	Q_3
Upper limit of medial group or the group in which any partition value lies	u, l_1, U, U_m

11.10 ANSWERS TO CHECK YOUR PROGRESS

A) 1) (a) 16, (b) 0.18

2) $M = U - \frac{\frac{N}{2} - C'}{f}$

3) Class Interval of median class is considered.

4) First Case 63. Second Case 64.

5) 30

6) 210.5

7) i) True, ii) False, iii) True, iv) False, v) False .

B) 3) i) False, ii) True, iii) True, iv) False, v) False, vi) False, vii) False, viii) False ix) True

4) i) minimum, ii) 2nd, 5th, 50th,

iii) 30%, less; iv) D_6 or P_{60}, v) Q_1, Q_3

11.11 TERMINAL QUESTIONS/EXERCISES

Questions

1) What is median? Explain its merits and limitations.
2) Explain the methods of computing median.
3) Compare the arithmetic mean and median as measures of average?
4) Compare and contrast between Quartiles, Deciles and Percentiles?

Exercises

1) The number of books issued at the counter of a university library on 10 different days are: 180, 95, 75, 70, 80, 102, 100, 94, 75, 400. Which average would represent this data best ? Calculate it.

   (Answer: Median 97.5)

2) Information on insurance claims for automobile accidents is given below. Determine the median.

Amount of Claim (Rs.)	Frequency
Less than 150	52
150 - 199.99	108
200 - 249.99	230
250 - 299.99	528
300 - 34999	663
350 - 399.99	816
400 - 449.99	993
450 - 499.99	825
500 and above	650

(Answer : Approximately Rs. 402)

3) Calculate the median from the following data, taking mean value as **45.5.**

Marks	No. of Students
70-80	10
60-70	10
50-60	20
40-50	
30-40	12
20-30	7
10-20	8
0-10	5

(Answer: 50)

4) Determine graphically the median from the following data. Obtain the range of marks obtained by middle 80% of the students.

Marks out of 60	No. of students
Less than 10	4
Less than 20	10
Less than 30	30
Less than 40	40
Less than 50	47
Less than 60	50

(Answer : 27.5, 11.7 to 47.1)

5) For a group of 500 students following information is available about the marks obtained out of 100:
Median = 45, $Q_1 = 23$, $Q_3 = 73$, $D_4 = 38$, $P_{63} = 60$, $P_{90} = 83$. And 8% of students have obtained less than 12 marks. 3% of students have obtained more than 95 marks.
Tabulate the data in class intervals.

Answer:

Marks :	0-12	12-23	23-38	38-45	45-60	60-73	73-83	83-95	95-100
No. of Students:	40	85	75	50	65	60	75	35	15

6) Find the missing frequencies if median is 25.

Daily Expenditure (Rs.)	Families
0-10	14
10-20	–
20-30	27
30-40	–
40-50	15

(Answer: 23, 21)

7) A laundry uses two different brands of washing machines. According to its past experience, the following results have been recorded :

Brand	Median Life	Mean Life
A	6,500 hours	6,000 hours
B	6,000 hours	6,500 hours

If both brands are of the same price, which brand should be purchased by the laundry.

8) Calculate Q_1, P_{30}, D_8 from the data given below:

Size of collar worn :	14	14.5"	15"	15.5"	16"
No. of Students :	20	37	43	26	14

(Answer : $Q_1 = 14.5"$; $P_{30} = 14.5"$; $D_8 = 15.5"$)

9) Determine graphically the valves of D_6, Median, P_{20}, Q, and Q_3 from the following data. Verify them by applying their respective formulae.,

Daily Wages (Rs.)	Workers
Below 105	
10 - 20	25
20 - 30	40
30 - 40	70'
40 - 50	90
50 - 60	40
60 - 70	20
Above 7010	

(Answer: $D_6 = 44.4$; Median 41.1; $P_{20} = 27.5$; $Q_1 = 30.7$; $Q_3 = 49.4$)

Mote: These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.

UNIT 12 MODE

Structure

12.0 OBJECTIVES

After studying this unit, you should be able to:

- define mode
- compute mode for different types of data
- locate mode graphically
- appreciate the limitations and uses of mode.

12.1 INTRODUCTION

As you know, among the measures of central tendency, there are some measures which are based on all items of the data and some other measures which are positional averages. In Unit 10 you have studied about arithmetic mean which is based on all the items of data. In Unit 11 you have studied about median which is a positional average. In this unit you will study about Mode which is another positional average. You will learn the meaning, methods of computation, locating it graphically, limitations and uses of mode.

12.2 WHAT IS MODE?

Mode is also a measure of central tendency. Mode is the value of a variate which is repeated most often in the data set. The genesis of the word 'mode' lies in the French word 'le mode' that means fashion. Mode is, therefore, considered to be the most common or most fashionable value.

Mode is often considered to be that value of the variate which occurs most frequently. But it is not exactly true for every frequency distribution. Rather it is that value of the variate around which the other items tend to concentrate most heavily. It shows the centre of concentration of the frequency in and around a given value. If is not the centre of gravity like mean. It is apositional measure similar to median. It is commonly denoted by M_o.

For example, take the case of a shopkeeper who sells shoes. He is interested to know the sizes of shoes which are commonly demanded. Were in such a situation, mean would indicate a size that may not fit any person. Median may not provide a representative size because of the unevenness in the distribution. It is the mode which will help in making a choice of approximate size for which an order can be placed.

12.3 COMPUTATION OF MODE

The method *of* computing mode is differentfor grouped data and ungrouped data, Now let us study those methods separately.

12.3.1 Ungrouped Data

For an ungrouped data mode is found out simply by inspection. The value that occurs most frequently in the given distribution is taken as a mode. For example, the ages (in years) of 10 boys are as follows: 5, 6, 4, 10, 7, 6, 9, 2, 8, 6. Here the number six appeared thrice. Therefore, mode age is six years.

Mode does not exist as such in some cases. For example, take the following data set : 5, 10, 15, 20, 25, 30. In this case there is no mode because none of the numbers is repeated.

In some cases there may be more than one mode. For example, one typist typed 10 pages and the number of mistakes per page are as follows: 5, 1, 0, 1, 2, 1, 2, 3, 2, 4. In this case, both the numbers 1 and 2 appear equal number of times. Therefore, there are two modes: 1 and 2. Similarly, the distribution can be a tri-modal or even multi-modal. For such distributions, the mode as a measure of central tendency has little significance. Mode has very limited use for ungrouped data.

12.3.2 Grouped Data

The method of computing mode is different between discrete distribution and continuous distribution. Let us now study those methods in detail.

Discrete Series

For discrete distribution, i.e., when the values of individual items are known, mode can be determined just by inspection. By inspection you can find out the value of the variate around which the items are most heavily concentrated. For example, study the following frequency distribution:

Size of Item :	20	21	22	23	24	25
Frequency :	15	20	25	45	30	12

In this frequency distribution, 23 has the highest frequency, implying that there is a heavy concentration of items at this value. Therefore, mode is 23.

In a series like this it is easy to obtain mode. Difficulty arises when nearly equal concentrations are found in two or more neighbouring classes; i.e., there is a small difference between the maximum frequency and the frequency preceding it or succeeding it. To locate a modal class in such situations, there is a need for Grouping and Analysis.

Grouping Table : A grouping table has six columns as explained below:

Column 1 : It is of class frequencies written against each class.

Column 2 : Frequencies are grouped in this column in two's, and totals are found. Then the highest total is marked or circled.

Column 3 : Leaving first frequency from the top, the remaining frequencies are again grouped in two's and the highest total is marked.

Column 4 : Starting from the top, frequencies are grouped in three's, their totals are obtained and the highest total is marked.

Column 5 : Leaving first frequency, they are again grouped in three's. Their totals are obtained and the highest total is marked.

Column 6 : Leaving the first two frequencies from the top, remaining frequencies are grouped in three's. Their totals are calculated and the highest total is marked.

Analysis Table : After preparing a grouping table, an analysis table is prepared. It is two-fold : 1) vertical (i.e., stubs) where the column numbers, as obtained in a grouping table, are taken and 2) horizontal (i.e., captions) where the values of the variate (or the classes) are taken. Now you take the grouping table, where you have marked or circled highest frequencies in every column, Take these circled frequencies in turns along with the corresponding values of the variate. In the analysis table under these values and in the row corresponding to relevant column number, tally bars are placed, The number of bars placed in each column of an analysis table are totalled. The maximum of these totals is marked. The value of the variate corresponding to it is the mode or the modal class. Let us study the preparation of grouping and analysis tables by taking an illustration.

Illustration 1

Find the mode (M_o) for the following information on the marks obtained by the students:

Marks :	55	60	61	62	63	64	65	66	68	70
No. of Students :	4	6	5	10	20	22	24	6	2	1

Solution

As you notice here, the difference between the highest frequency (i.e. 24) and the two frequencies preceding it (i.e., 22 and 20) is very small. The frequency which is next to the highest frequency (i.e., 6) also is very small. Therefore, grouping has to be done to ascertain the modal class.

Grouping Table

Marks	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
55	4		×		×	×
60	6	10		15		×
61	5		11		21	
62	10	15				35
63	20		30	52		
64	22	42			66	
65	24		46			52
66	6	30		32		
68	2		8		9	
70	1	3	×	×		×

Analysis Table

Col. No.	Marks 55	60	61	62	63	64	65	66	68	70
1							I			
2					I	I				
3						I	I			
4				I	I	I				
5					I	I	I			
6						I	I	I		
Total				1	3	5	4	1		

The highest total in the analysis table is five. The item corresponding to it is 64. Therefore, the mode (M_o) is 64. It may be noted here that the highest frequency (as shown in data) is for 65, whereas grouping and analysis tables indicated concentration of frequencies around 64. Thus, the correct value of mode is 64.

Continuous Series (i.e. data with class intervals)

In the case of contiquous series, (i.e. data with class intervals) which have equal class intervals throughout, there are two major steps in computing the mode.

Step 1 : Ascertain the modal class by preparing grouping table and analysis table exactly in the same way as discrete series. The minor difference in the procedure is that different classes of the given frequency distribution are taken vertically.

Step 2 : Having located correctly a modal class mode (M_o) is obtained by interpolation by using any of the following formulas:

a) $$M_o = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

where l = lower limit of the modal class

i = class interval

$\Delta_1 = f_1 - f_o$

$\Delta_2 = f_1 - f_2$

f_1 is the frequency of the modal class

f_0 is the frequency of the class preceding the modal class

f_2 is the frequency of the class succeeding the modal class.

By substituting the values of A, and A, in the above formula :

$$M_o = 1 + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

$$= 1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Note: If $(2f_1 - f_0 - f_2)$ is zero, the formula becomes meaningless. If any numerator or denominator becomes negative, then the formula does not give valid result. In that case it should be taken as:

$$M_o = 1 + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times i$$

where $|f_1 - f_0|$ and $|f_1 - f_2|$ means absolute values of the difference i.e., difference neglecting signs.

b) The mode also can be calculated by using the upper limit of the modal class.

$$M_o = u - \frac{f_1 - f_2}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

c) Where the modal class is other than the one containing the maximum frequency, the following formula is more suitable :

$$M_o = 1 + \frac{f_2}{f_0 + f_2} \times i$$

Notes :

1) If the very first class of the frequency distribution is the modal class, the f_0 is taken as zero. If modal class is the last group, then f_2 is taken as zero.

2) These formulas hold good only for the distributions with equal class intervals. Why is it so? The reason is simple. If two class intervals of size 10 and 20 have frequencies 15 and 18 respectively, then on simple comparison it appears frequency 18 is larger than 15. But mode is concerned with concentration of items. Concentration for the first group is 15/10 or 1.5 items per unit length of class interval. While in the second case it is only 18/20 or 0.9 items per unit length of class interval. Thus, from the point of view of determining mode, frequency 18 for class interval size 20 is less than the frequency 15 for the class interval size 10. Therefore, direct comparisons of frequencies can only be made when class intervals are equal.

3) For the distributions with unequal class intervals, first the class intervals are made equal assuming that frequencies are uniformly distributed or by combining groups and then apply the usual formula.

Illustration 2

For the following frequency table, calculate the mode:

Monthly Rent Paid (Rs.)	No. of Families Paying the Rent
20 - 40	6
40 - 60	9
60 - 80	11
80 - 100	14
100 - 120	20
120 - 140	15
140 - 160	10
160 - 180	8
180 - 200	7
	100

Solution

By inspection the modal class appears to be **100-120,** but let us verify by grouping.

Grouping Table

Monthly Rent (Rs.)	Col. 1	Col. 2	Col.3	Col. 4	Col.5	Col. 6
20 - 40	6		×		×	×
		15				
40 - 60	9			26		×
			20			
60 - 80	11				34	
		25				
80 - 100	14					(45)
			(34)			
100 - 120	(20)			(49)		
		(35)				
120 - 140	15				(45)	
			25			
140 - 160	10					33
		18				
160 - 180	8			25	×	
			15			
180 - 200	7	×			×	×

Analysis Table

Col. No.	Monthly Rent (Rs.)							
	20-40	40 - 60	60-80	80-100	100-120	120-140	140-160	160-180	180-200
1					I				
2					I	I			
3				I	I				
4				I	I	I			
5					I	I	I		
6			I	I	I				
Total			1	3	6	3	1		

The highest total being 6, the modal group is **100 - 120.**

Applying the formula :

$$M_o = l + \frac{f_1 - f_0}{2fi - f_0 - f_2} \times i$$

$$= 100 + \frac{20 - 14}{2(20) - 14 - 15} \times 20$$

$$= 100 + \frac{6}{11} \times 20$$

$$= 100 + 10.91$$

$$= 110.91$$

∴ Mode of monthly rent is Rs. **110.91**

Illustration 3
Calculate the mode from the following data:

Size	Frequency
0 - 9	3
10 - 19	4
20 - 29	8
30 - 39	7
40 - 49	6
50 - 59	3

Solution
By inspection, it is difficult to ascertain the modal class. Therefore, we have to resort to grouping.

Grouping Table

Size	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
0 - 9	3		×		×	×
10 - 19	4	7		15		×
20 - 29	(8)		12		(19)	
30 - 39	7	(15)				(21)
40 - 49	6		(13)	(16)	×	
50 - 59	3	9	×		×	×

Analysis Table

Col. No.	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59
1			I			
2			I	I		
3				I	I	
4				I	I	
5			I	I		
6			I	I	I	
Total			4	5	3	

Prom the analysis table, it is obvious that 30-39 is the modal class. But the maximum frequency lies in class 20-29. Therefore, a more suitable formula for calculating the mode is:

$$M_o = l + \frac{f_2}{f_0 + f_2} \times i$$

$$= 29.5 + \frac{6}{8+6} \times 10 \qquad (29.5 \text{ being real limit})$$

$$= 29.5 + \frac{60}{14}$$

$$= 29.5 + 4.29$$

$$= 33.79$$

Therefore, mode is 33.8, You may note that a different result will be obtained if mode is calculated by the following formula :

$$M_o = l + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times i$$

$$= 29.5 + \frac{|7-8|}{|7-8| + |7-6|} \times 10$$

$$= 29.5 + \frac{1}{1+1} \times 10$$

$$= 29.5 + 10/2$$

$$= 34.5$$

You should note that the mode is 34.5 under this method whereas under the earlier method it is 33.8. If you use the formula $M_o = 1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$ denominator will become zero and the numerator will be negative and, therefore, this formula is not applicable. It is important to note that unlike arithmetic mean and median, the different methods of calculating mode can give different results.

12.3.3 Smooth Data

When the data shows more or less uniform movement, it is called the smooth data. For such data mode can be obtained easily without using any of the above formulas. It can be worked out by'a very simple calculation. The rules to be followed for computing mode for smooth data are as under : When $f_0 = f_2$ i.e., the frequencies neighbouring the modal class frequency are equal, the mode is the mid-point of the two limits of the modal class. Study the following illustration carefully.

Size (x)	:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency (f)	:	1	5	15	20	15	6	1

The highest frequency being 20, the modal class here is 30-40. Since each of the two frequencies neighbouring the maximum frequency are equal (i.e., 15), the mode is the simple mean of 30 to 40.

Therefore, $M_o = \frac{30 + 40}{2} = 35$

You may verify whether the result obtained by this formula is the same as the result obtained by the methods suggested earlier for the grouped data, Whenever $f_0 = f_2$ and both f_0 and f_2 are less than f_1 this will always happen. When $f_0 \neq f_2$ (i.e., the two frequencies neighbouring the modal frequency are not equal) and the difference between the neighbouring frequency and the modal frequency is not very large, the mode is the weighted mean of the two limits—upper (u) and the lower (l) of modal class—the weights being the neighbouring frequencies falling on either side of a modal class. Therefore $M_o = \frac{lf_0 + uf_2}{f_0 + f_2}$. For an example, study the following illustration :

'Size	:	0-10	10-20	20-30	30-40	49-50	50-60	60-70
Frequency	:	500	610	740	748	745	690	500

Here the modal class is 30-40 corresponding to the highest frequency 748 (f,). Two neighbouring frequencies are 740 (f_0) and 745 (f_2) which are not equal and they do not differ much from f_1. The modal class is 30-40, '1' is 30 and 'u' is 40

$$\therefore M_o = \frac{30 \times 740 + 40 \times 745}{740 + 745}$$

$$= \frac{52,000}{1,485}$$

$$= 35.02$$

The result derived by this method will always be the same as obtained by using the formula : $M_o = 1 + \frac{f_i}{f_0 + f_2} \times i$ You may verify it.

Illustration 4

From the data given below, find the mode.

Age in Years	:	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
No. of Persons:		50	70	80	180	150	120	70	50

Solution

The highest frequency is in the group 35-40. But concentration of frequency appears to be around the group 40-45. So we do grouping for ascertaining the modal class.

Grouping Table

Ages	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
20 - 25	50		x		x	x
		120				x
25 - 30	70		1 0	200		
30 - 35	80				330	
		260				
35 - 40	180		330			410
40 - 45	150			450		
		270				
45 - 50	120				340	
			190			
50 - 55	70			x		240
		120				
55-60	50		x	x	x	

We observe here that class **40-45** participates in maximum frequency in Columns **2, 3, 4, 5** and 6, **(i.e.,** 5 times out of six columns) and class 35-40 participates only 4 times. You **may** verify it by **analysis** table.

using the formula $M_o = 1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

$$M_o = 40 + \frac{150 - 180}{2 \times 150 - 180 - 120} \times 5 = 40 + \frac{-30}{0} \times 5$$

So mode cannot be determined as $2f_1 - f_0 - f_2 = 2 \times 150 - 180 - 120 = 0$. Therefore, we will use the following formula :

$$M_o = 1 + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times i$$

$$= 40 + \frac{|150 - 180|}{|150 - 180| + |180 - 120|} \times 5$$

$$= 40 + \frac{30}{30 + 60} \times 5$$

$$= 40 + \frac{5}{3}$$

$$= 40 + 1.67$$

$$= 41.67$$

∴ **Modal** Age = 41.67

12.3.4 Empirical Method

In a symmetrical distribution (like the one taken in section 12.3.3) the values of mean, median and mode coincide. You **can verify** it. But in the case of distribution which is not symmetrical **(i.e.,** when frequencies at equal distance from central class are not equal), there are **two possibilities** :

1) **When** there is **greater** concentration in lower values, such distribution is known as **positively skewed distribution.** As shown **in** Figure 12.1, this type of distribution shall have a longer tail on right hand side. In this type of distribution, the value of the mean **is** the highest, **the** value of the mode is the lowest and median lies between mean and mode. The distance between mean and **median** is about one-third the distance between mean and mode.

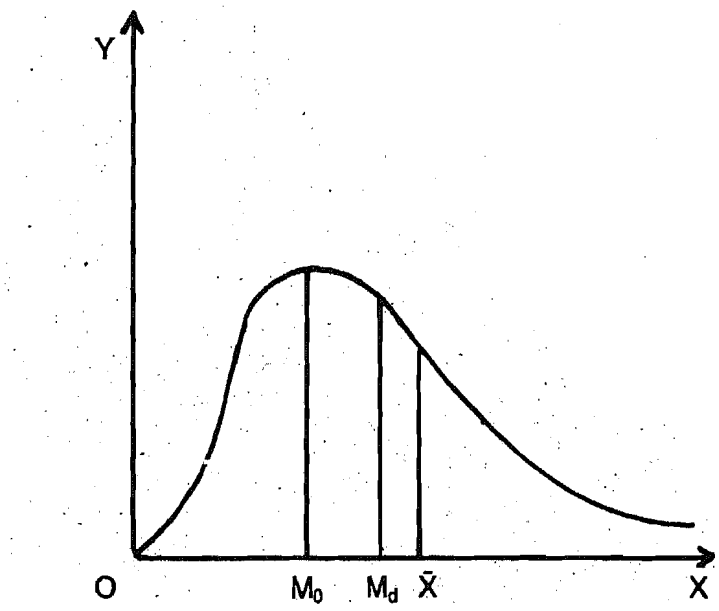


Figure 12.1 Positively Skewed Distribution

2) There may be greater concentration of the items in higher values. Such a distribution is known as negatively skewed distribution. Study Figure 12.2 carefully. You should note that this type of distribution has a longer tail on left band side. Here the mean will be the lowest, the mode will be the highest and median lies about one-third the distance from mean towards mode.

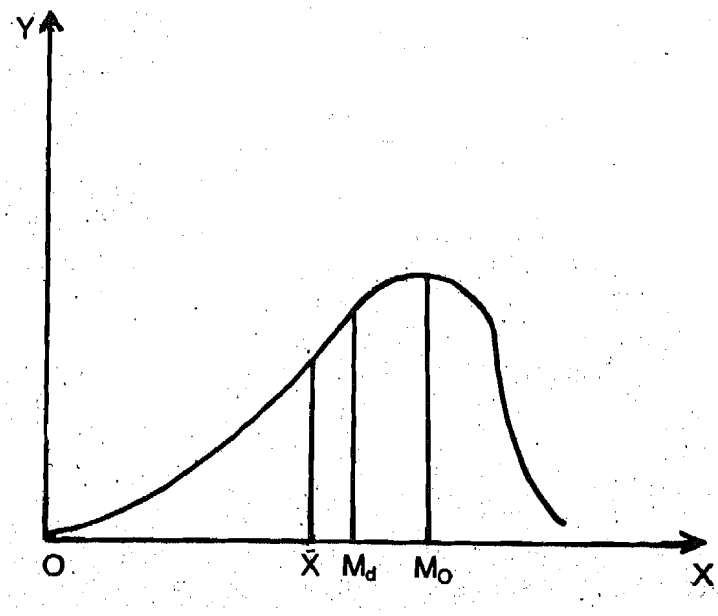


Figure 12.2 Negatively Skewed Distribution

The relationship between mean, mode and median in the two situations explained above is established by Karl Pearson using the following formula':

Mean − Mode = 3 (Mean − Median). This gives

Mode = Mean − 3 (Mean − Median)

= '3 Median − 2 Mean

or $M_o = 3M_d - 2\bar{X}$

This is an empirical relationship because this has been observed to be true in most of the moderately skewed data. It may not be true in extreme situations. A mathematical proof for it is not possible. In this situation where the mode is ill-defined, its value can be obtained empirically by using the above formula.

Illustration 5

Find the mode from the following table :

Size of the Items	Frequency
40 - 49	7
50 - 59	9
60 - 69	10
70 - 79	6
80 - 89	13
90 - 99	10
100 - 109	12
110 - 119	7

Solution

By inspection, the modal class is not clear. Hence, we have to do grouping and analysis.

Grouping Table

Size	Col. 1	Cal. 2	Col. 3	Col. 4	Col. 5	Col. 6
40 - 49	7				x	x
		16				
50 - 59	9		19	26		x
60 - 69	10					
		16				29
70 - 79	6		19			
80 - 89	13	23		29		
90 - 99	10		22		35	
100 - 109	12			x		29
		19				
110 - 119	7		x	x	x	

Analysis Table

Col. No.	60-69	70 - 79	80 - 89	90 - 99	100 - 109	110 - 119
1			I			
2			I	I		
3				I	I	
4		I	I	I		
5			I	I	I	
6	I	I	I	I	I	
Total	1	2	5	5	3	

In the analysis table maximum total **5** occurs twice. The mode, therefore, is ill-defined and is to be determined empirically by using the formula: $M_o = 3M_d - 2\bar{x}$. You may check yourself that here Median = **83.84** and $\bar{x} = 80.14$.

$\therefore M_o = 3\,(83.84) - 2\,(80.14)$

$= 251.52 - 168.28$

$= 91.24$

$\therefore$ Mode = 91.24

Check Your Progress A

1) Define mode.

.....

.....

.....

2) State the various formulas for the computation of mode.,

.....

.....

.....

.....

.....

3) What is the empirical relationship between arithmetic mean. median, and mode?

.....

.....

.....

.....

.....

4) For a frequency distribution, the mean is 26.8 and the median is 27.9. Find the value of mode.

.....

.....

.....

.....

5) State whether the statements given below are True or False.

i) Mode is a unique value in a distribution.

ii) Coniputation of mode neglects the extreme values of the distribution.

iii) Mode of a data cannot be greater than arithmetic mean.

iv) Mode is always found in a class with highest frequency.

v) Even though mode can be computed from data, it is not a mathematical average.

6) Fill in the Blanks:

i) When the purpose is to know the point of the ..... concentration, mode is preferred.

ii) Mode and median are ..... measures.

iii) There are two modes in the following series : 1, 2, 1, 2, 3, 2, 0, 1. They are ..... and .....

iv) In a series when nearly equal concentration, found in two or more neighbouring classes, we prepare ..... and ..... tables to find mode.

v) If $2f_1 - f_0 - f_2$ is zero, the mode is obtained by the formula .....

vi) Read the following data :

X :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
F :	2	8	12	18	12	5	3

Here the mode is the simple mean ..... of and .....

vii) Read the following data:

X :	10-20	20-30	30-40	40-50	50-60	60-70
F :	100	125	220	228	222	150

In this case'mode is a weighted mean of thk two limits of modal class 40 and 50, the weights being ..... and .....

viii) If two values in a given data set occur more often than any others, the distribution, is said to be .....

12.4 GRAPHICAL DETERMINATION OF MODE

You have studied various methods of computing the mode. In fact, like median, mode also can be determined graphically., Determination of mode graphically involves the following procedure:

1) Draw a histogram for the given frequency distribution. A partial histogram can also be drawn by using only three classes — pre-modal, modal and post-modal. You have studied about histogram in Unit 9.

2) The top right comers of the highest rectangle (modal class rectangle) and the preceding rectangle are joined by a straight line. Similarly, the top left corners of the highest rectangle and the rectangle just on its right are joined by a straight line.

3) Draw a perpendicular to x-axis from the point of intersection of these two straight lines.

4) The point where it meets the x-axis gives the value of the mode.

Let us understand this proce'dure clearly through an illustration.

Illustration 6

Find the mode of the following data graphically and also check the result through calculation:

Class Interval :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	4	18	30	42	24	10	3

Solution

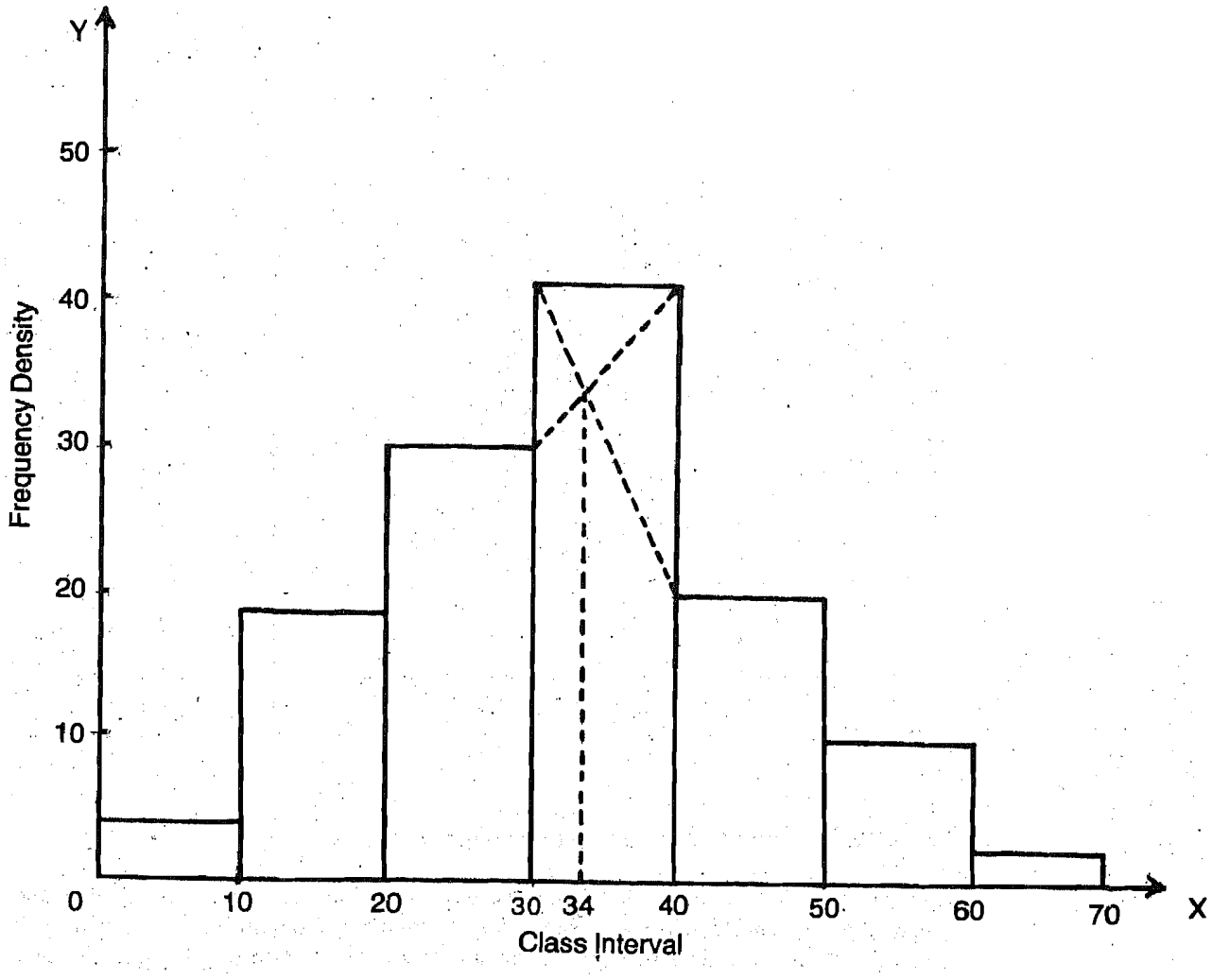


Figure 12.3. Histogram and Class Interval Determination of Mode

By using the usual formula $M_o = 1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

$$M_o = 30 + \frac{42 - 30}{84 - 30 - 24} \times 10$$

$$= 30 + \frac{12}{30} \times 10$$

$$= 30 + 4$$

$$= 34$$

You must note that the value of mode obtained here is the same as the value obtained graphically. But, if you compute the mode by formula $M_o = 1 + \frac{f_2}{f_0 + f_2} \times i$ the result will not be the same as obtained by graph. The logic behind the graphic method and formula based on f_0, f_1, f_2 is same. The details of the logic are beyond the scope of this course.

There is one limitation of the graphical method of determining mode. When modal class is adjoining to the class with highest frequency, mode cannot be determined graphically in the modal class. It can only be determined from a class with highest frequency. Thus, mode calcula'ted graphically will not be a proper mode. To understand this, let us determine mode graphically for the data in Illustration 3 discussed earlier.

Solution

. Here the class intervals are of inclusive type. So they have to be first converted to real limits before drawing the histogram. Now Iook at Figure 12.4 carefully.

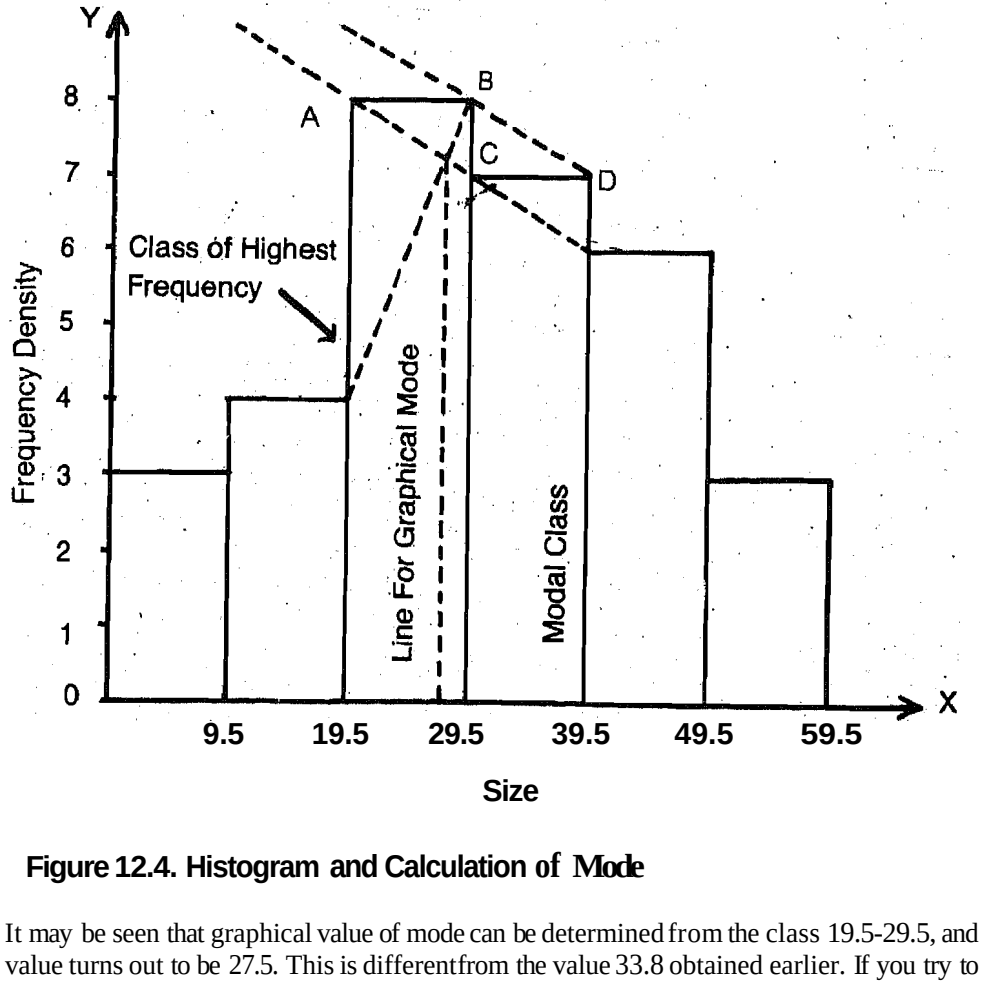


Figure 12.4. Histogram and Calculation of Mode

It may be seen that graphical value of mode can be determined from the class 19.5-29.5, and value turns out to be 27.5. This is different from the value 33.8 obtained earlier. If you try to determine the mode in the modal group, you have to join points A to C and B to D. The two lines AC and BD do not intersect in the modal class. Hence mode cannot be determined in the modal class by graphical method.

If the modal group and the group with highest frequency are not adjoining and separated by

two to three groups, the mode can be determined graphically in both the groups. Out of these two modes, mode of first preference can be decided by looking to the height of the perpendicular drawn from the point of intersection to the x-axis. Such distributions can be termed as **bi-modal.** Let us take an illustration to explain this.

Illustration 7

The following distribution gives 'over time work' done by 100 employees of a company during a month. Determine the mode graphically.

Over time Hours :	10-12	12-14	14-16	16-18	18-20	20-22	22-24	24-26	26-28	28-30
No. of Employees :	3	5	16	21	17	6	4	23	3	2

Solution

By preparing grouping and analysis tables, you can easily verify that the modal class is 16-18. But the highest frequency is in the group 24-26 which is at a distance of 3 groups from modal class. Now look at Figure 12.5 for histogram and graphical determination of mode.

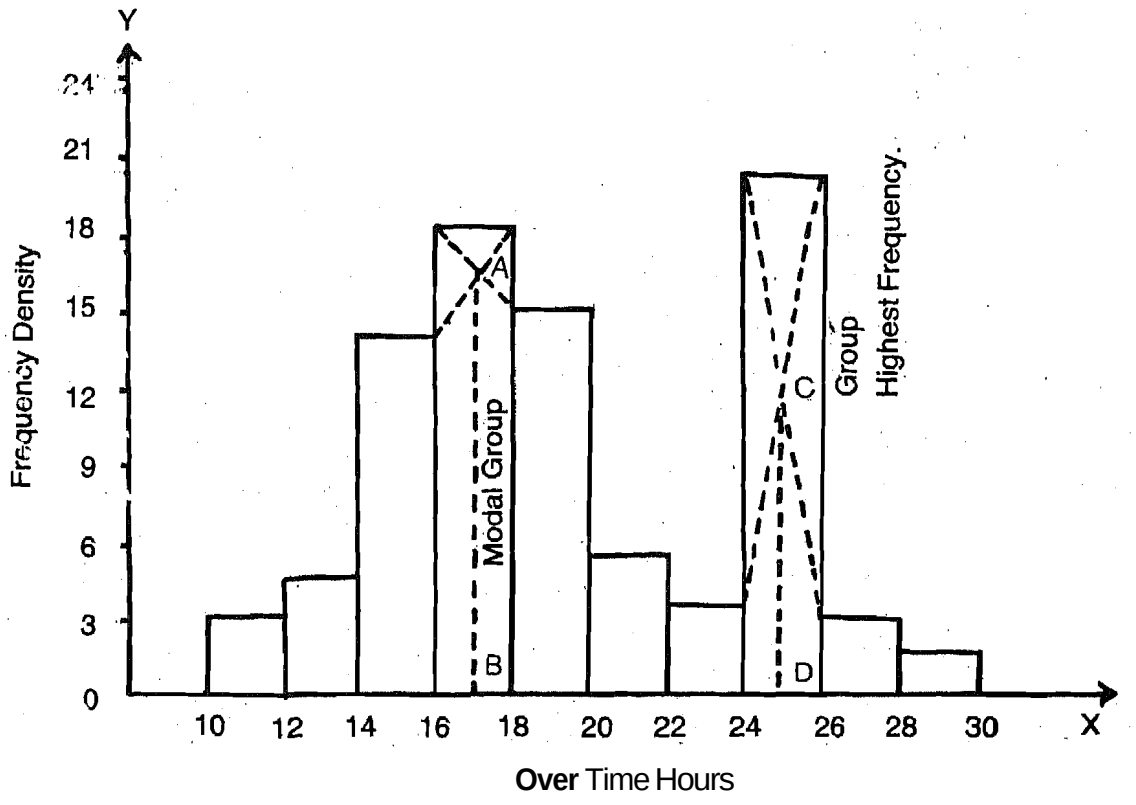


Figure 12.5. Histogram **and** Calculation of Mode

Point B in the modal class gives one mode. This is 16.9, approximately. Point D in the class with highest frequency gives another mode. This is 25.1, approximately. The length of the lines AB and CD gives the extent of concentration of items at two points B and D. As the length AB is greater than CD, there is a greater amount of concentration of frequencies at mode B than at mode D. Thus, the value of B (i.e. 16.9) is the mode of first preference value and D value (i.e. 25.1) will be taken as mode of second preference. By computation method also mode can be determined in both the groups. The values obtained will be approximately the same as the one determined by graph. As pointed out earlier such a data may be termed as bi-modal. However, **a perfect bi-modal data is one in which concentration of frequencies at the two modes is exactly equal.**

1125 MERITS AND LIMITATIONS OF MODE

Merits

1) In certain situations mode is the only suitable average, e.g., modal size of garments, modal size of shoes, modal wages, modal balance of depositors in a bank, etc.

2) It is used to describe qualitative phenomena. For instance, if a printing press turns out five impressions which we rate very sharp, sharp, sharp, blurred and sharp, then the modal value is sharp.

3) For the preference of consumers' product, the modal preference is regarded. A restaurant owner who specialises in one dish may wish to know the modal preference of his potential clientele.

4) In the case of skewed distribution, mode is the indicator of the point of heaviest concentration.

5) It is very profitably used in market research.

6) Even if one or more classes are open-ended, mode can be used.

Limitations

1) Too often, there is no modal value. It is a useless measure, when there are more than one mode.
2) It is not capable of further algebraic treatment.
3) It is an ill-defined measure. Therefore, different formulas yield somewhat different answers.
4) It is not based on all the items of the data.
5) The value of the mode is affected'significantlyby the size of the class-intervals,
6) Although a mode is the value of a variate that occurs most frequently, its frequency does not represent a majority of the total frequencies.

Check Your Progress B

1) Why is it usually better to calculate a mode from grouped rather than ungrouped data?
   a) The ungrouped data tend to be bi-modal.
   b) The mode for the grouped data will be the same, regardless of the skewness of the distribution.
   c) Extreme values have less effect on grouped data.
   d) The chance of an unrepresentative value being chosen as the mode is reduced.

2) In which of the cases would a mode be most useful as an indicator of central tendency?
   a) Every value in a data set occurs exactly once.
   b) All but three values in a data set occur once, three values occur 100 times each.
   c) All values in a data set occur 100 times each.
   d) Every observation in a data set has the same value.

3) When bell-shaped distribution is symmetrical and has one mode, the highest point on the curve is referred to as a) Range; b) Mode; c) Median; d) Mean; e) All of these; f) b, c, and d, but not a.

4) State whether the following statements are True or False.
   i) Graphical method and computation methods of finding a mode always give identical values.
   ii) As mode can be computed from the data, it is capable of algebraic treatment.
   iii) Mode at times can be used to describe qualitative phenomena.
   iv) Mode plays an important role in checking the symmetry of the data.
   v) If it is not possible to compute mode, you cannot find it graphically also.

5) Fill in the blanks:

i) If the mean and median of a moderately asymmetrical series are 26.8 and 27.9 respectively, the most probable mode will be .....

ii) The approximate value of mode is 52 with mean 58 and median.....

iii) If the data set has only one mode and it is less than the mean, it can be concluded that the graph of the data is skewed to the .....

iv) For a moderately skewed distribution, the empirical relation is given as M_o = ...

v) The mode can be graphically determined by constructing the ..... and using the ..... rectangle and two ..... rectangles.

vi) For preference of consumers' product, the .....preference is considered.

vii) Mode suffers from sampling.....

12.6 SOME ILLUSTRATIONS

Illustration **8**

Estimate the value of arithmetic mean if mode is 15.3 and median is 14.2

Solution

The empirical relation between mean, median and mode is:

$$M_o = 3M_d - 2\bar{x}$$

Substituting the values of M_o and M_d

$$15.3 = 3 \times 14.2 - 2\bar{x}$$

$$2\bar{x} = 42.6 - 15.3$$

$$2\bar{x} = 27.3$$

$$\bar{x} = 13.65$$

Illustration **9**

With the help of empirical relation between M_o, M_d, and $\bar{x}$ show **that**

i) $M_d = M_o + \frac{2}{3}(\bar{x} - M_o)$

ii) $\bar{x} = M_d + \frac{1}{2}(M_d - M_o)$

Solution

The empirical relation between mean, median and mode is:

$$M_o = 3M_d - 2\bar{x}$$

i) $M_o = 3M_d - 2\bar{x}$

$$M_o + 2\bar{X} = 3M_d$$

$$\frac{1}{3}(M_o + 2\bar{x}) = M_d$$

$$M_d = \frac{1}{3}M_o + \frac{2}{3}\bar{x}$$

$$= M_o - \frac{2}{3}M_o + \frac{2}{3}\bar{x}$$

$$= M_o + \frac{2}{3}(\bar{x} - M_o)$$

$$\therefore M_d = M_o + \frac{2}{3}(\bar{x} - M_o)$$

$$\text{Median} = \text{Mode} + \frac{2}{3}(\text{Mean} - \text{Mode})$$

ii) $M_o = 3M_d - 2\bar{x}$

$$2\bar{x} = 3M_d - M_o$$

$$\bar{x} = \frac{3}{2}M_d - \frac{1}{2}M_o$$

$$\text{Mean} = \text{Median} + \frac{1}{2}(\text{Median} - \text{Mode})$$

Illustration 10

The following table gives the age (in years) of employees of a firm. The modal age is 32 years. Find the missing frequency.

Age in Years :	20-25	25-30	30-35	35-40	40-45
No of Employees:	5		18	9	6

Solution

Let us assume that the missing frequency is 'F'. As the mode is 32, the modal group is 30-35.

Now $M_o = 1 + \frac{f_1 - f_0}{2f1 - f_0 - f_2} \times i$

where $1 = 30$, $f_0 = F$, $f_1 = 18$, $f_2 = 9$, $i = 5$ and $M_o = 32$

Substituting the x-values:

$$32 = 30 + \frac{18 - F}{2 \times 18 - F - 9} \times 5$$

$$2 = \frac{18 - F}{27 - F} \times 5$$

$$54 - 2F = 90 - 5F$$

$$3F = 36$$

$$F = 12$$

∴ Missing frequency is 12.

Illustration 11

Calculate mode from the data given below:

Profit (Rs. in lakhs) :	0-5	5-10	10-20	20-30	30-50
No. of Companies :	4	6	15	18	20

Solution

Here the class intervals are not equal. In such cases two methods can be used:
i) Rewriting the data with equal class intervals, ii) Using empirical relationship.

i) On combining the first two groups, class intervals will become 0-10. Next two class intervals are of size 10. The last class interval is of size 20. It can be divided into two i.e. 30-40 and 40-50. Assuming frequencies as uniformly distributed, both such groups will have frequencies of 10 each. Thus, the given data can be written as:

Profit (Rs. in lakhs) :	0-10	10-20	20-30	30-40	40-50
No. of Companies :	10	15	18	10	10

It is clear that the modal class is 20-30

$$\text{Now Mode} = 1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Substituting the values of 1, f_0, f_1, f_2, and i

$$M_o = 20 + \frac{18 - 15}{2 \times 18 - 15 - 10} \times 10$$

$$= 20 + \frac{3}{11} \times 10$$

$$= 20 + 2.7$$

$$= 22.7$$

∴ Mode of the profit is Rs. 22.7 lakhs.

ii) You may verify arithmetic mean = Rs. 24.3 lakhs and the median is Rs. 23.6 lakhs

Now mode = 3 Median − 2 Arithmetic Mean

∴ Mode = 3 x 23.6 x − 2 x 24.3

= 70.8 − 48.6

= 22.2

∴ Mode of profit is Rs. 22.2 lakhs.

You may note that the modes amved at by the two methods differ. It is because, as you know, mode is not rigidly defined.

Illustration **12**

As a manager of a transport company you want to buy 100 tyres from either producer A or producer B. The price of the two tyre types is same. The following information is available about the average distance run by these two types of tyres :

Firm	Average distance run	
	Arithmetic Mean (km.)	Mode (km.)
Type A	**35,000**	**32,000**
Type B	**32,000**	35,000

i) which type would you buy?

ii) If you want to buy one tyre for your own car, will your decision be the same?

Solution

i) AM × No. of Items = Total value of items. So if you buy tyres from producer A, the total distance run by all 100 tyres would be 100 x 35,000 = 35,00,000 kms. If you buy from producer B, the total distance run would be 100 x 32,000 = 32,00,000. As the total distance run in first case is greater, you would prefer tyres of producer **A.**

ii) When you are buying only one tyre, it is not necessary that the tyre bought will give the same mileage as arithmetic mean. On the other hand, it is quite likely that the tyre you bought may give mileage equal to mode, the value around which you have maximum concentration of items. As the mode of producer B is higher in this case, you will prefer producer B product.

It may be noted that when large number of tyres are bought, some tyres may give mileage equal to arithmetic mean and others may give more than arithmetic mean. If the selection is done randomly, the mean of the distance run by the selected tyres would be almost same as the mean value claimed by the producer. Hence, in the case (i) arithmetic mean was used to assess which type of purchase gives greater service,

12.7 LET US SUM UP

The mode is the value of the variate around which the other items tend to concentrate most heavily. It can be computed for both ungrouped and grouped data. However, for ungrouped data it has a limited use. For a discrete distribution, mode is that value of the variate around which the items are most heavily concentrated. Where there are nearly equal concentrations in two or more neighbouring classes to a class with highest frequency, it is'difficult to determine the mode. In such cases 'grouping and analysis tables' are prepared to ascertain the modal class. For a continuous distribution, after having located a modal class, mode is calculated by using different interpolaltive formulas. For the data set displaying uniform movement, mode is obtained by using simple rules where it could be either a simple mean of two limits of a modal class or a weighted mean of them. For moderately skewed distribution, mode is obtained by an empirical relationship $M_o = 3 M_d - 2 \bar{x}$ Mode also can be graphically determined by constructing the histogram and using the highest rectangle and two of its neighbouring rectangles.

Mode is very useful in situations like finding a modal size of shoes, modal size of garments, modal wages, etc. It is also used to describe the qualitative phenomenon and to indicate modal preference of consumers for consumer products. Mode suffers from certain limitations such as incapability of further algebraic treatment, ill-defined nature, non-existence, presence of more than one mode, etc.

12.8 KEY WORDS AND SYMBOLS

Analysis Table: The table which helps to ascertain the modal class showing the maximum frequency occurring in different columns.

Bi-modal Distribution: A distribution of data in which two values occur more frequently than the rest of the values in the data set.

Empirical Relationship of Averages: The relationship that exists between averages in a moderately skewed distribution viz., $M_o = 3M_\alpha - 2\bar{x}$

Grouping Table: The table which has six columns, used for ascertaining a modal class.

Mode: The value of the variate around which the other items tend to concentrate most heavily.

Negatively Skewed Distribution: The distribution wherein there is a greater concentration in higher values with a longer tail on left hand side.

Positively Skewed Distribution: The distribution where there is a greater concentration in lower values with a longer tail on the right hand side.

List of symbols

In addition to the list of symbols given under Units 10 & 11, following is the list of symbols used in connection with Mode. The list is on the same lines as is Unit 10.

Difference between modal and next (higher values side) frequency	$f_1 - f_2$, Δ_2, d_2. Where Δ_2 and d_2 are always taken as positive.
Difference between modal and previous (lower values side) frequency	$f_1 - f_0$, Δ_1, d_1. Where Δ_1 and d_1 are always taken as positive.
Frequency of a group next to (higher values side) modal group.	f_2
Frequency of a group previous to (lower values side) modal group.	f_0, f_1 (when modal frequency not denoted by f_1)
Frequency of the modal group	f_1, f_m, f_{mo}, f.
Lower limit of the modal group	l, l_1, L, L_{M_o}
Mode	M_o, Z
Upper limit of the modal group	u, l_2, U, U_m, U_{M_o}

12.9 ANSWERS TO CHECK YOUR PROGRESS

A) 4) 30.1

5) i) False; ii) True; iii) False; iv) False; v) True

6) i) highest; ii) positional; iii) 1, 2; iv) grouping analysis

v) $M_o = l + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times i$ OR $M_o = l + \frac{f_2}{f_0 + f_2} \times i$

vi) 30.40; vii) 220,222; viii) bi-modal;

B) 1) d

2) b

3) f

4) i) False; ii) False; iii) True; iv) True; v) True

5) i) 30.1; ii) 56; iii) right; iv) $3M_e - 2\bar{x}$; v) histogram, highest, neighbouring; vi) modal; vii) instability

12.10 TERMINAL QUESTIONS/EXERCISES

Questions

1) 'Arithmetic Mean, Median and Mode all try to give one main characteristic of the data but in their own way'. Discuss.

2) What is mode? Explain its limitations and uses as a measure of average?

Exercises

I) Find the modal age of married women at first child birth:

Age (Years)	13	14	15	16	17	18	19	20	21	22	23	24	25
No. of Women	37	162	343	390	256	433	161	355	'65	85	49	49	40

(Answer : 18 years)

2) From the following information regarding the wage distribution in a certain factory, determine the modal age:

Weekly Wage (Rs.)	No. of Employees
20 - 40	8
40 - 60	12
60 - 80	20
80 - 100	30
100 - 120	40
120 - 140	35
140 - 160	18
160 - 180	7
180 - 200	5

(Answer: Rs. 113.33)

3) Find the modal size of shoes from the following information:

Size of Shoes	1	2	3	4	5	6	7	8	9	10
Frequency	10	5	13	6	23	32	14	35	8	7

(Answer : 6)

4) The following table gives the relative distribution of sales calls made on Amar Phannaceuticals in the past month. Find the modal calls.

No. of Sales Calls	0	1	2	3	4	5 or more
Relative Frequency:	0.21	0.18	0.38	0.19	0:03	0.01

(Answer: 2 sales calls)

5) Calculate the mode for the following data:

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	24	42	56	66	108	130	154

(Answer: 71.34)

6) Determine the most common salary graphically for the following data and verify it by using an appropriate formula:

Salary (more than Rs.)	100	150	200	250	300	350	400	450
No. of Employees:	100	98	93	83	43	23	12	5

(Answer: 280)

7) Determine the mode graphically and also by computation.

Weight less than (kgs,)	80	85	90	95	100	105	110	115	120	125
No. of Employees	0	5	13	'30	55	75	85	93	120	125

(Answer : 98.1)

8) Obtain the mode for the following distributions without using the usual formulas:

i)

x :	0 - 10	10 - 20	20-30	30-40	40-50	50-60	60-70
f :	1	6	15	20	15	6	1

ii)

x :	48-52	52-56	56-60	60-64	64-68	68-72	72-76
f :	4	8	16	18	15	4	2

(Answers: i - 35, **ii** - 61.93)

9) Estimate the median when arithmetic mean is 27.9 and mode is 25.2. Give the assumptions, if any.

(Answer: 27)

10) What are the modal values of the following distributions?

a)

Hair Colour	:	Black	Brunette	Red Head	Blonde
Frequency	:	I I	24	6	18

b)

Blood Group	:	AB	O	A	B
Frequency	:	4	12	35	16

(Answers: a - Brunette; b - A)

Note : These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.

UNIT 13 GEOMETRIC, HARMONIC AND MOVING AVERAGES

Structure

13.0 OBJECTIVES

After studying this unit, you should be able to:

- define and compute geometric mean and harmonic mean'
- enumerate the properties of geometric mean and harmonic mean
- appreciate the limitations and uses of geometric mean and harmonic mean
- explain the concept of moving average
- use moving average in determining trend of time series
- make a choice of suitable average in a given situation.

13.1 INTRODUCTION'

As you know, the averages can be classified as mathematical averages, positional averages and special averages. You have already studied about arithmetic mean which belongs to the category of mathematical averages, median and mode, which belong to the category of positional averages. In this unit you will study about the two other mathematical averages viz., Geometric Mean and Harmonic Mean. You will also study special average viz., Moving Average, and how to choose a suitable average amongst all types of averages in a given situation.

13.2 . GEOMETRIC MEAN

In the situations where we deal with quantities that change over a period of time, we may be interested to know the average rate of change. In such cases the simple arithmetic mean is not suitable and we have to resort to the geometric mean.

13.2.1 Computation

Like other averages, computation procedure of geometric mean is different for grouped data and ungrouped data. Now let us study these methods.

8) Obtain the mooe for the following distributions without using the usual formulas:

i)

x :	0 - 10	10-29	20-30	30-40	40-50	50-60	60-70
f :	I	6	15	20	15	6	1

ii)

x :	48-52	52-56	56-60	60-64	64-68	68-72	72-76
f :	4	8	16	18	15	4	2

(Answers: i - 35, **ii** - 61.93)

9) Estimate the median when arithmetic mean is 27.9 and mode is 25.2. Give the assumptions, if any.

(Answer: 27)

10) What are the modal values of the following distributions?

a)	Hair Colour	:	Black	Brunette	Red Head	Blonde
	Frequency	:	11	24	6	18
b)	**Blood Group**	:	AB	**O**	**A**	B
	Frequency	:	4	12	35	16

(Answers: a - Brunette, b - **A)**

Note : These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.

UNIT 13 GEOMETRIC, HARMONIC AND MOVING AVERAGES

Structure

13.0 OBJECTIVES

After studying this unit, you should be able to:

- define and compute geometric mean and harmonic mean
- enumerate the properties of geometric mean and harmonic mean
- appreciate the limitations and uses of geometric mean and harmonic mean
- explain the concept of moving average
- use moving average in determining trend of time series
- make a choice of suitable average in a given situation.

13.1 INTRODUCTION

As you know, the averages can be classified as mathematical averages, positional averages and special averages. You have already studied about arithmetic mean which belongs to the category of mathematical averages, median and mode, which belong to the category of positional averages. In this unit you will study abaut the two other mathematical averages viz., Geometric Mean and Harmonic Mean. You will also study special average viz., Moving Average, and how to choose a suitable average amongst all types of averages in a given situation.

13.2 GEOMETRIC MEAN

In the situations where we deal with quantities that change over a period of time, we may be interested to know the average rate of change. In such cases the simple arithmetic mean is not suitable and we have to resort to the geometric mean.

13.2.1 Computation

Like other averages, computation procedure of geometric mean is different for grouped data and ungrouped data. Now let us study these methods.

Ungrouped data

If there are two items in the data series, the square root of the product of these two items is the geometric mean. If there are three items, the cube root of the product of three items is their geometric mean. If there are 'n' items in the series, its geometric mean is the nth root of the product of those items. Let us express it symbolically:

$$\text{Geometric Mean} = \sqrt[n]{x_1\, x_2 \ldots\ldots x_n}$$

where X_1, X_2, X_n refer to the 'n' items of the series. For example, we have three numbers 4, 8, and 16, the geometric mean of these three numbers would be:

$$G.M = \sqrt[3]{4 \times 8 \times 16}$$
$$= \sqrt[3]{512}$$
$$= 8$$

Thus, geometric mean is an average based on the product of items. When the number of, items is three or more, finding their product and extracting its roots becomes difficult. Therefore, computations can be simplified by the use of logarithm. The procedure is as follows:

1) Obtain the logarithm of the different values of the variable and take their total $\Sigma \log x$.
2) Divide it by 'n' (the number of items) and take the antilogarithm of the value so obtained. That gives the Geometric Mean.

Symbolically it can be expressed as follows:

$$\log G.M. = \frac{1}{n} \log (x_1 x_2 \ldots\ldots x_n)$$
$$= \frac{\log x_1 + \log x_2 + \ldots. \log x_n}{n}$$
$$= \frac{\Sigma \log x}{n}$$

Therefore, $G.M. = \text{Antilog} \frac{\Sigma \log x}{n}$

For example, geometric mean of four numbers 20, 65, 83 and 135 will be:

$$G.M. = \text{Antilog} \frac{\log 20 + \log 65 + \log 83 + \log 135}{4}$$
$$= \text{Antilog} \frac{1.3010 + 1.8129 + 1.9191 + 2.1303}{4}$$
$$= \text{Antilog } 1.7908$$
$$= 61.77$$

Illustration 1

Compared to the previous year, the overhead expenses went up by 32% in 1987, by 40% in 1988 and by 50% in 1989. Calculate the average rate of increase in overhead expenses over the three years.

Solution

The increase in overhead expenses is 32%, 40% and 50% in 1987, 1988 and 1989 respectively. This means successively expenses become 132%, 142% and 150% of the previous level. Therefore, at the end of three years the final level will be $\frac{132 \times 140 \times 150}{100 \times 100}$ per cent of the original level.

As these figures are multiplicative in nature, their average will be given by geometric mean,

$x_1 = 132\ x_2 = 140\ x_3 = 150$ and $n = 3$

$$\text{Now G.M.} = \text{Antilog} \frac{\Sigma \log x}{n}$$
$$= \text{Antilog} \frac{\log 132 + \log 140 + \log 150}{3}$$
$$G.M. = \text{Antilog} \frac{2.1206 + 2.1461 + 2.1761}{3}$$
$$= \text{Antilog} \frac{6.4428}{3}$$

= Anti log 2.1476
= 140.5

On an average overhead expenses become 140.5% of previous year's level. Therefore, average rate of increase in overhead expenses is 40.5% (i.e., 140.5 − 100).

Grouped Data

You know how to compute geometric mean for ungrouped data. Now we should discuss the procedure for grouped data. As you know, the grouped data can be in the form of either discrete series or continuous series, we have to follow different procedures for these two types of series.

Discrete Series: When the data is grouped data i.e., in the form of a frequency distribution, the geometric mean is computed as follows:

$$G.M. = \sqrt[n]{x_1^{f_1} x_2^{f_2} \ldots x_r^{f_r}}$$

where $x_1, x_2, \ldots, x_r$ are the different values of the variate x with their respective frequencies $f_1, f_2, \ldots f_r$ and $n = f_1 + f_2 + \ldots f_r = \Sigma f$

$$\log G.M. = \frac{1}{n}(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_r \log x_r)$$

$$= \frac{1}{n}(\Sigma f \log x)$$

$$G.M. = \text{Antilog}\left(\frac{\Sigma f \log x}{n}\right)$$

Continuous Series: The only change in the earlier formula of geometric mean is that you replace 'x' by 'm' which is the mid-value of classes.

Here $G.M. = \text{Antilog}\left(\frac{\Sigma f \log m}{n}\right)$

The procedure followed in both the formulas is as under:

1) Convert the given values of variate x or the mid-value (m) in the case of continuous series into logarithms
2) Multiply them with the respective frequencies and get the product $\Sigma f \log x$ or $\Sigma f \log m$, as the case may be.
3) Divide the product by the total frequency $\Sigma f = n$ and take the antilogarithm of the value so obtained.

Illustration 2

Find out the geometric mean for the following data :

Size	Frequency
7.5 - 10.5	5
10.5 - 13.5	9
13.5 - 16.5	19
16.5 - 19.5	23
19.5 - 22.5	7
22.5 - 25.5	4
25.5 - 28.5	1

Solution

Class interval	Mid-Point (m)	log m	f	f.log m
7.5 - 10.5	9	0.9542	5	4.7710
10.5 - 13.5	12	1.0792	9	9.7128
13.5 - 16.5	15	1.1761	19	22.3459
16.5 - 19.5	18	1.2553	23	28.8719
19.5 - 22.5	21	1.3222	7	9.2554
22.5 - 25.5	24	1.3802	4	5.5208
25.5 - 28.5	27	1.4314	1	1.4314
Total			68	81.9092

$$\log G.M. = \left(\frac{\Sigma f \log x}{n}\right)$$

$$= \frac{1}{68} \times 81.9092$$

$$= 1.2045$$

$$G.M. = \text{Antilog } 1.2045$$

$$= 16.02$$

Geometric Mean **for** Computing Average **Rate of** Change

More often we are interested in the average rate of change in a variable between any two time periods such as annual rate of increase in population, annual rate of increase in GNP, average rate of increase in profit, etc. The methods of computing such rates is similar to that of finding the geometric mean.

For a given series assume P_o is the value at the beginning of the period and P_n is the value at the end of the period. Now, the average growth rate (r) can be obtained by using the following compound interest formula:

$P = P_o (1 + r)^n$, where 'n' is the time-span

$$(1 + r)^n = \frac{P_n}{P_o}$$

$$(1 + r) = \sqrt[n]{\frac{P_n}{P_o}}$$

$$r = \sqrt[n]{\frac{P_n}{P_o}} - 1$$

Illustration 3

The population of a country was 300 millions in 1951. It became 520 millions in 1969. Calculate the percentage compund rate of growth per annum.

Solution

Here P_o is 300: P_n is 520 and n is 18. Let 'r' be the growth rate per annum.

$$\text{Now } 1 + r = \sqrt[n]{\frac{P_n}{P_o}}$$

$$= \sqrt[18]{\frac{520}{300}} \text{ using logarithms}$$

$$\log (1 + r) = \frac{\log 520 - \log 300}{18}$$

$$1 + r = \text{Antilog}\left(\frac{2.7160 - 2.477}{18}\right)$$

$$= \text{Antilog}\left(\frac{0.2389}{18}\right)$$

$$= \text{Antilog } 0.0133$$

$$= 1.031$$

$$r = 1.031 - 1$$

$$= 0.031$$

$\therefore$ Percentage compound growth rate is $100 \times r = 3.1\%$

13.2.2 Weighted Geometric Mean

Like weighted arithmetic mean, we can also calculate the weighted geometric mean. The computational procedure is as follows:

$$\text{Weighted G.M.} = \sqrt[\Sigma w]{x_1^{w_1}, x_2^{w_2}, \ldots . x_N^{w_N}}$$

Where $x_1, x_2, \ldots . x_N$ are the values of the variate and $W_1, W_2, \ldots . W_N$ are the corresponding weights.

Taking logarithms,

$$\text{Log Weighted G.M.} = \frac{W_1 \log x_1 + W_2 \log x_2 + ... + W_n \log x_n}{\Sigma W}$$

$$\text{or log Weighted G.M.} = \frac{\Sigma W \log x}{\Sigma W}$$

$$\text{Weighted G.M.} = \text{Antilog}\left[\frac{\Sigma W \log x}{\Sigma W}\right]$$

illustration 4

Calculate the weighted Geometric Mean from the following information:

Group	Index No.	Weight
Food	300	40
Fuel	200	10
Cloth	250	10
House Rent	150	15

Group	Index No.	Weight	Log x	W. Log x
Food	300	40	2.4771	99.084
Fuel	200	10	2.3010	23.01
Cloth	250	10	23979	23.979
House Rent	150	15	2.1761	32.6415
		$\Sigma W = 75$		$\Sigma W \log x = 178.7145$

$$\text{Weighted G.M.} = \text{Antilog}\left[\frac{\Sigma W \log x}{\Sigma W}\right]$$

$$= \text{Antilog}\left[\frac{178.7145}{75}\right]$$

$$= \text{Antilog } 2.3829$$

$$= 241.50$$

Therefore weighted geometric mean of index numbers is 241.50

13.2.3 Properties of Geometric Mean

Geometric mean has the following important properties:

1) In a given series, if each item is substituted by geometric mean of the series, the product of the items remains unaltered. For example, the geometric mean of the items.4, 8 and 16 is 8. Therefore, $4 \times 8 \times 16 = 8 \times 8 \times 8 = 512$.

2) The value of geometric mean balances the ratio deviations of the observations from it. In other words, the geometric mean of two numbers 'a' and 'b' is 'G', and the two ratios a : G and G : b are equal. It means a/G is equal to G/b. For example, geometric mean of 4 and 16 is $\sqrt{4 \times 16}$ or 8. The ratio 4/8 and 8/16 should be equal, which is a fact:

3) It lends itself to algebraic treatment. If geometric means of two or more groups are given, the geometric mean of the combined group can be obtained, as follows:

$$\text{Combined G.M.} = \text{Antilog}\left[\frac{N_1 \log GM_1 + N_2 \log GM_2 + ... N_K \log GM_K}{N_1 + N_2 + ... + N_K}\right]$$

where GM_1 = Geometric mean of the first group
GM_2 = Geometric mean of the second group
GM_K = Geometric mean of the Kth group.

For example, let 100 items have GM = 50 and 200 items have GM = 40. Then the combined geometric mean will be :

$$\text{Weighted G.M.} = \text{Antilog}\ \frac{100 \log 50 + 200 \log 40}{300}$$

$$= \text{Antilog}\left[\frac{100 \times 1.6990 + 200 \times 1.6021}{300}\right]$$

= Antilog 1.6344
= 43.09

4) As compared to arithmetic mean, the geometric mean is less affected by large items. It may be stated that the geometric mean has bias towards small items while arithmetic mean has bias towards large items. For example, let us take the five items : 2, 3, 5, 10 and 100.

$$\text{Arithmetic mean} = \frac{2+3+5+10+100}{5}$$

$$= 24$$

$$\text{Geometric mean} = \text{Antilog}\left[\frac{\log 2 + \log 3 + \log 5 + \log 10 + \log 100}{5}\right]$$

$$= \text{Antilog}\left[\frac{0.3010 + 0.4771 + 0.6990 + 1.0000 + 2.0000}{5}\right]$$

$$= \text{Antilog } 0.8954$$

= 7.86 approximately

You may note that arithmetic mean is 24 which is sufficiently larger than geometric mean 7.86. So geometric mean has a tendency to be pulled towards small items, while arithmetic mean has a tendency to be pulled towards large items.

13.2.4 Uses and Limitations

Uses

1) For computing the averages of ratios and percentages, geometric mean is the most suitable average.

2) As it has bias towards lower values, it is particularly useful when a given phenomenon has a limit for lower values but no such limit for upper values. For example, price cannot be below zero.

3) In the construction of index numbers, geometric mean is considered to be the best average. It is especially used in developing Fisher's Ideal Formula that satisfies time reversal and factor reversal tests. (The study of these concepts is beyond the scope of this course.)

4) When large weights are desired to be assigned to small items and small weights to be assigned to large items, it is a more suitable average than arithmetic mean.

Limitations

1) Even if the single item of the given series is zero, geometric mean will be zero. Hence, it cannot be computed. For example, geometric mean of the three items 0, 10, 100 will be $\sqrt[3]{0 \times 10 \times 100} = 0$.

2) If any of the items is negative, geometric mean does not exist.

3) The computational procedure is difficult especially when the items are very large.

4) Its bias for lower values obstructs its use in the situations where disparities are to be highlighted as in the case of income distribution.

Check Your Progress A

1) Money invested in NSC VI issue becomes double in 6 years. What is the percentage rate of growth per year?

.....

.....

.....

.....

2) Marks secured by 70 students in a test (maximum marks 75) are given below. Compute geometric mean and compare it with arithmetic mean.

Marks	:	5 - 15	15 - 25	25 - 35	35 - 45	45 - 55	55 - 65
No. of Students	:	12	15	25	10	6	2

.....

.....

.....

.....

.....

.....

.....

3) The price of a commodity increased by 5% from 1978 to 1979. 8 % from 1979 to 1980 and 77% from 1980 to 1981. The average increase from 1978 to 1981 is quoted as 26% and not 30%. Verify this statement.

.....

.....

.....

.....

4) A machine is assumed to depreciate 40% in value in the first year, 25% in the second year and 10% per annum for the next three years. Each percentage is calculated on the diminishing value. What is the average percentage depreciation for the five years?

.....

.....

.....

.....

5) State whether the following statements are True or False.

i) Geometric mean is the 'n'th root of the sum of 'n' items.

ii) Geometric mean is the suitable method of averaging ratios and percentages.

iii) It is possible to calculate geometric mean for all types of data.

iv) Value of weighted geometric mean will always be greater than simple geometric mean.

v) Geometric mean is not a good method for averaging when large items are to be given importance.

6) Fill in the blanks:

i) The geometric mean of the individual values 1,2,4,8,16,32,64,128 and 256 is.....

ii) If there are 10 items in the series, its geometric mean is the..... th root of the product of these items.

iii) In computing the average ratio of increase in GNP, ... measure is more suitable.

iv) Geometric mean has bias towards ..... values.

v) If the single item of the given series is zero, geometric mean will be.....

vi) If the price of a commodity doubles in a period of 4 years, the average percentage increase per year is .....

vii) The annual rates of growth of output of a factory in 5 years are 5, 7.5, 2.5, 5 and 10 per cent respectively. Then the compound growth rate per annum for the period is.....

13.3 HARMONIC MEAN

As you know, generally, the data is in varied forms. The manner in which the data is given counts heavily for judging the appropriateness of the use of the measures of central

tendency. For example, when the total distance is constant and the speed per unit time is given harmonic mean is a more appropriate measure to find out the average speed. Suppose the data is given in terms of articles produced per hour and we are interested in knowing the average time per unit, then harmonic mean is preferable.

13.3.1 Computation

The method of computing harmonic mean is different for ungrouped data and grouped data. Now let us study these methods separately.

Ungrouped Data

If there are 'n' values of variate x viz., $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ their harmonic mean (HM) is calculated as follows:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots\ldots+\frac{1}{x_n}} = \frac{n}{\Sigma\frac{1}{x}}$$

$$\frac{1}{\text{Harmonic Mean}} = \frac{\Sigma\frac{1}{x}}{n}$$

Rewriting it

$$\text{H.M.} = \text{Reciprocal of } \left[\frac{\Sigma\frac{1}{x}}{n}\right]$$

= Reciprocal of (A.M. of reciprocals of n values $x_1, x_2, \ldots, x_n$)

Therefore, **harmonic mean is the reciprocal of the arithmetic mean of reciprocals**. For example, harmonic mean of two values 12 and 15 can be computed as follows:

$$\text{H.M.} = \frac{2}{\frac{1}{12}+\frac{1}{15}}$$

$$= \frac{2}{\frac{5+4}{60}}$$

$$= \frac{120}{9}$$

$$= 13.34$$

Illustration 5

A motorist travelled for three days at the rate of 480 kms. a day. On the first day he travelled for 10 hours at a speed of 48 kms per hour, on the second day he travelled for 12 hours at a speed of 40 kms. per hour and on the third day for 15 hours at a speed of 32 kms. per hour. What was his average speed?

Solution

Here the total distance travelled per day is constant, and time and speed are variable. We are required to compute the average speed. Therefore, harmonic mean is the appropriate average.

$$\text{H.M.} = \frac{3}{\frac{1}{48}+\frac{1}{40}+\frac{1}{32}}$$

$$= \frac{3}{\frac{37}{480}}$$

$$= \frac{3\times480}{37}$$

= 39 kms per hour (approximately).

Here, how does the harmonic mean become the appropriate average? It can be verified easily as below:

The total distance travelled in 3 days = 480 + 480 + 480 = 3 × 480 kms.

The total time taken = 10 + 12 + 15 = 37

$\therefore$ The average speed $= \frac{3\times480}{37} = 39$ kms per hour approximately.

Now you should note that the result obtained by this logical method is equal to the harmonic mean. Hence, in averaging speeds, when total distance is constant and time is variable, harmonic mean is the appropriate average.

Grouped Data

As you know, there are two types of grouped data: 1) Discrete series, and 2) Continuous series. Now let us study the methods of computing harmonic mean for these two types of data sets.

Discrete Series : For a discrete series, harmonic mean is calculated as follows:

$$\text{H.M.} = \frac{n}{\Sigma f (\text{reciprocals of } x)}$$

$$= \frac{n}{\Sigma f \frac{1}{x}}$$

$$= \text{Reciprocal } \frac{\Sigma f \frac{1}{x}}{n}$$

where symbols have their usual meaning.

The procedure to be followed here is, as follows:

1) Take the reciprocal of various values of variate x.
2) Multiply the reciprocals by the respective frequencies and obtain the total product i.e. $\left(\Sigma f \frac{1}{x}\right)$
3) Take a ratio of the total frequency (n) to $\Sigma f \frac{1}{x}$

Illustration 6

A person buys 10 kgs of commodity A at the rate of 2 kg per rupee, 20 kg of commodity B at the rate of 5 kg per rupee and 30 kg of commodity C at the rate of 10 kg per rupee. Find the average price in kgs per rupee.

Solution

We have to find out the average price. So let us denote the items to be averaged out as 'x'. The quantities bought are similar to frequencies. So denote them by 'f'. Now harmonic mean would be calculated as below:

Commodity	Price in kg per Rupee (x)	Quantity bought (f)	$\frac{1}{x}$	$f\frac{1}{x}$
A	2	10	0.5	5.0
B	5	20	0.2	4.0
C	10	30	0.1	3.0
Total		$n = \Sigma f = 60$		$\Sigma f \frac{1}{x} = 12.0$

$$\text{H.M.} = \frac{n}{\Sigma f \frac{1}{x}}$$

$$= \frac{60}{12.0}$$

$$= 5.0$$

Therefore, the average price is 5 kgs per rupee.

Note: You may ask why harmonic mean has been calculated in this illustration. To find out average price you need total money spent and the total quantity (kgs) bought. Then the average price in kgs per rupee will be the total quantity bought divided by total money spent. Column 1/X gives price of one kg in rupees and column f gives the quantity bought. So column $f\frac{1}{x}$ gives total money spent in buying quantity 'f' of different commodities. Now Σf or n gives the total kg bought by spending total money $\Sigma f \frac{1}{x}$. Hence the required average is $\frac{n}{\Sigma f \frac{1}{x}}$ which is same as harmonic mean.

From this illustration also you should note that while averaging prices expressed in quantit units, the correct average is the harmonic mean. In general, we can say that while finding t combined effect of the items to be averaged, if their reciprocals are used, harmonic mean is the right method of averaging.

Continuous Series: The computational procedure for continuous series is the same as prescribed for discrete series. The only difference is that in the case of continuous series we take the reciprocals of the mid-values (m) of different classes. Then multiply them with the respective class frequencies and obtain the total of that product i.e. (Σ f/m). Then take the ratio of total frequency (n) to the total product obtained.

$$\therefore \text{H.M.} = \frac{n}{\Sigma f/m} \text{ or Reciprocal } \frac{\Sigma f \frac{1}{m}}{n}$$

Illustration 7

Calculate harmonic mean for the following information:

Class Interval	f
0 - 10	5
10 - 20	8
20 - 30	10
30 - 40	2
40 - 50	7
50 - 60	6
60 - 70	3

Solution

Class Intervals	f	Mid-values (m)	1/m	f/m
0-10	5	05	0.2	1.0
10-20	8	15	0.067	0.536
20-30	10	25	0.04	0.40
30-40	2	35	0.029	0.348
40-50	7	45	0.022	0.154
50-60	6	55	0.018	0.108
60-70	3	65	0.015	0.045
	51			Σ f/m = 2.591

$$\text{H.M.} = \frac{n}{\Sigma f/m}$$

$$= \frac{51}{2.591}$$

$$= 19.68$$

13.3.2 Weighted Harmonic Mean

There are situations where we need to calculate weighted harmonic mean rather than simple harmonic mean. For example, a person walks first 10 kms at a speed of 4 kms an hour, next 5 kms at 3 kms an hour, and then 4 kms at 2 kms an hour. His average speed is to be found out. The kilometres walked by him at three phases would be considered as weights. The formula to be used here is:

$$\text{Weighted H.M.} = \frac{\Sigma W}{\Sigma \frac{W}{x}} \text{ where 'W' refers to weights}$$

$$\text{Weighted H.M.} = \text{Reciprocal} = \frac{\Sigma W \frac{1}{x}}{\Sigma W}$$

In the above example

x :	4	3	2
w :	10	5	4

$$\text{weighted H.M.} = \frac{10 + 5 + 4}{\frac{10}{4} + \frac{5}{3} + \frac{4}{2}}$$

$$= \frac{19}{2.5 + 1.67 + 2}$$

$$= \frac{19}{6.17}$$

$$= 3.08 \text{ kms per hour}$$

In this illustration, the weighted harmonic mean is the appropriate method. It can be verified by calculating the average speed by ordinary arithmetic method.

Case	Distance	Speed	Time taken	Hours
First	10 kms	4 km. p.h.	10/4	2.50
Second	5 kms	3 km. p.h.	5/3	1.67
Third	4 kms	2 km. p.h.	4/2	2.00
Total	19 Kms			6.17

Average speed = 19/6.17 = 3. 08 kms per hour. The two results are exactly the same. So when harmonic mean is to be calculated for items which differ in relative importance also, weighted harmonic mean should be calculated.

Illustration 8

Mr. Rakesh started for a village at a distance of six kms. He travelled in his car at a speed of 40 kms per hour. After travelling for 4 kms the car stopped running. He then travelled in a rickshaw at a speed of 10 kms per hour. After travelling a distance of 1.5 kms, he left the rickshaw and covered the remaining distance on foot at a speed of 4 kms per hour. Find his average speed per hour and verify the result.

Solution

Here speeds are $x_1 = 40$, $x_2 = 10$, $x_3 = 4$ and the weights are the distance travelled i.e., $w_1 = 4$, $w_2 = 1.5$, $w_3 = 0.5$

$$\text{Weighted H.M.} = \frac{\Sigma W}{\Sigma W/x}$$

$$= \frac{4 + 1.5 + 0.5}{\frac{1}{40} \times 4 + \frac{1}{10} \times 1.5 + \frac{1}{4} \times 0.5}$$

$$= \frac{6}{0.1 + 0.15 + 0.125}$$

$$= \frac{6}{0.375}$$

$$= 16$$

Therefore, the average speed of Rakesh is 16 kms per hour. Let us verify the answer by calculating the time taken.

Mode of Conveyance	Distance	Speed	Time Taken
Car	4 kms	40 km.p.h.	6 minutes
Rickshaw	1.5 kms	10 km.p.h.	9 minutes
On Foot	0.5 kms	4 km.p.h.	7.5 minutes
Total	6 kms		22.5 minutes

In 22.5 minutes he covered 6 kms. Therefore, in 60 minutes he would cover 16 kms (i.e. $6 \times 60 / 22.5$).

13.3.3 Properties of Harmonic Mean

1) If each value of the variate is replaced by harmonic mean, the total of reciprocals of values of the variate remains the same.
2) Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the individual observations.
3) Like arithmetic mean and geometric mean, it lends itself to further algebraic treatment.
4) Amongst the three means (viz., arithmetic mean, harmonic mean and geometric mean), harmonic mean is the least i.e., AM ≥ GM ≥ HM.

To illustrate this, let us calculate the harmonic mean of five items 2, 3, 5, 10 and 100, and compare it with the arithmetic mean and geometric mean.

$$HM = \frac{5}{\frac{1}{2}+\frac{1}{3}+\frac{1}{5}+\frac{1}{10}+\frac{1}{100}}$$

$$= \frac{5}{0.50+0.33+0.20+0.10+0.01}$$

$$= \frac{5}{1.14} = 4.39$$

As computed earlier under properties of G.M., the arithmetic mean is 24 and geometric mean is 7.86. This illustrates that for a set of positive items AM > GM > HM. This property may also be stated as that harmonic mean has bias towards small items.

Note: When all the given items have exactly the same value, then only AM = GM = HM. In such a case, median and mode will also be equal to this common value.

13.3.4 Uses and Limitations

Uses

1) For the rates and ratios involving speed, time and distance, harmonic mean is used to find out the average speed.

2) For the rates and ratios involving price and quantity (both amount of money spent and the units per rupee are given), harmonic mean is used. In general, if reciprocals of items are used in obtaining their combined effect, harmonic mean is to be used for averaging them.

3) In a given data set if there are a few large values, the reciprocals will tone down the effect of large numbers. In such cases harmonic mean is to be used.

4) When it is desired to assign greater weight to smaller values and smaller weight to larger values of a variate, its use is recommended.

Limitations

1) It is difficult to compute and understand.

2) It cannot be computed when one or more items are zeros. In fact in such a case HM will be always zero whatever may be the value of other items. For example, harmonic mean of 0, 10 and 100 will be ; $\frac{3}{\frac{1}{0}+\frac{1}{10}+\frac{1}{100}} = \frac{3}{\infty+0.10+0.01} = \frac{3}{\infty} = 0$

Note : The sign ∞ means 'infinity'. It is the concept of the greatest number.

3) To assign the largest weight to the smallest item, it is not always a desirable feature and has a limited scope in the analysis of economic data.

13.4 HARMONIC MEAN VERSUS ARITHMETIC MEAN

In order to derive averages of the rates and ratios (that involve speed, time and distance or price, quantity and amount of money spent, etc.) making a choice between the harmonic mean and arithmetic mean is not very easy. In some situations harmonic mean seems to be more proper, whereas in other situations harmonic mean is found more suitable to derive the correct answer. Such a choice mainly depends on the nature of the data. Based on it, some general guidelines for a judicious choice can be prescribed.

1) For the rates and ratios involving speed, time and distance, if the distance is given, harmonic mean is preferred. But if the time is given, arithmetic mean will be more suitable. In general, if the given ratios are in the form of x units per y, use harmonic mean when X's are given, and use arithmetic mean when Y's are given. Let us understand it more clearly through an illustration.

Illustration 9

A person travels 100 kms distance by car at an average speed of 30 kms per hour. Then he makes return trip at an average speed of 20 kms per hour. What is his average speed?

Solution

Here the speed is given in kms per hour and the total distance travelled is also known (i.e., 100 kms each side). Therefore, weighted harmonic mean with equal weight 100 each or simple harmonic mean is a more suitable average.

$$\text{H.M.} = \frac{2}{\frac{1}{20}+\frac{1}{30}}$$

$$= \frac{2}{\frac{3+2}{60}} = \frac{2\times 60}{5}$$

= 24 kms per hour

Now let us slightly change the above information. Suppose for the same trip the person travels at 30 kms per hour for half of the time and at 20 kms per hour for the other half of the time. Since the times of the trip are given, arithmetic mean will be chosen as an average. Further as two time periods are equal, simple arithmetic mean is suitable.

$$\text{Arithmetic Mean} = \frac{30+20}{2} = 25 \text{ km. p. h.}$$

You can verify here whether the arithmetic mean is the correct average or not. With the arithmetic mean speed of 25 kms per hour, he can cover 200 kms in 8 hours. If he travels for half of the time i.e., 4 hours with a speed of 30 kms per hour and 4 hours with a speed of 20 kms per hour he would cover exactly 200 kms. Hence, in this case the correct average speed is arithmetic mean.

2) The second distinguishing point is that the arithmetic mean is affected by the extreme items, whereas harmonic mean is more sensitive to low values. Therefore, for an uneven distribution use of arithmetic mean is not suggested, whereas for the analysis of economic data, use of harmonic mean is not used.

Check Your Progress B

1) What is harmonic mean?

.....

.....

.....

.....

2) Monthly expenditure of a group of students is given below. Compute the harmonic mean. 125, 75, 10, 130, 45, 500, 150, 80, 65, 100.

.....

.....

.....

.....

3) Compute the harmonic mean from the following data:

Size of Items	0-10	10-20	20-30	30-40	40-50
Frequency	5	10	7	3	2

.....

.....

.....

.....

.....

.....

4) An investor buys Rs. 1,200 worth of shares in a company each month. During the first five months, he bought shares at a price of Rs. 10, Rs. 12, Rs. 15, Rs. 20 and Rs. 24 per share. What is the average price paid per share?

.....

.....

.....

.....

.....

5) A person, to reach his native place, covers first 1200 kms by train at an average speed of 80 kms per hour. Then 20 kms by bus at a speed of 40 kms per hour and finally 5 kms by cycle rickshaw at an average speed of 8 kms per hour. What is the average speed for the total journey?

.....

.....

.....

.....

.....

.....

6) State whether the following statements are True or False.

i) Harmonic mean is the reciprocal of arithmetic mean.

ii) Harmonic mean is the best measure to average rates and speeds in all situations.

iii) Only mathematical measures of average can be calculated by using weights.

iv) If an item in the data has a value of 0, harmonic mean does not exist.

v) To find the average for the price in rupees per item, when the number of items purchased is given, harmonic mean is the right average.

7) Fill in the blanks with the appropriate word given in the bracket:

i) ..... is the reciprocal of the mean of reciprocals. (Mean/Harmonic mean)

ii) The harmonic mean is such that if each value of the variate is replaced by it, the total of ..... of the value of the variate remains the same. (reciprocal/product)

iii) Amongst the three averages (AM, GM, and HM), harmonic mean is the ..... (largest/least)

iv) When it is desired to assign ..... weights to small values, harmonic mean is recommended. (smaller/greater)

v) If the given ratios are in othe form of x units per y, use..... when x's are given. (Harmonic mean/arithmetic mean)

vi) If the speed of aeroplane and the distances travelled are given, to find the average speed ..... harmonic mean is used. (simple/weighted).

13.5 MOVING AVERAGE

13.5.1 What is Moving Average?

While considering matters such as trend of prices, sales, profits, etc., a particular type of average known as moving average is used. It is a measure of trend (long-term tendency of the data) in the time series data. Moving average is an arithmetic average of data arising over a period of time and is calculated by replacing the first item in the average by the newly arising item. Each moving average is based on values covering a fixed time span which is called "**period of moving averages**".

The successive averaging process does a smoothing operation in the time series data, i.e., it irons out fluctuations of uniform period and intensity. They can be completely eliminated by choosing the period of moving average that coincides with the period of the cycle i.e.,

periodic movements. Even if the periodic movement is absent in the time series, the irregularities of data can be reduced to a large extent by moving average process.

13.5.2 Computation

In the computation of moving average, the period of moving average is a very important factor. For example, for yearly values A, B, C, D, E, and F, the three yearly moving averages can be computed as shown in Table 13.1.

Table 13.1

Computation of Moving Averages

Yearly Values	3 Yearly Moving Totals	3 Yearly Moving Averages
A	.....	.....
B	(A + B + C)	(A + B + C)/3
C	(B + C + D)	(B + C + D)/3
D	(C + D + E)	(C + D + E)/3
E	(D + E + F)	(D + E + F)/3
F	.....	.....

We can have either an odd period of moving average (e.g. 3 years, 5 years, 7 years) or an even period of moving average (i.e. 2 years, 4 years, 6 years). The period of moving average is generally determined in the light of the length of the cycle in the data. Ordinarily the moving average period ranges between 3 to 10 years for business series.

Odd Period of Moving Average

When period of moving average is odd (say 3 years, 5 years, 7 years, etc.) the moving average is associated with mid-points of relevant time interval. Study Table 13.2 carefully to understand the procedure.

Table 13.2

Computation of Odd Period Moving Average

Years	Sales ('000 tonnes)	3Yearly Moving Totals	3 Yearly Moving Averages
1977	15	-	-
1978	25	72	24
1979	32	81	27
1980	24	75	25
1981	19	60	20
1982	17	-	-

You should note that the moving average for the first three years (1977, 1978, and 1979) i.e., 72 is associated with the middle year 1978. Having dropped the first year, the moving average of the next three years i.e., 1978, 1979 and 1980 is placed against 1979; and so on. You must also note that moving average for the first year and the last year in the given data cannot be obtained. If the period of moving average is 5 years, moving average for the first two years and last two years cannot be obtained.

Even Period of Moving Average

If the period of moving average is even (say 4 years, 6 years, 8 years, etc.) the moving totals and moving averages would not coincide with the original time period. It would not be possible to place moving average exactly against some year. Therefore, you have to resort to 'centering'. Centering is done in a manner that helps coincide the moving average with the original data. Study Illustration 10 carefully and understand the procedure involved in centering.

Illustration 10

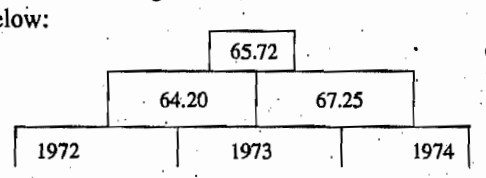
Compute 4 yearly moving averages for the following data:

Years	: 1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Sales (Rs. in '000)	: 75	60	54	69	86	65	63	80	90	72

Solution

Years	Sales (Rs. in .000s)	4 Yearly Moving Total	4 yearly Moving Average	4 yearly Movin Average Centered
1971	75	—	—	—
1972	60	—	—	—
		258	64.20	
1973	54			65.72
		269	67.25	
1974	69			67.88
		274	68.50	
1975	86			69.62
		283	70.75	
1976	65			72.12
		294	73.50	
1977	63			73.50
		298	74.50	
1978	80			75.37
		305	76.25	
1979	90	—	—	—
1980	72	—		

The total 258 of the first four figures (years 1971 to 1974) and their average 64.20 is written against the middle of this time period i.e., middle of the years 1972 and 1973. This middle time period is a specially designed year taking last six months from 1972 and the first six months from 1973. Similarly, the total 269 corresponding to years 1972 to 1975 and their average 67.25 is written against the specially designed year i.e., the mid-year of 1973 and 1974. This process continues till the last average 76.25 and the total 305 is noted against the mid-year of 1978 and 1979. To find out the first centred moving average 65.72 (i.e., a figure of moving average which will coincide with the year 1973), we have to find the mid- value of 64.20 and 67.25, the first two figures in Column 4. This can be easily seen with the help of diagram given below:



The diagram shows that the figure which coincides with the year 1973 will come from half of 64.20 and half of 67.25, which means that it is the mean of the two moving averages. This mean value 65.72 is, therefore, called **centred moving average** and is entered in the last column. The various centred moving averages are, thus, calculated by taking successively mean of the two consecutive figures from Column 4.

13.6 CHOICE OF A SUITABLE AVERAGE

Starting from Unit 10, we have discussed various averages viz., mean, mode, median, geometric mean, etc. You have studied merits, demerits, and specific uses of each of these averages separately. Now we should know how to make choice of a suitable average for a given purpose. Examining from the point of view of essential qualities of a good measure of central tendency, arithmetic mean appears to be the best measure as it possesses the largest number of these qualities. Given the situation, however, the choice of a suitable average poses a problem. If the choice is not proper, the conclusions will not be much dependable. With an improper choice of an average, the comparative scene that emerges will be far from reality. Therefore, while making the choice of an average, you should keep in mind the following aspects.

1) **The Purpose:** The choice is to be made in accordance with the purpose that an average is designed to serve. If the purpose is to give all the items of the series an equal importance, arithmetic mean will be a proper average. If the purpose is to find the most common or most fashionable item, the mode will be a suitable average. If the purpose is to locate a position of an item in relation to other items, it would be the median that serves the purpose. When small items are to be given a little more importance than big items, the choice falls on geometric mean. If sufficiently greater weights are to be assigned to smaller values, harmonic mean should be used.

2) **Nature and the Form of the Data Set:** If the distributions are skewed, mode or mean will be preferred. For an open-ended distribution, again mode or median would be more suitable. In case of j-shaped or reverse j-shaped distribution i.e., which highly deviate from symmetry, the median is the most important average. Price distribution and income distribution are two examples of it. If the data is evenly spread out and does not display wide variations, the arithmetic mean will be an appropriate average. Average cost of production is an example of it. When the ratios or percentages are to be averaged, geometric mean is the most appropriate measure. The data set in which the value of a variable is compared with another variable which is constant, harmonic mean is the most suitable average. Examples are — varying speed with constant distance, varying quantities bought per rupee, etc.

3) **Amenability to further Algebraic Treatment:** If an average is to be used for further algebraic treatment, arithmetic mean is considered to be the best as it is very widely used.

4) **Qualitative Phenomena :** For the characteristics which are qualitative in nature such as honesty, beauty, intelligence, etc., median seems to be a proper average.

5) **Special Purposes:** For calculating trend in time-series analysis, the moving average would be the most suitable average.

Though the above considerations act as a guiding principle in making a choice of a suitable average, in many cases it is arbitrary. If the higher value is required to prove the hypothesis, it is tempting to use the measure which gives the higher value. Since we can select the measures of central tendency to suit our fancy, there is a possibility of selecting the average which produces the result we want. When used unscrupulously or incompetently, the user is at fault not the tool.

Check Your Progress C

1) Fill in the blanks with the appropriate word given in the bracket.

   i) .... is a measure of trend and cyclic variations. (Average/Moving average)

   ii) The fluctuations in the time series can be completely eliminated by choosing proper ..... of moving average. (period/length)

   iii) For even period of moving average ..... is resorted to. (averaging/centering)

2) Suggest suitable averages for the following cases:

   i) The sizes of garments sold out in a garment shop.

   ii) The distribution with open-ended classes.

   iii) Average rate of growth required from the figures of percentage changes from year to year.

   iv) The distance is fixed but the speeds are varying.

   v) When it is desired that the absolute sum of the deviations from that average is minimum.

   vi) When it is desired to iron out irregular fluctuations in the time-series data.

   vii) If the purpose is to give small items more importance than the big items.

   viii) If the distribution has extreme items.

   ix) The cost of living index is to be constructed.

   x) In making the preference of furniture style from the available styles.

13.7 LET US SUM UP

There are a few other measures of central tendency such as geometric mean and harmonic mean which are used in specific situations. For averaging ratios or percentages, geometric mean is used. If there are "N" items in the series, its geometric mean is the Nth root of the product of these items. When the number of items is more, logarithm of geometric mean is taken to facilitate computation, Geometric mean is computed for both ungrouped and grouped data, and also for discrete and continuous series by using different formulas.

Geometric mean is very widely used for computing average rate of change in the variable during a particular time span. Weighted geometric mean also can be calculated which is used in the construction of index numbers. Geometric mean has some mathematical properties that enhance its use in averaging ratios and percentages. It also suffers from certain limitations. It does not work if the value of an item is zero or negative.

The data sets in which the value of a variable is compared with another variable which is constant, harmonic mean is used. For example, harmonic mean is used for averaging rates and ratios involving speed, time and distance. It is the reciprocal of the arithmetic mean of reciprocals of the individual observations. It can be computed for ungrouped and grouped data and also for the discrete and continuous series. Like weighted geometric mean weighted harmonic mean also can be calculated. There are situations in which it is difficult to make a choice between harmonic mean and arithmetic mean as an average. If the given ratios are in the form of x units per y, use harmonic mean when x's are given and use arithmetic mean when y's are given.

While considering the trend of prices, sales, profits, etc., moving average is used. It is a series of overlapping simple averages that does a smoothing operation in the time-series data. Computation of moving average depends on the period of moving average which can be odd or even. For an odd period, the moving average is associated with mid-points of relevant time interval. It is not possible for an even period and hence centering is done.

The choice of a suitable average depends on the purpose that an average designed to serve such as the nature and the form of the data set, its amenability to further algebraic analysis, etc. The use of measures of central tendency, however, is to be made cautiously and competently.

13.8 KEY WORDS

Geometric Mean : If there are N items in the series, the geometric mean is the Nth root of their product.

Harmonic Mean: The reciprocal of the arithmetic mean of reciprocals of the individual observations.

Moving Average: A series of overlapping averages each being an ordinary arithmetic mean of data arising over a period, and written against the middle year. This is designed to approximate the trend of a series.

Trend: The general long-term tendency of the time-series data.

13.9 ANSWERS TO CHECK YOUR PROGRESS

A) 4) 121.3

1) 12.3 % approximately

2) GM = 25.3 marks, AM = 28.4 marks

4) 29%

5) i) False, ii) True, iii) False, iv) False, v) True

6) i) 16, ii) 10, iii) GM, iv) middle, v) zero, vi) 18.7, vii) 5.9

B) 2) HM = 50.55

3) 13

4) Rs. 14.63

5) 75.45 km.p.h.

6) i) False, ii) False iii) True, iv) True, v) False

7) i) Harmonic mean, ii) reciprocal, iii) least, iv) greater, v) harmonic mean, vi) weighted

C) 1) i) Moving average, ii) period, iii) centering

2) i) Mode, ii) Median, iii) Geometric Mean, iv) Harmonic Mean, v) Median vi) Moving average, vii) Harmonic Mean, viii) Median, ix) Weighted average x) Mode

13.10 TERMINAL QUESTIONS/EXERCISES

Questions

1) Compare arithmetic mean, geometric mean and harmonic mean to point out their relative merits and limitations.

2) Explain the method of finding a moving average. What purpose does it serve?

3) How do you make a choice of suitable measure of central tendency?

Exercises

1) If the population has doubled itself in twenty years, is it correct to say that the rate of growth has been 5% per annum? If not, what is the true rate of growth?

   (**Answer:** No. 1.035%)

2) The annual growth rate of production of a factory in 5 years is 5.0, 7.5, 5.0, 2.5 and 10 per cent, respectively. What is the compound rate of growth of production per annum for the period?

   (**Answer:** 5.9% per annum)

3) Geometric mean of 8 items is 3 and geometric mean of 12 items is 11. What will be the geometric mean for all 20 items?

   (**Answer:** 6.54)

4) Find the Harmonic mean for the following data;
   i) 1, 2, 3, 4, 5, 6, 7, 8, 9
   ii) 1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9

   (**Answer:** i) 3.184; ii) 4.505)

5) You take a trip which entails travelling 900 miles by train at an average speed of 60 km. p.h.; 3,000 miles by boat at an average speed of 25 km.p.h.; 4,000 km. by plane at 350 km. p.h. ; and finally 15 miles by taxi at 25 km. p.h. What is the average speed for the entire distance?

   (**Answer:** 31.6 km. p.h.)

6) Calculate the (i) 3 yearly, and (ii) 4 yearly moving averages for the following data on prices.

Years	: 1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Price (Rs.):	6	8	11	7	10	11	15	11	12	13	17

7) Give a specific example of your own for each of the following cases:
   i) The median is preferred to the arithmetic mean.
   ii) The harmonic mean is preferred to the arithmetic mean.
   iii) The mode will be preferred to the median.
   iv) The arithmetic mean is preferred to the harmonic mean.
   v) No average would be meaningful.

Note: These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.

SOME USEFUL BOOKS

Elhance, D.N. and veena Elhance, 1988. *Fundamentals of Statistics*, Kitab Mahal : Allahabad, (Chapter 8)

Gupta, C.B. 1988. *An Introduction to Statistical Methods*, Vikas Publishing House: New Delhi. (Chapter 9)

Gupta, S.P. 1989, *Elementary Statistical Methods*, Sultan Chand & Sons: New Delhi. (Chapter 7)

Sancheti, D.C., and Kapoor, V.K., 1989, *Statistics Theory, Methods and Applications*, Sultan Chand & Sons: New Delhi. (Chapter 5)

Shenoy, G.V. Srivastava V.K. and S.C. Sharma, 1988. *Business Statistics* Wiley Eastern: New Delhi, (Chapter 4)

Simpson, G, and Kafka, F. *Basic Statistics*, Oxford & IBH Publishing: New Delhi. (Chapters 10-12)

ECO-07 ELEMENTS OF STATISTICS
Course Components

BLOCK	UNIT NO.	PRINT MATERIAL
1		**Basic Statistical Concepts**
	1	Meaning and Scope of Statistics
	2	Organising a Statistical Survey
	3	Accuracy, Approximation and Erros
	4	Ratios, Percentages and Rates
2		**Collection, Classification and Presentation of D**
	5	Collection of Data
	6	Classification of Data
	7	Tabular Presentation
	8	Diagrammatic Presentation
	9	Graphic Presentation
3		**Measures of Central Tendency**
	10	Concept of Central Tendency and Mean
	11	Median
	12	Mode
	13	Geometric, Harmonic and Moving Averages
4		**Measures of Dispersion and Skewness**
	14	Measures of Dispersion-I
	15	Measures of Dispersion-II
	16	Measures of Skewness

UNIT 14 MEASURES OF DISPERSION - I

Structure

14.0 OBJECTIVES

After studying this unit, you should be able to :

- explain the concept of dispersion and significance of measuring it,
- differentiate between absolute and relative measures of variation,
- compute several measures of dispersion such as the range, quartile deviation and mean deviation for different types of data, and
- decide the use of appropriate measures under different situations.

14.1 INTRODUCTION

In Units 10 to.13 you have studied about different measures of central tendency. But central tendency alone is not sufficient to analyse the data. For more meaningful analysis of the data. it is necessary to study dispersion i.e., the spread of the data or the extent ti, which items deviate from central tendency . In this unit, you will study the meaning and significance of dispersion. You will also learn in detail about the three measures of dispersions viz., range, quartile deviation and'mean deviation.

14.2 WHAT IS DISPERSION?

The word, dispersion is used to denote the degree of heterogeneity in the data. It is an important characteristic indicating the extent to which observations vary among themselves. A measure of dispersion describes the spread or scattering of individual values around the central value. To understand the concept of dispersion clearly. study Illustration 1 carefully.

Illustration 1

Daily Sales of Three Different Firms *(in Rs.)*

Firm A	Firm B	Firm C
60,000	62,500	51,000
60,000	60,000	32,000
60,000	52,250	22,1100
60,000	56,500	18,000
60,000	60,500	27,000
60,000	68,250	2,10,000
$\bar{X}_A = 60,000$	$\bar{X}_B = 60,000$	$\bar{X}_C = 60,000$

Since the average sales of firms A, B, and C are the same, we are likely to conclude that the three distributions of the sales are similar. But you should note that the variations in the sales are different from firm to firm. Daily sales are the same for all the days in the case of Firm A whereas there is some variation in the daily sales oi Firm B and greater amount of variation for Firm C. Here, although these three data sets have the same mean, they differ in terms of scatter of items. Therefore, different sets of data may have the same measure of central tendency, but may differ greatly in terms of spread or scatter of the items i.e. dispersion.

The word dispersion can be interpreted in another sense also. When all items of the data are not equal to central tendency, then the various items differ from central tendency by a certain amount. Dispersion gives, on an average, by how much amount items differ from central tendency. You may note that in the case of Firm B, deviations of individual sales from the mean sale (i.e., 60,000) are much smaller than the deviations of Firm C. This implies that the average of the deviations from the mcan sales will be smaller for Firm B compared to Firm C. In other words. Firm B has smaller dispersion than Firm C.

14.3 SIGNIFICANCE OF MEASURING DISPERSION

Measures of variations (dispersion) are calculated to serve the following purposes :

1 Measuring variability determines the reliability of an avarage by pointing out to what extent the average is representative of the entire data. In Illustration I discussed earlier, mean sales Rs. 60,000 is the perfect representative of sales for different days tor Firm A. In case of Firm B, the variation is low as the mean sale is quite close to sales figures of different days. Therefore, in this case, the mean can be considered as representative of the sales for each day. But in case of Firm C the variation in individual figures is very large so the average of Rs. 60,000 is hardly a representative of all high' and low figures such as Rs. 2,10,000 and Rs. 18,000.

2 Measures of dispersion enable comparisons of two or more distributions with regard to their variability.

3 Another purpose of measuring variability is to determine the nature and cause of variation in order to control the variation itself.

4 Measuring variability facilitates the use of other statistical measures like correlation, regression, statistical inference, ctc.

14.4 PROPERTIES OF A GOOD MEASURE OF DISPERSION

As you know, a measure of dispersion is the average of the deviations of items from its mean i.e., it is an average of second order. Hence, it should also possess all the qualities of a good measure of an average. According to Yule and Kendall the qualities of a good measure of dispersion are as follows :

1 Statistical measures are used even by layman. So complicated definitions and calculations are not desirable. It should be simple to understand and easy to. calculate.

2 It should be rigidly defined. For the same data, all the methodsshould produce the same answer. Different methods of computation leading to different answers is not proper.

3 It should be based on all items. When it is based on all items, it will produce a more representative value. Thus, good measure of dispersion should be based on the entire data.

4 It should be amenable to further algebric treatment. This means combining groups, calculations of missing values, adjustment for wrong entries, etc., which are possible without the knowledge of actual values of all items. Such treatment should be possible with a good measure of dispersion also.

5 It should have sampling stability. It means that the average difference between the values obtained from the sample and the corresponding values from the population should be the least. If it is so far a measure of dispersion, it is the best measure.

6 It should not be unduly affected by the extreme items. Extreme items, many times, are not true representatives of the data. So their presence should not affect the 'calculation to a large extent.

This list is not a complete-list of the properties of a good measure of dispersion. But these are the most important characteristics which a good measure of dispersion should possess.

14.5 ABSOLUTE AND RELATIVE MEASURES OF DISPERSION

The measure of dispersion which are expressed in terms of the original units of data are termed as **Absolute Measures.** For example, in Illustration 1 discussed earlier, the daily sales of the Firm B range between Rs. 52,250 to Rs. 68,250. So the spread of the data is of the order Rs. 68,250–52,250 or Rs. 16,000. This is the absolute measure of the spread of the sales. Such measures expressed in units of data are not suitable for comparing the variability of the distributions or series expressed in different units of measurement. **Relative Measures** of dispersion, on the other hand, are obtained as ratios or percentages. Therefore, relative measures are pure numbers independent of the unit of measurement. A measure **of** relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average or the selected items of the data. Hence, it is also known as **Coefficient of Dispersion.** For example, in Illustration 1 discussed earlier, if one expresses the spread Rs. 16,000 as the ratio of average sales Rs. 60,000 i.e., 16,000160,000 it becomes a relative measure. This value is a simple number and has no specific units of measurement with it. Similarly, the spread Rs. 16.000 could also be expressed as the ratio of sum of two extreme sales i.e., $\frac{16,000}{52.250 + 68.150}$, This will also give a relative measure of the spread of the sales.

Sometimes, even when data are in the same units, the comparison of variation by absolute measure of variation is not worth comparing. A variation of one kilometer (1,00,000 cm) in measuring distance from Delhi to Bombay is hardly of any significance. But a variation of 10 cm in measuring a piece of cloth of 1.40 meters is of very great significance. So whenever comparisons of variability in two sets of data are donc, it is always donc in terms of relative measures.

Check Your Progress A

1 What is the meaning of the term Dispersion'?

.....

.....

.....

2 Differentiate bctwccn absolute measures and relative measures of dispersion.

.....

.....

.....

.....

State whether the following statements are True or False.

i) Variation means only the spread of items of a data set.

ii) The only purpose of measuring dispersion is to assess the reliability of an average,

iii) Variability itself may be some times used in decision making.

iv) Comparison of two sets of data can only be done by relative measures of dispersion.

v) All absolute measures of dispersion are pure numbers.

vi) A good measure of dispersion should preferably use all the items of the data.

14.6 MEASURES OF DISPERSION

The following measures of absolute dispersion are in common use :

1 Based on selected items of the data

i) Range — spread for entire data
ii) Inter Quartile Range — spread for middle 50% data. More commonly Quartile Deviation is used in its place, which is half of inter quartile range.

2 Based on all items of the data

i) Mean Deviation — mean of the absolute deviations from central tendency.
ii) Standard Deviation or Root Mean Square Deviation about arithmetic mean

3 A Graphic Method — Lorenz Curve.

The relative measures of dispersion corresponding to the measures of'absolute dispersion are :

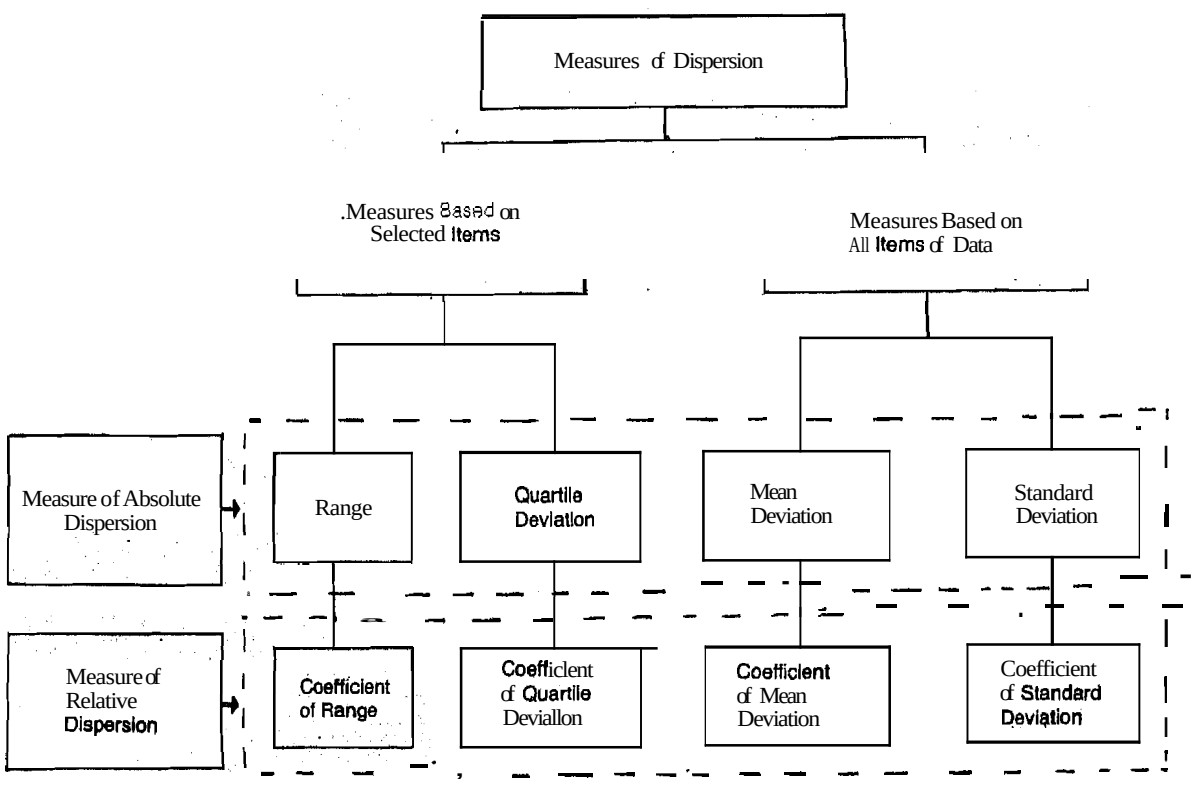
Absolute Measures of Dispersion	Relative Measures of Dispersion
i) Range	Coefficient of Range
ii) Quartile Deviation	Coefficient of Quartile Deviation
iii) Mean Deviation	Coefficient of Mean Deviation
iv) Standard Deviation	Coefficient of Standard Deviation

Coefficient of standard deviation when expressed in percentages is called coefficient of variation.

Study Figure 14.1 carefully for classification of measures of dispersion. You will study Range, Quartile Deviation and Mean Deviation later in this unit, and Standard Deviation and Lorenz Curve in Unit 15.

Figure 14.1

Classification of Measures of Dispersion



14.7 RANGE

The range is defined as the difference between the highest (numerically largest) value and the lowest (numerically smallest) value in a set of data.

Thus, Range $= X_{max} - X_{min}$

Where, X_{max} = highest value,' X_{min} = lowest value.

From Illustration 1 discussed earlier, consider the daily sales data for the three firms and compute the range.

For Firm **A,** Range = 60,000 − 60,000 = 0

For Firm B, Range = 68,250 − 52,250 = 16,000

For Firm C, Range = 2,10,000 − 18,000 = 1,92,000

The interpretation of the value of range is very simple. In this illustration, the variation is zero in case of daily sales for Firm **A,** the variation is small in case of Firm B, and the variation is very large in case of Firm C.

For grouped data, the range may be approximated as the difference between the upper limit of the largest class and the lower limit of the smallest class. The relative measure corresponding to range, called the coefficient of range, is obtained by expressing range as the ratio of sum of two extreme items. In this case ratio is not expressed in terms of average, as the range does not depend on average. It relates only to two selected items of the data. So the coefficient of range is defined as :

$$\text{Coefficient of Range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$$

Study Illustration 2 carefully and understand the procedure involved in the computation of the coefficient of range.

Illustration 2

Calculate the coefficient of range from the following data :

Sales (Rs. in Lakhs)	No. of Dnys
30 – 40	12
40 – 50	18
50 – 60	20
60 – 70	19
70 – 80	13
80 – 90	8

Solution

Range $= X_{max} - X_{min}$

$= 90 - 30$

$= 60.$

$$\text{Coefficient of Range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$$

$$= \frac{90 - 30}{90 + 30}$$

$$= \frac{60}{120}$$

$= 0.5.$

The range is very easy to calculate and it gives us some idea about the variability of the data. Since only two extreme values are used for computing range, it is a crude measure of variation.

The concept of range is extensively used in statistical quality control. Range is helpful in studying variations in the, rices of shares, debentures and agricultural commodities which are very sensitive to price changes. The range is a good indicator for weather forecast.

14.8 QUARTILE DEVIATION

Quartile deviation is defined as half the difference between the upper and lower qualities. You have already studied the methods of computing Quartiles in Unit 11.

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

Where Q_1 is the first quartile and Q_3 is the third quartile.

As the difference between Q, and Q_3 is the distance between the two quartiles, this may be called Inter Quartile Range and half of this, Semi-Inter Quartile Range is called Quartile Deviation.

Quartile Deviation (QD) is dependent on the two quartilcs, and docs pot take into account the variability of the largest 25% and the smallest 25% of observations. It is, therefore, unaffected by extreme values. Another advantage of quartile deviation is that it is the only measure of variability which can be used for open-end distribution. The main limitation of quartile deviation is that it does not depend on the magnitudes of all observations. It is based on the middle 50% of the observations.

The relative measure of dispersion based on quartile deviation is called coefficient of quartile deviation. The coefficient of quartile deviation is defined as :

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

This is because quartiles are also two selected items of the data and they have nothing to do with the average of the data. Study the following Illustrations carefully, you will understand the procedure involved in the calculation of Quartile Deviation.

Illustration 3

Calculate quartile deviation and its coefficient from the following data :

Weight (in Kgs):	60	61	62	63	65	70	75	80
No. of Workers :	1	3	5	7	10	3	1	1

Solution

Computation of Quartile Deviation and its Coefficient

Weight in Kgs.	Frequency	Cumulative Frequency
60	1	1
61	3	4
62	5	9
63	7	16
65	10	26
70	3	29
75	1	30
80	1	31 = n

Q_1 = Size of $\frac{1}{4}(n + 1)$ or 8th observation

= 62 Kgs. (because 8th observation falls in this category)

Q_3 = Size of $\frac{3}{4}(n + 1)$ or 24th observation

= 65 Kgs. (because 24th observation falls in this category)

$$\therefore \text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{65 - 62}{2}$$

= 1.5 Kgs.

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{65 - 62}{65 + 62}$

$= \frac{3}{127}$

$= 0.024.$

Illustration 4

Calculate semi-interquartile range and its coefficient from the following data :

Marks	:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Students	:	11	18	25	28	30	33	22	15	22

Solution

To compute quartile deviation, we need the values of the first quartile and the third quartile which can be obtained from the following table :

Marks	Frequency (f)	Cumulative Frequency (c.f.)
0-10	11	11
10-20	18	29
20-30	25	54
30-40	28	82
40-50	30	112
50-60	33	145
60-70	22	167
70-80	15	182
80-90	22	204

Q_1 has $\frac{N}{4}$ observations i.e., $\frac{204}{4} = 51$ observations, below it. So Q_1 lies in the 20-30 class.

$$Q_1 = l + \frac{\frac{N}{4} - c}{f} \times i$$

Where l = lower limit of quartile class
c = cumulated frequency preceding the quartile class
f = simple frequency of the quartile class
i = class-interval of quartile class

$\therefore Q_1 = 20 + \frac{51 - 29}{25} \times 10$

$= 28.8.$

Q_3 has $\frac{3N}{4}$ observations i.e., $3 \times \frac{204}{4} = 153$ observations, below it. So Q_3 lies in the 60-70 class.

$$Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times i$$

$= 60 + \frac{153 - 145}{22} \times 10$

$= 63.64.$

Semi-inter Quartile Range or Quartile Deviation is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$
$$= \frac{63.64 - 28.8}{2}$$
$$= \frac{34.84}{2}$$
$$= 17.42 \text{ marks.}$$

The relative measure corresponding to quartile deviation, called the coefficient of quartile deviation, it calculated as follows :

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$
$$= \frac{63.64 - 28.8}{63.64 + 28.8}$$
$$= 0.37.$$

Illustration 5

Compute an appropriate measure of dispersion for the data given below:

Monthly Expenditure (Rs.)	No. of Families
Below 850	12
850 - 900	16
900 - 950	39
950 - 1,000	56
1,000 - 1,050	62
1,050 - 1,100	75
1,100 - 1,150	30
1,150 and above	10

Solution

Since the frequency distribution has open-end class, quartile deviation will be the most appropriate measure of dispersion.

Monthly Expenditure (Rs.)	No. of Families	Cumulative Frequency
Below 850	12	12
850 - 900	16	28
900 - 950	39	67
950 - 1,000	56	123
1,000 - 1,050	62	185
1,050 - 1,100	75	260
1,100 - 1,150	30	290
1,150 and above	10	300 = n

Q_1 has $\frac{N}{4}$ observations i.e., $\frac{300}{4} = 75$ observations, below it. So Q_1 lies in the class 950 - 1,000.

$$Q_1 = l + \frac{\frac{N}{4} - c}{f} \times i$$
$$= 950 + \frac{\frac{300}{4} - 67}{56} \times 50$$
$$= \text{Rs. } 957.14.$$

Q_3 has $\frac{3N}{4}$ observations i.e., $\frac{3 \times 300}{4} = 225$ observations, below it. So Q_3 lies in the class 1,050 - 1,100.

$$Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times i$$

$$= 1,050 + \frac{225 - 185}{75} \times 50$$

$$= \text{Rs. } 1,076.67.$$

$$\therefore \text{ Q.D. } = \frac{Q_3 - Q_1}{2}$$

$$= \frac{1,076.67 - 957.14}{2}$$

$$= \text{Rs. } 59.76.$$

Check Your Progress B

1 **Distinguish between the absolute and relative measures of dispersion.**

.....

.....

.....

.....

.....

2 **Define quartile deviation.**

.....

.....

.....

.....

3 **Distinguish between range and the coefficient of range.**

.....

.....

.....

.....

.....

4 **Compute the range and quartile deviation for the following data on the number of patients treated at the Hospital emergency room per day.**

45, 50, 36, 59, 28, 42, 55, 67, 33, 35, 40, 50

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

5 Compute range, quartile deviation and related coefficients from the following data:

Size	:	5-7	8-10	11-13	14-16	17-19
Frequency	:	14	24	38	20	4

.....

.....

.....

.....

.....

.....

6 State whether the following statements are True or False :

i) Range ignores the distribution of values between the two extremes.

ii) Range can be calculated for an open-end data.

iii) Quartile deviation is same as the inter percentile range of P_{25} P_{75}.

iv) Quartile deviation is a best measure of dispersion for open-end data.

v) Relative measure of Quartile Deviation is based on Median.

14.9 MEAN DEVIATION

As you know, one of the characteristics of an ideal measure of dispersion is that it should be based on all items. From this point of view, range and quartile deviations are not ideal as they are not based on all the observations of the data. But, the measure of mean (or average) deviation is ideal in this sense as this measure is based on all observations in the given data set. This measure is computed as the arithmetic mean of the absolute deviations of the individual observations from the average of the given data. The average which is frequently used in computing the mean deviation is mean or median, though sometimes mode can also be used. Absolute deviations means the deviations are treated as positive regardless of the actual sign. Given the observations $X_1, X_2, \ldots, X_n$, in order to find Mean Deviation about average A, we first obtain the deviations $(X_1 - A), (X_2 - A), \ldots, (X_n - A)$. Some of these deviations may be positive and some negative. If we write $|X_i - A|$ to denote the positive value of $(X_i - A)$, whatever be the actual sign, the sum of these absolute deviations is $|X_1 - A| + |X_2 - A| + \ldots\ldots\ldots + |X_n - A| = \sum |X - A|$. The arithmetic mean of these absolute deviations is the Mean Deviation.

$\therefore$ Mean Deviation (about A) $= \frac{1}{n} \sum |X - A|$

i) When A is the arithmetic mean $(\overline{X})$,

Mean Deviation (about $\overline{X}$) $= \frac{1}{n} \sum |X - \overline{X}|$

ii) When A is the median (M_e)

Mean Deviation about $M_e = \frac{1}{n} \sum |X - M_e|$

iii) When A is mode (M_o)

Mean Deviation about $M_o = \frac{1}{n} \sum |X - M_o|$

If the data given is in frequency distribution form, taking anology from calculation of arithmetic mean, mean deviation (M.D.) about $\overline{X} = \frac{1}{n} \sum f|X - \overline{X}|$. Similarly, mean deviation about $M_e = \frac{1}{n} \sum f|X - M_e|$. Mean deviation about $M_o = \frac{1}{n} \sum f|X - M_o|$.

Usually, the name of average from which deviations are taken is always mentioned with mean deviation. But in some texts, when deviations are taken from $\overline{X}$, instead of mean deviation about $\overline{X}$ only 'Mean Deviation' has been stated as its name. An **important property of Mean Deviation is that it has the minimum value when deviations are taken from median, i.e., Mean Deviation about median is the least.**

The relative measure corresponding to the mean deviation, called the **coefficient of mean deviation,** is obtained by dividing mean deviation by the particular average used in computing the mean deviation. Thus, if mean deviation has been computed from median, the coefficient of mean deviation shall be obtained by dividing the mean deviation by the median. Coefficient of M.D. about $M_e = \frac{\text{M.D. about Median}}{\text{Median}}$.

Similarly, a coefficient of M.D. about $\overline{X} = \frac{(\text{M.D. about } \overline{X})}{\overline{X}}$

Mean deviation is based on all observations and hence takes into account the variability of each of the items in the data set. However, the practice of neglecting signs and taking absolute deviations makes it difficult to be treated algebrically.

Although the average deviation is a good measure of variability, its use is limited. If one desires only to measure and compare variability among several sets of data, the average deviation may be used. The concept of mean deviation will become clear, if you study the following illustrations carefully.

Illustration 6 .

Calculate the mean deviation of the following values about the Median :

18, 25, 63, 59, 29, 72, 17, 25, 105, 87.

Solution

Since there are ten observations which in an even number, the median is the average of the two middlemost observations, when arranged in order of magnitude

17, 18, 25, 25, **29, 59,** 63, 72, 87, 105

$\text{Median} = \frac{29 + 59}{2} = 44$

Calculations for Mean Deviation

X	$\lvert X - \text{Median}\rvert$ i.e., $\lvert X - 44\rvert$
18	26
25	19
63	19
59	15
29	15
72	28
17	27
25	19
105	61
87	43
Total	$\lvert X - M_e\rvert = 272$

Mean Deviation about Median $= \frac{1}{n} \sum \lvert X - M_e \rvert$

$= \frac{1}{10} \times 272$

$= 27.2.$

Illustration 7

Find'the Mean Deviation about Mean of the following series :

X	:	10	11	12	13	14	Total
Frequency		3	12	18	12	3 .	4 8 '

15

Solution

Calculation of Mean Deviation about $\overline{X}$

X	f	fx	$\lvert X - \overline{X} \rvert$ i.e, $\lvert X - 12 \rvert$	$f\lvert X - \overline{X} \rvert$
10	3	30	2	6
11	12	132	1	12
12	18	216	0	0
13	12	156	1	12
14	3	42	2	6
Total	48	576	—	36

Note : $\overline{X} = \frac{\sum fx}{n} = \frac{1}{48} \times 576 = 12$

Mean Deviation (about Mean) $= \frac{1}{n} \sum f\lvert X - \overline{X} \rvert$

$= \frac{36}{48} = 0.75.$

Illustration 8

From the following grouped data relating to the sales of 100 Companies, find out Coefficient of Mean Deviation by using mean $(\overline{X})$.

Sales (Rs. '000)	No. of Companies
40 – 50	5
50 – 60	15
60 – 70	25
70 – 80	30
80 – 90	20
90 – 100	5

Solution

To construct average deviation, we have to construct the following table :

Sales (Rs. '000)	Mid Values (X)	No. of Companies (f)	fX	$\lvert X - \overline{X} \rvert$ i.e., $\lvert X - 71 \rvert$	$f\lvert X - \overline{X} \rvert$
40 - 50	45	5	225	26	130
50 – 60	55	15	825	16	240
60 – 70	65	25	1,625	6	150
70 - 80	75	30	2.250	4	120
80 - 90	85	20	1.700	14	280
90 - 100	95	5	475	24	120
Total		n = 100	$\sum fX = 7,100$	$\sum f\lvert X - \overline{X} \rvert = 1,040$	

$\overline{X} = \frac{1}{n} \sum fx = \frac{7,100}{100} = 71$

Mean Deviation (about Mean) $= \frac{1}{n} \sum f\lvert X - \overline{X} \rvert$

$= \frac{1}{100} \times 1.040 = 10.40$ or Rs. 10.4 thousands.

Coefficient of Mean Deviation $= \frac{\text{Mean Deviation about } \overline{X}}{\overline{X}}$

$= \frac{10.40}{71} = 0.146.$

Illustration 9

The following is the age.distribution of 80 LIC Policy holders insured through an agent. Calculate the coefficient of mean deviation from the median.

Age Group (in Years)	Frequency
16 – 20	8
21 – 25	15
26 – 30	13
31 – 35	20
36 – 40	11
41 – 45	7
46 – 50	3
51 – 55	2
56 – 60	1

Solution

Calculation of Mead Deviation from Median

Age-Group (In Years)	Frequency (f)	Cumulative Frequency (Cf)	Class Mid-Point (M)	$\|X - M_d\|$ i.e., $\|X - 31.5\|$	$f\|X - M_d\|$
16 – 20	8	8	18	13.5	108.0
21 – 25	15	23	23	8.5	127.5
26 – 30	13	36	28	3.5	45.5
31 – 35	20	56	33	1.5	30.0
36 – 40	11	67	38	6.5	71.5
41 – 45	7	74	43	11.5	80.5
46 – 50	3	77	48	16.5	49.5
51 – 55	2	79	53	21.5	43.0
56 – 60	1	80	58	26.5	26.5
Total	n = 80				$\sum f\|X - M_d\| = 582.0$

Median has $\frac{N}{2}$ or 40 observations below it. So it lies in the class of 31-35 or 30.5 – 35.5 (in terms of real limits).

$$\text{Median} = l + \frac{\frac{N}{2} - c}{f} \times i$$

$$= 30.5 + \frac{40 - 36}{20} \times 5 = 31.5 \text{ years.}$$

$$\text{Mean Deviation (about Median)} = \frac{1}{n} \sum f |X - M_d|$$

$$= \frac{1}{80} \times 582 = 7.275 \text{ years.}$$

$$\text{Coefficient of Mean Deviation about median} = \frac{\text{M.D. about } M_d}{M_d}$$

$$= \frac{7.275}{31.5} = 0.23.$$

14.10 LET US SUM UP

Dispersion represents the Spread or the scatterness of the data. It is also used to denote the average of deviation of items from some measure of central tendency, Dispersion is calculated to assess the reliability of an average or to compare variability

of two or more data or to control the variation itself. A good measure of dispersion should be based on all observations, should easily be calculated, least affected by sampling fluctuations and amenable to further algebric treatment. Relative measures of dispersion are computed to compare variability in two or more sets of data. They are obtained by expressing absolute measures of dispersion as the ratio of the appropriate average or the sum of two selected items of the data.

The various measures of dispersion in common use are range, quartile deviation, mean deviation and standard deviation. Range is defined as the difference between the highest and the lowest items of the data. It gives the spread of entire data. Quartile deviation is half the difference between Q_1 and Q_3 and is based on middle 50% items only. Mean deviation is the arithmetic mean of the absolute deviations of items from a measure of central tendency, which could be mean or median or sometimes even mode.

Quartile deviation is a suitable measure for open-end data. Range is useful when extreme items are important such as in quality control, price study or meteorological data. As mean deviation is based on all items, in most of the cases it is a better representative of the variability of the data than the other two measures.

14.11 KEY WORDS AND SYMBOLS

Inter Quartile **Range** : A measure of dispersion which considers the spread in the middle 50%. It is $(Q_3 - Q_1)$ of the data.

Mean Deviation : The arithmetic mean of the absolute deviations from the mean median or the mode.

Quartile Deviation : One-half the distance between the first and the third quartiles.

Range : The difference between the largest and the smallest value in a set of data.

List **of Symbols**

In addition to the list of symbols given in Block 3, following are the additional symbols used in connection with measures of dispersion given in this Unit.

Mean Deviation : M.D., 6, A.D.

Mean Deviation about Arithmetic Mean : M.D. $(\bar{X})$, 6 $\bar{X}$, AD $\bar{X}$

Mean Deviation about Median : M D M_d, δM_d, A D M_d

Mean Deviation about Mode : M D (M_o), 6 M_o, δz, AD M_o

Quartile Deviation : Q.D, Q

Range : R

14.12 ANSWERS TO CHECK YOUR PROGRESS

A 3 i) False ii) False iii) True iv) True v) False vi) True

B 4 Range = 39, Q.D. = 9.25

5 Range = 14, Coefficient of Range = 0.58, Q.D. = 2.25, Coefficient of Q.D. = 0.101

6 i) True ii) False iii) False iv) True v) False

14.13 TERMINAL QUESTIONS/EXERCISES

Questions

1 What is the mean deviation? Review its advantages and disadvantages.

1 Calculate quartile deviation and mean deviation about $\overline{X}$ for the following data :

Age (in Years) :	20	30	40	50	60	70	80
No. of Members :	3	61	132	153	140	51	3

(Answer : Q.D. = 10, M.D. $\overline{X}$ = 9.52)

2 A frequency distribution for the duration of 20 long distance telephone calls are shown below :

Calls Duration	Frequency
4 but less than 8	4
8 but less than 12,	5
12 but less than 16	7
16 but less than 20	2
20 but less than 24	1
24 but less than 28	1
Total	20

Compute the mean, median and quartile deviation.
(Answer : Mean = 12.8, Median = 12.6, Q.D. = 3.3)

3 Calculate the mean deviation about Median and coefficient of mean deviation from the following data :

Sales (Rs. '000)	No. of Companies
Less than 20	3
Less than 30	9
Less than 40	20
Less than 50	23
Less than 60	25

(Answer : M.D. about M_e = 8.9, Coefficient of M.D. about M_e = 0.29)

4 A. survey of domestic consumption of electricity gave the following distribution of the units consumed. Compute the quartile deviation and its coefficient.

No. of Units	No. of Consumers
Below - 200	9
200 - 400	18
400 - 600	27
600 - 800	32
800 - 1,000	45
1,000 - 1,200	38
1,200 - 1,400	20
1.400 and above	11

(Answer : Q.D. = 520.6, Coefficient of Q.D. = 0.317)

5 Calculate the mean deviation about the mean and median from the following data :

Class Interval :	0-9	10-19	20-29	30-39	40-49	50-59
Frequency :	15	36	53	42	17	2

(Answer : M.D. $\overline{X}$ = 9.10, M.D. M_e = 9.08)

6 Calculate the mean deviation about Mode and its coefficient for the following data :

No. of Defects per Item	Frequency
0 - 5	18

5 – 10	32
10 – 15	50
15 – 20	75
20 – 25	125
25 – 30	150
30 – 35	100
35 – 40	90
40 – 45	80
45 – 50	50

(Answer : **M . D** M_o = 9.02, Coefficient **M . D** M_o = 0.338)

7) **Compute the mean deviation and its coefficient for the following data :**

No. of Shares Applied for	No. of Applicants
50–100	2,500
100 – 150	1,500
150–200	1,300
200 – 250	1,100
250 – 300	900
300 – 350	750
350 – 400	675
400–450	525
450 – 500	450

(Answer : **M . D** M_e = 102.13, Coefficient of M.D. M_e = 0.011)

Note : These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University, These are for your practice only.

UNIT 15 MEASURES OF DISPERSION - II

Structure

15.0 OBJECTIVES

After studying this unit, you should be able to

- define and compute standard deviation and coefficient of variation for different kinds of data,
- explain the merits and limitations of standard deviation,
- draw Lorenz Curve and graphically determine the extent of unequalities of items,
- compare different measures of dispersion and use them at appropriate places.

15.1 INTRODUCTION

In Unit 14 you have learnt about three measures of dispersion viz., range, quartile deviation and mean deviation. The first two are based on two selected items of the data and the third measure is computed by using the values of each and every item. But in the calculation of mean deviation the negative signs of the deviations of items from a central tendency are ignored. As we ignore the signs which arise during calculations, mean deviation suffers from certain limitations. There is another measure of dispersion viz., standard deviation, which takes care of the problem of signs. In this unit you will learn about the method of computing standard deviation and its coefficient for different kinds of data, their merits, limitations and uses. You will also learn the Lorenz Curve which is the graphical method of finding the dispersion.

15.2 STANDARD DEVIATION

As discussed earlier, while computing the mean deviation we ignore the negative signs of the deviations of the items from the central tendency. This is because in dispersion we are interested only in knowing how much, on an average, items deviate from central tendency irrespective of the fact that items are less than or more than central tendency. This ignoring of signs which arise during calculations, introduces some limitations on the measure. A mathematical solution for ignoring signs is squaring. As the square of any negative item becomes positive, a new measure of dispersion is defined in which deviations are first squared (to ignore the signs) and then averaged out. The value so obtained gives the average of the squares of the deviations and not of deviations directly. So, finally a square root of this value is extracted. Thus the result obtained will give an indirect average of deviations. As this measure is calculated by finding square root of the mean of the squares of the deviations of items from central tendency, it is called Root Mean Square Deviation. Like mean deviation, root mean square deviation can also be calculated by

5 – 10	32
10 – 15	50
15 – 20	75
20 – 25	125
25 – 30	150
30 – 35	100
35 – 40	90
40 – 45	80
45 – 50	50

(Answer : M.D. M_o = 9.02, Coefficient **M . D** M_o = 0.338)

7) **Compute the mean deviation and its coefficient for the following data :**

No. of Shares Applied for	No. of Applicants
50 – 100	2,500
100 – 150	1.500
150 – 200	1,300
200 – 250	1,100
250 – 300	900
300 – 350	750
350 – 400	675
400 – 450	525
450 – 500	450

(Answer : M.D. M_e = 102.13, Coefficient of **M . D** M_e = 0.011)

Note : These.questions and exercises will.help you to understand the unit better. Try to write answers for them..But do not submit your answers to the University, These are for your practice only,

UNIT 15 MEASURES OF DISPERSION - II

Structure

15.0 OBJECTIVES

After studying this unit, you should be able to

- define and compute standard deviation and coefficient of variation for different kinds of data,
- explain the merits and limitations of standard deviation,
- draw Lorenz Curve and graphically determine the extent of unequalities of items,
- compare different measures of dispersion and use them at appropriate places.

15.1 INTRODUCTION

In Unit 14 you have learnt about three measures of dispersion viz., range, quartile deviation and mean deviation. The first two are based on two selected items of the data and the third measure is computed by using the values of each and every item. But in the calculation of mean deviation the negative signs of the deviations of items from a central tendency are ignored. As we ignore the signs which arise during calculations, mean deviation suffers from certain limitations. There is another measure of dispersion viz., standard deviation, which takes care of the problem of signs. In this unit you will learn about the method of computing standard deviation and its coefficient for different kinds of data, their merits, limitations and uses. You will also learn the Lorenz Curve which is the graphical method of finding the dispersion.

15.2 STANDARD DEVIATION

As discussed earlier, while computing the mean deviation we ignore the negative signs of the deviations of the items from the central tendency. This is because in dispersion we are interested only in knowing how much, on an average, items deviate from central tendency irrespective of the fact that items are less than or more than central tendency. This ignoring of signs which arise during calculations, introduces some limitations on the measure. A mathematical solution for ignoring signs is squaring. As the square of any negative item becomes positive, a new measure of dispersion is defined in which deviations are first squared (to ignore the signs) and then averaged out. The value so obtained gives the average of the squares of the deviations and not of deviations directly. So, finally a square root of this value is extracted. Thus the result obtained will give an indirect average of deviations. As this measure is calculated by finding square root of the mean of the squares of the deviations of items from central tendency, it is called Root **Mean Square Deviation.** Like mean deviation, root mean square deviation can also be calculated by

subtracting arithmetic mean or median or mode. Out of these three values, in every data, root mean square deviation about arithmetic mean is the least. So it is called Standard Deviation. Now let us study the meaning, methods of computation, merits and limitations of standard deviation.

15.2.1 Meaning

Standard deviation may be defined as the square root of the arithmetic mean of the squares of deviations of given observations from their arithmetic mean. It is usually denoted by the Greek letter σ (sigma). The major steps involved in the computation-- of standard deviation are as follows:

1) Compute the arithmetic average of the given series.

2) Calculate the deviations of various items from the arithmetic mean.

3) Compute the squares of all the individual deviations.

4) Total the squared deviations and devide the sum by the number of items.

5) Square root of the resultant figure is the standard deviation of the series.

For a set of N observations $X_1, X_2, \ldots, X_n$ with mean $\overline{X}$, deviations from mean will be $(X_1 - \overline{X}), (X_2 - \overline{X}), \ldots, (X_n - \overline{X})$. So the square of deviations from mean will be $(X_1 - \overline{X})^2, (X2 - \overline{X})^2, \ldots, (X_n - \overline{X})^2$. Therefore, the mean square deviations from the arithmetic mean will be:

$$\frac{1}{n}\left[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right] = \frac{1}{n}\sum (X - \overline{X})^2$$

Hence Standard Deviation $(\sigma) = \sqrt{\frac{1}{n}\sum (X - \overline{X})^2}$

This way of defining also helps us in understanding the method of calculating standard deviation.

Illustration **1** .

Find out standard deviation for a set of item's : **3,** 4, 6, 7, 15, 25.

Solution

Computation of Standard Deviation

Item No.	X	$(X - \overline{X})$	$(X - \overline{X})^2$
1	3	–7	49
2	4	–6	36
3	6	– 4	16
4	7	–3	9
5	15	5	25
6	25	15	225
Total	$\Sigma X = 60$	$\Sigma(X - \overline{X}) = 0$	$\Sigma (X - \overline{X})^2 = 360$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{60}{6} = 10$$

Standard Deviation $(\sigma) = \sqrt{\frac{1}{n}\Sigma (X - \overline{X})^2} = \sqrt{\frac{360}{6}} = \sqrt{60}$

$= 7.7$ approximately

The definition given above needs some adjustment for grouped data. If the data is grouped, the squares of the deviations of items (or the mid values of the class interval) from their A.M./are first multiplied by their respective frequencies, then added and divided by the total of frequencies. Therefore, standard deviation for a grouped data will be given by $\sqrt{\frac{1}{n}\Sigma f (X - \overline{X})^2}$ or $\sqrt{\frac{1}{n}\Sigma f (m - \overline{X})^2}$ where 'm' is the mid point of the class intervals.

The square of standard deviation is called variance. Thus, for an ungrouped data, variance is given by $\frac{1}{n}\Sigma(X - \overline{X})^2$ and for grouped data it is given by $\frac{1}{n}\Sigma f(X - \overline{X})^2$ or $\frac{1}{n}\Sigma f(m - \overline{X})^2$ (when class intervals are given). Let us take some illustrations to understand the definition and steps involved in the calculation of standard deviations.

Illustration 2

Calculate the standard deviation and variance from the following data

x	:	10	12	14	16	18	20	22
f	:	3	5	9	16	8	7	2

Solution

Calculation of Standard Deviation

x	f	fX	$(X - \overline{X})$	$f(X - \overline{X})^2$
10	3	30	−6	108
12	5	60	−4	80
14	9	126	−2	36
16	16	256	0	0
18	8	144	2	32
20	7	140	4	112
22	2	44	6	72
	N = 50	$\Sigma fX = 800$		$\Sigma f(X - \overline{X})^2 = 440$

$$\overline{X} = \frac{\Sigma fX}{n}$$

$$= \frac{800}{50} = 16$$

$$\sigma = \sqrt{\frac{1}{n}\Sigma f(X - \overline{X})^2}$$

$$= \sqrt{\frac{1}{50} \times 440} = 2.97$$

$$\text{Variance} = \frac{1}{n}\Sigma f(X - \overline{X})^2$$

$$= \frac{1}{50} \times 440 = 8.8$$

Illustration 3

Find out the standard deviation and variance from the following frequency distribution :

Marks	:	0-4	4-8	8-12	12-16
No. of Students	:	4	8	2	1

Solution

Calculation of standard deviation and variance :

Marks	f	m (Mid-point)	fm	$(m - \overline{X})$	$(m - \overline{X})^2$	$f(m - \overline{X})^2$
0 - 4	4	2	8	−4	16	64
4 - 8	8	6	48	0	0	0
8-12	2	10	20	4	16	32
12 - 16	1	14	14	8	64	64
Total	N = 15		$\Sigma fm = 90$			$\Sigma f(m - \overline{X})^2 = 160$

$$\overline{X} = \frac{\sum fX}{n}$$

$$= \frac{90}{15} = 6$$

$$\sigma = \sqrt{\frac{1}{n}\sum f(m - \overline{X})^2}$$

$$= \sqrt{\frac{1}{15} \times 160} = 3.27 \text{ approximately}$$

$$\text{Variance} = \frac{1}{n}\sum f(X - \overline{X})^2$$

$$= \frac{160}{15} = 10.67 \text{ approximately.}$$

15.2.2 Computation

There are two methods of calculating standard deviation : i) direct method and 2) short-cut method. Let us study these two methods:

Direct Method : It is a method in which calculations are made by directly using the definitions. So this method is same as given in the previous sections. So the formulae for finding standard deviation under this method are:

For ungrouped data $\sigma = \sqrt{\frac{1}{n}\sum (X - \overline{X})^2}$

For grouped data $\sigma = \sqrt{\frac{1}{n}\sum f(X - \overline{X})^2}$ or $\sqrt{\frac{1}{n}\sum f(m - \overline{X})^2}$

When the size of items and their numbers are not large, we can find the standard deviation by directly using the sum of items and the sum of the squares of the items. The extra step of first calculating arithmetic mean and then finding the deviations of items from arithmetic mean is not necessary. The formulas used are given below.

For ungrouped data $\sigma = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$ or $\sqrt{\frac{\sum X^2}{n} - \overline{X}^2}$ as $\overline{X} = \frac{\sum X}{n}$

For grouped data $\sigma = \sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2}$ or $\sqrt{\frac{\sum fX^2}{n} - \overline{X}^2}$ as $\overline{X} = \frac{\sum fX}{n}$

It can be proved mathematically that the second set of formulas and the first set given earlier are identical, and give the same result.

Short-cut **Method** : When data is huge or arithmetic mean comes out in fractions, standard deviation can also be calculated by taking deviations from assumed mean (A). The formulas are given below :

For ungrouped data $\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ where $d = X - A$

For grouped data $\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$ or $C\sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2}$

Where d' is the step deviation and is given by $d' = \frac{X - A}{C}$ and C is common factor in the x-a or 'd' column. Let us take some illustrations to explain the computations involved in using these formulas.

Illustration 4

Calculate standard deviation for the following series by direct method and short-cut method using assumed mean as 32.

Serial No.	:	1	2	3	4	5	6	7	8	9	10
Size	:	20	22	27	30	31	32	35	40	45	48

Solution

Direct **Method** : Calculation of Standard Deviation

S. No.	X	$(X - \overline{X})$	$(X - \overline{X})^2$
1	20	−13	169
2	22	−11	121
3	27	− 6	36
4	30	− 3	9
5	31	− 2	4
6	32	− 1	1
7	35,	2	4
8	40	7	49
9	45	12	144
10	48	15	225
	$\Sigma X = 330$		$\Sigma (X - \overline{X})^2 = 762$

Now $\overline{X} = \frac{\Sigma x}{n} = \frac{330}{10} = 33$

$$\sigma = \sqrt{\frac{\Sigma (X - \overline{X})^2}{n}} = \sqrt{\frac{762}{10}} = \sqrt{76.2} = 8.73.$$

Short-cut Method : Calculation of Standard Deviation

S. No.	X	$d = X - 32$	d^2
1	20	−12	144
2	22	−10	100
3	27	− 5	25
4	30	− 2	4
5	31	− 1	1
6	32	0	0
7	35	3	9
8	40	8	64
9	45	13	169
10	48	16	256
		$\Sigma d = 10$	$\Sigma d^2 = 772$

Now $\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$

$$= \sqrt{\frac{772}{10} - \left(\frac{10}{10}\right)^2} = \sqrt{77.2 - 1}$$

$$= \sqrt{76.2} = 8.73.$$

You may note that the results obtained by both the methods are the same.

Illustration 5

Calculate the standard deviation and variance from the following frequency distribution by direct and short-cut methods using 14 as assumed mean.

X	:	10	12	14	16	18	20	22
f	:	3	5	9	16	8	7	2

Solution

Calculation of Standard Deviation and Variance: -

Direct Method

X	f	fX	$(X - \overline{X})$	$f(X - \overline{X})$	$f(X - \overline{X})^2$
10	3	30	−6	−18	108
12	5	60	−4	−20	80
14	9	126	−2	−18	36
16	16	256	0	0	0
18	8	144	2	16	32
20	7	140	4	28	112
22	2	44	6	12	72
	N = 50	∑fX = 800		∑f (x − x̄) = 0	∑f (x − x̄)² = 440

$$\text{Now } \overline{X} = \frac{\sum fx}{n} = \frac{800}{50}$$

$$= 16$$

$$\sigma = \sqrt{\frac{1}{n}\sum f(X - \overline{X})^2}$$

$$= \sqrt{\frac{440}{50}} = \sqrt{8.8}$$

$$= 2.97$$

Variance $= \sigma^2 = 8.8$.

Short-cut Method

X	f	d = X − 14	fd	fd^2
10	3	−4	−12	48
12	5	−2	−10	20
14	9	0	0	0
16	16	2	32	64
18	8	4	32	128
20	7	6	42	252
22	2	8	16	128
			∑fd = 100	$\sum fd^2 = 640$

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$$

$$= \sqrt{\frac{640}{50} - \left(\frac{100}{50}\right)^2}$$

$$= \sqrt{12.8 - 4}$$

$$= \sqrt{8.8}$$

$$= 2.97$$

Variance $= \sigma^2 = 8.8$.

You may note that when arithmetic mean is in whole numbers, there is not much simplification in calculations by short-cut method.

Illustration 6

The profits (in Rs. lakhs) earned by 100 companies during 1987-88 are shown below. Compute (a) Mean, (b) Variance, and (c) Standard Deviation by using items and their squares.

Profits (Rs. lakhs)	No. of Companies
20-30	4
30-40	8
40-50	18
50-60	30
60-70	15
70-80	10
80-90	8
90-100	7

Solution

Class	Mid-point (X)	Frequency (f)	fX	fX^2
20 – 30	25	4	100	2,500
30 – 40	35	8	280	9,800
40 – 50	45	18	810	36,450
50 – 60	55	30	1,650	90,750
60 – 70	65	15	975	63,375
70 – 80	75	10	750	56,250
80 – 90	85	8	680	57,800
90 – 100	95	7	665	63,175
Total		100	5,910	3,80,100

a) Mean $(\overline{X}) = \frac{\Sigma fx}{n} = \frac{5,910}{100}$

$= 59.10.$

b) Variance $= \sigma^2 = \frac{\Sigma fX^2}{n} - \left(\frac{\Sigma fX}{n}\right)^2 = \frac{380100}{100} - \left(\frac{5910}{100}\right)^2$

$= 3801.00 - 3492.81$

$= 308.19$

c) Standard Deviation $= \sqrt{\text{Variance}} = \sqrt{308.19}$

$= 17.56.$

In this illustration you may notice that by using sums of items and their squares the total calculations involved are large. This method is a direct method in the sense that we have used the items directly and not calculated their deviations from any value. This method may be used only when size of items are small and their total number is also small.

Illustration 7

Calculate Mean and Standard Deviation from the following distribution :

Class-Interval	:	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	:	4	8	8	16	12	6	4

Solution

Let us use the short-cut method, a method which is most commonly used and involves least amount of lengthy calculations. Like calculations of arithmetic mean the assumed mean is taken as one of the mid-points which is towards the middle and corresponds to a high frequency. The deviations so obtained are divided by the common factor, if any. When we divide them by the common factor, this method is also called **step deviation method.**

Calculations of mean and standard deviation.

Class-Interval	f	Mid-point (X)	$d = X - A$ $(X - 45)$	$d' = \frac{d}{10} = \frac{d}{C}$	fd′	fd'^2
10–20	4	15	−30	−3	−12	36
20–30	8	25	−20	−2	−16	32
30–40	8	35	−10	−1	−8	8
40–50	16	45	0	0	0	0
50–60	12	55	+10	1	12	12
60–70	6	65	+20	2	12	24
70–80	4	75	+30	3	12	36
Total	n = 58				$\Sigma fd' = 0$	$\Sigma f(d')^2 = 148$

Here assumed mean (A) is 45, and common factor (C) is 10 which is also the class interval.

$$\text{Mean } \overline{X} = A + \frac{Cfd'}{n} \times C$$

$$= 45 + \frac{0}{38} \times 10 = 45$$

$$\text{Standard Deviation} = C \times \sqrt{\frac{\Sigma fd'^2}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$$

$$= 10 \times \sqrt{\frac{148}{58} - \left(\frac{0}{50}\right)^2}$$

$$= 10 \times \sqrt{2.552} = 1.597 \times 10 = 15.97.$$

Illustration 8

Find the standard deviation of the following distribution :

Class-Interval :	0-500	500-1000	1000-1500	1500-2000	2000-3000
Frequency	90	218	86	41	15

Solution

Calculation of Standard Deviation

Class-Interval	Frequency (f)	Mid-point (m)	$d' = \frac{m-750}{250}$	fd′	$f(d')^2$
0 - 500	90	250	-2	-180	360
500 - 1000	218	750	0	0	0
1000 - 1500	86.	1250	2	172	344
1500 - 2000	41	1750	4	164	656
2000 - 3000	15	2500	7	105	735
Total	n = 450	—	—	261	2095

Here, assumed mean A is 750 and common factor C is 250.

$$\text{S.D.} = C \times \sqrt{\frac{\Sigma fd'^2}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$$

$$= 250 \times \sqrt{\frac{2095}{450} - \left(\frac{261}{450}\right)^2}$$

$$= 250 \times \sqrt{4.6556 - (0.58)^2}$$

$$= 250 \times \sqrt{4.3192} = 519.6 \text{ approximately.}$$

You may note that when class intervals are not equal the step deviation d' may not be integers in order i.e., 1, 2, 3, ... or $-1, -2, -3$, ... etc.

Check Your Progress A

1) Define standard deviation.

.....

.....

.....

.....

2) Write the formulae used for computing standard deviation.

.....

.....

.....

.....

.....

.....

3) Compute standard deviation for the following set of observations.

245, 322, 192, 310. 231

.....

.....

.....

.....

4) Calculate the variance for the following data:

Value:	130-139	140-149	150-159	160-169	170-179	180-189	190-199
f :	1	4	14	20	22	12	2

.....

.....

.....

.....

.....

.....

.....

5) State whether the following statements are True or False.

i) Variance can be called as mean square deviation.

ii) Standard deviation can be negative.

iii) Root mean square deviation can have more than one value in a given data.

iv) Standard deviation is not a particular case of root mean square deviation.

v) Different methods of calculating standard deviation will give different results.

15.2.3 Properties

You have learnt the meaning and methods of computing standard deviation. Now let us study the important properties of standard deviation.

1) The value of standard deviation remains the same if each of the observations in a series is increased or decreased by a constant value. Thus, if $Y = X + K$, where K is a constant quantity, then standard deviation Y is equal to standard deviation of X. In other words, standard deviation is independent of change of origin.

For example :

	X	$X - \overline{X}$	$(X - \overline{X})^2$	Let $Y = X + 10$	$(Y - \overline{Y})$	$(Y - \overline{Y})^2$
	1	−2	4	1+10 = 11	−2	4
	2	−1	1	2+10 = 12	−1	1
	3	0	0	3+10 = 13	0	0
	4	1	1	4+10 = 14	1	1
	5	2	4	5+10 = 15	2	4
Total	15	0	10	65	0	10

Arithmetic Mean of $X = \frac{\sum X}{n} = \frac{15}{5} = 3$

$$\sigma \text{ of } X = \sqrt{\frac{\sum (X - \overline{X})^2}{n}} = \sqrt{\frac{10}{5}} = \sqrt{2}$$

$$\text{A.M. of } Y = \frac{\sum Y}{n} = \frac{65}{5} = 13$$

$$\sigma \text{ of } Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{n}}$$

$$= \sqrt{\frac{10}{5}} = \sqrt{2}$$

Hence, S.D. of X = S.D. of Y.

2) For a given series, if each observation is multiplied or divided by a constant value, standard deviation will also be similarly affected. Thus, if $Y = A X$, where A is a constant, then S.D. of Y = (S.D. of X) x A.

For example,

X	$X - \overline{X}$	$(X - \overline{X})^2$	$Y = 10X$	$Y - \overline{Y}$	$(Y - \overline{Y})^2$
1	−2	4	10	−20	400
2	−1	1	20	−10	100
3	0	0	30	0	0
4	1	1	40	10	100
5	2	4	50	20	400
15	0	10	150	0	1,000

$$\overline{Y} = \frac{\sum Y}{n} = \frac{150}{5} = 30$$

$$\sigma \text{ of } Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{n}} = \sqrt{\frac{1000}{5}} = \sqrt{200} = 10\sqrt{2}.$$

$\therefore \sigma$ of Y = 10 (σ of x)

Thus, you may conclude that the standard deviation is independent of any change of origin but is not independent of the change of scale.

3) For a given set of observations, standard deviation is never less than mean deviation about arithmetic mean and quartile deviation. In fact mean deviation is $\frac{4}{5}\sigma$ and quartile deviation is $\frac{2}{3}\sigma$ for normal data.

4) If two groups contain n_1 and n_2 observations with means $\overline{X}_1$ and $\overline{X}_2$ and standard deviation σ_1 and σ_2 respectively, then the standard deviation of the combined group can be determined. It is given by:

$$\sigma_{12} = \sqrt{\frac{(n_1\sigma_1^2 + n_2\sigma_2^2) + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$$

Where σ_{12} = combined standard deviation of the two groups

$$d_1 = \overline{X}_{12} - \overline{X}_1 \; ; \; d_2 = \overline{X}_{12} - \overline{X}_2$$

X_{12} = combined arithmetic mean of the two groups.

To understand the properties 3 and 4, study Illustrations 13 and 14 given under Section 15.4 ('Some Illustrations') presented later in this unit.

5) Root mean square deviation calculated about a value other than arithmetic mean will always be higher than standard deviation. For explaining this let us again take the values of X same as under (1) above, and calculate root mean square about 4, a value different from mean ($\overline{X}$) which is 3.

X	:	1	2	3	4	5
X−4	:	−3	−2	−1	0	1
$(X-4)^2$	:	9	4	1	0	1

Now $\sum(X-4)^2 = 15$

$\therefore$ Root Mean Square Deviation about 4 $= \sqrt{\frac{\sum(X-4)^2}{n}}$

$$= \sqrt{\frac{15}{5}}$$

$$= \sqrt{3}.$$

But standard deviation of X is $\sqrt{2}$. So, root mean square deviation about a value other than arithmetic mean is greater than standard deviation.

6) In an ordinary type data or normal type data (the meaning of normal data will be explained in moredetail in Unit 16) the number of items between the range A.M. $\pm$ a is about 68%, in the range A.M. $\pm$ 2 σ is about 95% and in range A.M. $\pm$ 3 σ is almost all the items of the data lie.

To explain it, let us consider the data of Illustration 5. For this data A.M. is 16 and σ is 2.97. So the range A.M. $\pm$ a will be 16 $\pm$ 2.97 or 13.03 to 18.97. In the data, number of items lying between 13.03 to 18.97 are 9+16+8 or 33 i.e., 66% of total items (i.e., 50) which is quite close to 68%. Similarly, the range A.M. $\pm$ 2 a will be 16 $\pm$ 2 $\times$ 2.97 or 10.06 to 21.94.

All items except the items of the first and the last group fall in this range. Thus, total number of items in the range 10.06 to 21.94 are 45 i.e., 90%, a value not very much different from 95%. You can also verify whether or not 100% items lie within the range A.M. $\pm$ 3 σ.

The percentages of items' lying between different ranges calculated above are not exactly the same as stated in the property. This only points out that the data of Illustration 5 is not perfectly normal but is quite close to it.

15.2.4 Merits and Limitations

Merits : Among all the measures of dispersion, standard deviation is considered superior because it possesses almost all the requisites of a good measure of dispersion. Standard deviation had the following merits :

i) It is rigidly defined and is based on all observations of the series.

ii) The unique property which makes standard deviation superior to other measures of dispersion is that it is amenable to algebric treatment. Thus, if we are given the number of observations, mean and standard deviation for each of several groups, we can easily calculate the standard deviation of the composite group.

iii) Standard deviation is least affected by the fluctuations of sampling.

iv) In a normal distribution, the mean ± S.D. covers 68.36% of the values whereas only 50% values are covered by quartile deviation and 57% by mean deviation. Because of this reason, standard deviation is called a 'standard measure'.

Limitations : The main Iimitations or demerits of standard deviation as a measure of dispersion are as follows:

i) The major limitation of SD is that it cannot be used for comparing the dispersion of two or more series of observations given in different units. A coefficient of standard deviation has to be defined for this purpose.

ii) The process of squaring deviations from mean and then taking the square-root of the mean of these squared deviations seems to be a complicated affair.

In fact this gives rise to another limitation i.e., standard deviation is very much affected by the extreme values. The process of squaring deviations give undue importance to large deviations from arithmetic mean which are obtained only from extreme iterns and it gives less importance to items which are nearer to mean.

iii) The standard deviation cannot be computed for a distribution with open-cnd classes.

15.3 COEFFICIENT OF VARIATION

The coefficient of variation, also known as coefficient of standard deviation expressed in percentages, is based on the ratio of the standard deviation to the arithmetic mean of a series. Thus, coefficient of variation may be expressed. as:

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Arithmetic Mean}} \times 100$$

The coefficient of variation is a relative measure of dispersion and is usually expressed in the form of percentage. So it can be conveniently used for comparing the variability or dispersion between the two sets of the observations given in different units or if units are same, have wide variations in the average value. It may. thus, be used to measure or compare the precision of two or more sets of observations.

To understand this point let us take an example. Suppose we measure the distance between Delhi and Bombay and make a deviation of 1 km or 1,00,000 cm in the actual distance of 1540 kms. This deviation is of hardly any significance as compared to a deviation of 10 cm in measuring a piece of one meter cloth. This fact is not revealed when 1,00,000 cm deviation in first case is compnrcd directly with 10 cm deviation of the second case. As 1,00,000 cm is larger than 10 cm one may be tempted to conclude that deviation of measurement in first case is very much important. But if we compute coefficients, the picture becomes clear. In first case coefficient is only $\frac{1}{1540} \times 100 = 0.065\%$ and in the second case the coefficient is $\frac{10}{1000}$ x 100 or 1%. So. deviations in second case is relatively larger. Thus, whenever comparisons of variation is to be done it must be done in terms of coefficient of variation only.

Illustration 9

The following is the record of goals scored by Team A in a football season.

No. of goals scored in a match	: 0	1	2	3	4
Number of matches	: 1	9	7	5	3

For Team B, the average number of goals scored per match was 2.5 with a standard deviation of 1.25 goals. Find which team is more consistent.

Solution

Computation of Arithmetic Mean and Standard Deviation of Team A

No. of Goals	No. of Matches (f)	Deviation (d)	fd	fd^2
0	1	−2	−2	4
1	9	−1	−9	9
2	7	0	0	0
3	5	1	5	5
4	3	2	6	12
	n = 25		$\Sigma fd = 0$	$\Sigma fd^2 = 30$

Arithmetic Mean of Team A : $= A + \frac{\Sigma fd}{n}$

$= 2 + \frac{0}{25} = 2$

Standard Deviation of Team A : $= \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2}$

$= \sqrt{\frac{30}{25} - \left(\frac{0}{25}\right)^2}$

$= \sqrt{1.2 - 0} = \sqrt{1.2}$

$= 1.1.$

Coefficient of Variation of Team A :

$= \frac{S.D.}{A.M.} \times 100$

$= \frac{1.1}{2} \times 100 = 55\%.$

Coefficient of Variation of Team B :

$= \frac{S.D.}{A.M.} \times 100 = \frac{1.25}{2.5} \times 100 = 50\%.$

The coefficient of variation of Team B is less than that of Team A. So, Team B is considered to be more consistent than Team A.

Illustration 10

From the data given below, state which series is more variable :

Variable	Series A	Series B
10–20	10	18
20–30	18	22
30–40	32	40
40–50	40	32
50–60	22	18
60–70	18	10

Solution

Computation of Arithmetic Mean and Standard Deviation of Series A

Class-Interval (Variable)	Mid-Value	Frequency (f_1)	Step Deviation (d_1)	f_1d_1	$f_1d_1^2$
10–20	15	10	−2	−20	40
20–30	25	18	−1	−18	18
30–40	35	32	0	0	0
40–50	45	40	1	40	40
50–60	55	22	2	44	88
60–70	65	18	3	54	162
		140		100	348

Here **A, is 35** and C_1 is 10.

$$\overline{X}_1 = A_1 + \frac{\sum f_1 d_1}{n_1} \times C_1$$

$$= 35 + \frac{100}{140} \times 10$$

$$= 35 + 7.143 = 42.1 \text{ approximately.}$$

$$\sigma_1 = C_1 \times \sqrt{\frac{\sum f_1 d_1^2}{n_1} - \left(\frac{\sum f_1 d_1}{n_1}\right)^2}$$

$$= 10 \times \sqrt{\frac{348}{140} - \left(\frac{100}{140}\right)^2}$$

$$= 10 \times \sqrt{2.486 - 0.510}$$

$$= 1.406 \times 10 = 14.06$$

$$\textbf{C.V.}(\text{Seric } \mathbf{A}) = \frac{\sigma_1}{\overline{X}_1} \times 100$$

$$= \frac{14.06}{42.1} \times 100 = 33.3\%.$$

Computation of Arithmetic Mean and Standard Deviation of Series B

Class-Interval (Variable)	Mid-Value	Frequency (f_2)	Step Deviation (d_2)	f_2d_2	$f_2d_2^2$
10–20	15	18	−2	−36	72
20–30	25	22	−1	−22	22
30–40	35	40	0	0	0
40–50	45	32	1	32	32
50–60	55	18	2	36	72
60–70	65	10	3	30	90
		140		40	288

Here A_2 = 35 and C_2 = 10.

$$\overline{X}_2 = A_2 + \frac{\sum f_2 d_2}{n_2} \times C_2$$

$$= 35 + \frac{40}{140} \times 10$$

$$= 35 + 2.85 = 37.85$$

$$\sigma_2 = C_2 \times \sqrt{\frac{\sum f_2 d_2^2}{n_2'} - \left(\frac{\sum f_2 d_2}{n_2}\right)^2}$$

$$= 10 \times \sqrt{\frac{288}{140} - \left(\frac{40}{140}\right)^2}$$

$$= 10 \times \sqrt{2.057 - 0.0784}$$

$$= 10 \times \sqrt{1.9786}$$

$$= 1.406 \times 10$$

$$= 14.06$$

$$\text{C.V. (Series B)} = \frac{\sigma_2}{\bar{X}_2} \times 100$$

$$= \frac{14.06}{37.85} \times 100$$

$$= 37.1\%.$$

Since the coefficient of variation of Series B is higher than that of Series **A,** Series B is more variable. In this illustration you may notice that standard deviation of both the series is same i.e., **14.06.** From this fact we should not conclude that two series have same variation. The difference inarithmeticmean **has** to be taken into account for correct interpretation.

15.4 SOME ILLUSTRATIONS

Illustration 11

4 state government decided to give old age pension to people over sixty years of age. The scales of pension were fixed as follows:

Age Group	Rs. per month
60–65	250
65–70	300
70–75	350
75–80	400
80–85	450

The ages of 25 persons who secured, the pension rights are given below:

74 62 84 72 83 72 81 64 71 63 61 60 61 67 74
64 79 73 75 76 69 78 66 67 68

Calculate the monthly average pension payable and standard deviation.

Solution

Classifying the Data

Age Group	Talley	Frequency
60–65	𝍸 11	7
65-70	𝍸	5
70–75	𝍸	6
75–80	1111	4
80–85	111	3
		25

Calculation of Monthly Average Pension Payable and the Standard Deviation

Scale of Pension (Rs.)	$d' = \left(\frac{X-350}{50}\right)$	f	fd'	fd'²
250	−2	7	−14	28
300	−1	5	−5	5
350	0	6	0	0
400	1	4	4	4
450	2	3	6	12
		25	−9	49

Here A = 350, C = 50, $\sum f = n = 25$, $\sum fd' = -9$ and $\sum fd'^2 = 49$

$$\bar{X} = A + \frac{\sum fd'}{n} \times C$$

$$= 350 - \frac{9}{25} \times 50 = 332$$

$$\sigma = C \times \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2}$$

$$= 50 \times \sqrt{\frac{49}{25} - \left(\frac{-9}{25}\right)^2} = 1.353 \times 50 = 67.65$$

Thus, the monthly average pension is Rs. 332 and standard deviation is Rs. 67.65.

Correcting Incorrect Value of Standard Deviation : Sometimes it so happens that while calculating mean and standard deviation we unconsciously copy out wrong observations. For example, an observation 25 may be posted as 52. There is a simple procedure to correct the incorrect values of mean and standard deviation. For obtaining correct mean, we find out correct $\sum X$ by deducting the wrong observations from the original $\sum X$ and adding to it the correct observations. Similarly, for calculating correct standard deviation we obtain the value of correct $\sum X^2$. The above procedure is clarified in Illustration 12.

Illustration 12

The mean and standard deviation of a set of 100 observations were worked out as 40 and 5, respectively. For one of the observations, by mistake, the value is taken as 50 instead of 40. Find the correct mean and variance.

Uncorrected $\sum X = n\bar{X} = 100 \times 40 = 4{,}000$

Correct $\sum X = 4{,}000 - 50 + 40 = 3{,}990$

Correct Mean $= \frac{3{,}990}{100} = 39.90$

Now Variance $= \frac{\sum X^2}{n} - \bar{X}^2$

As standard doviation (tr) is 5, variance (ti') $= 5^2 = 25$

Substituting the given values

$$25 = \frac{\sum X^2}{100} - (40)^2$$

$$25 = \frac{\sum X^2}{100} - 1{,}600$$

$$25 = \frac{\sum X^2 - 1{,}60{,}000}{100}$$

$$2500 = \sum X^2 - 1{,}60{,}000$$

$$\sum X^2 = 1{,}60{,}000 + 2{,}500$$

$$= 1{,}62{,}500.$$ This is the value of 'original $\sum X^2$.

Correct $\sum X^2 = 1{,}62{,}500 - (50)^2 + (40)^2$

$$= 1{,}62{,}500 - 2{,}500 + 1{,}600 = 1{,}61{,}600$$

$$\text{Correct Variance} = \frac{\text{Correct}\,\Sigma X^2}{n} - (\text{Correct}\,\overline{X})^2$$

$$= \frac{1,61,600}{100} - (39.9)^2$$

$$= 1,616 - 1,592.01 = 23.99$$

Thus correct mean is 39.9 and correct variance is 23.99.

Combined Standard Deviation : Just as it is possible to compute combined mean of two or more groups, we can also compute combined standard deviation of two or more groups. Combined standard deviation of two groups is denoted by 12 and is computed as follows :

$$\sigma_{12}^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2} \text{ or } \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}$$

Where σ_{12} = Combined standard deviation

σ_1 = Standard deviation of first group

σ_2 = Standard deviation of second group

$d_1 = (\overline{X}_1 - \overline{X}_{12})$

$d_2 = (\overline{X}_2 - \overline{X}_{12})$

$\overline{X}_1$ = Mean of first group

$\overline{X}_2$ = Mean of second group

$\overline{X}_{12}$ = Combined mean.

The above formula can also be extended to find out the standard deviation of three or more groups.

Illustration 13

For a group of 50 male workers, the mean and standard deviation of their daily wages are Rs. 72 and Rs. 9 respectively. For another group of 40 female workers these are Rs. 54 and Rs. 6 respectively. Find the standard deviation for the combined group of 90 workers.

Solution

In thisdata, $n_1 = 50$ and $n_2 = 40$

$\overline{X}_1 = 72$ and $\overline{X}_2 = 54$

$\sigma_1 = 9$ and $\sigma_2 = 6$

Combined mean for group of 90 $(\overline{X}_{12}) = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}$

$$= \frac{50 \times 72 + 40 \times 54}{90}$$

$$= \frac{3,600 + 2,160}{90} = 64$$

Combined Standard Deviation for the group of 90 :

$$\sigma_{12} = \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}$$

Now $d_1 = 64 - 72 = -8$ and $d_2 = 54 - 72 = -18$

$$\sigma_{12}^2 = \frac{50(81+64) + 40(36 + 324)}{90} = \frac{7,250 + 14,400}{90}$$

$$= \frac{21,650}{90} = 240.56$$

$\therefore \sigma_{12} = \sqrt{240.54} = 15.51$

You may note that the combined mean of the two groups has a value in between the means of the two groups but the combined standard deviation has a value much greater than the greatest of the given standard deviations. Combined mean will always be in between the range of the given means, but there is nothing wrong in getting combined standard deviations with a value outside the range of the given standard deviation. In fact, greater the difference between the given means, the combined standard deviation will be more away from the largest given standard deviations. When all the given groups have equal means, then only the combined standard deviation will be between the range of the given standard deviations.

Illustration 14

Calculate mean deviation about mean for data given previously in Illustration 7 and show that mean deviation is less than standard deviation.

Solution

Calculation of Mean Deviation

Class Interval	Frequency (f)	Mid Points (m)	$m-\overline{X}$	$f\|m-\overline{X}\|$ $f\|m-45\|$
10–20	4	15	−30	120
20–30	8	25	−20	160
30–40	8	35	−10	80
40–50	16	45	0	0
50–60	12	55	10	120
60–70	6	65	20	120
70–80	4	75	30	120
Total	58			720

.From Illustration 7, we have' $\overline{X} = 45$ and $\sigma = 15.97$

Mean Deviation about $\overline{X} = \frac{\sum f|m-\overline{X}|}{n} = \frac{720}{58} = 12.41$

Therefore, mean deviation about $\overline{X}$ is less than standard deviation. You should note that mean deviation about mean will always be less than standard deviation whatever may be the data.

Check Your Progress B

1) Indicate whether the following statements are True of False.

   i) Standard deviation is free from all those defects with which the other measures suffer.

   ii) The variance and the coefficient of variation are the same.

   iii) Root mean square deviation about arithmetic mean is the least.

   iv) Reducing each and every item by 5 will reduce standard deviation also by 5.

   v) In an ordinary data, standard deviation is less than Quartile deviation.

2) Fill in the blanks :

   i) Mean deviation is ..... than standard deviation.

   ii) $\overline{X} \pm \sigma$ includes .....per cent of the items.

   iii) If in' a series coefficient of variation is 64 and mean is 10, the standard deviation shall be .....

   iv) Variance is always ..... negative.

   v) If each of the 10 values of a set are equal to 5, the standard deviation will be equal to .....

3) Tick the correct answer :

a) The measures based on every item of the series:
   i) range
   ii) standard deviation
   iii) quartile deviation
   iv) **all** of them

b) One of the measures of dispersion which is more useful in case of open-end distributions:
   i) range
   ii) mean deviation
   iii) standard deviation
   iv) quartile deviation

c) Standard deviation is always computed from:
   i) mean
   ii) mode
   iii) median
   iv) Geometric Mean

d) Which of the following measures is least affected by extreme items:
   i) quartile deviation
   ii) range
   iii) standard deviation
   iv) mean deviation

e) Mean deviation is:
   i) less than S.D.
   ii) more than S.D.
   iii) not related to S.D.
   iv) equal to Standard Deviation

f) Coefficient of variation is given by:
   i) $\frac{\sigma}{\bar{X}}$
   ii) $\frac{\bar{X}}{\sigma}$
   iii) $\frac{\bar{X}}{\sigma} \times 100$
   iv) $\frac{\sigma}{\bar{X}} \times 100$

15.5 LORENZ CURVE

Lorenz Curve, is devised by Dr. Max O. Lorenz who is a famous economic statistician. It is a graphic method of studying dispersion. This curve which was originally used by him to measure the distribution of wealth and income; is now also used to study the distribution of profits, wages and turnover, etc.

The following steps are involved in the construction of Lorenz Curve:

i) The total values of items are obtained from various groups.

ii) Total values of items and frequencies corresponding to various groups are then cumulated in less than type and converted into percentages.

iii) On the X-axis start the scale from 100 and go upto 0 and use it for the percentage of cumulative frequencies (X).

iv) On Y-axis start the scale from 0 and go upto 100, and plot the percentage of the total values of the items (Y).

v) Draw a diagonal line joining the 0 on X-axis with 100 on Y-axis. This is known as **line of equal distribution.** Any point on this diagonal shows the same percentage on X as on Y.

vi). Plot the various points corresponding to X and Y and join them. The line so obtained, unless all items exactly equal, will always forms a curve below the line of equal distribution. The area between line of equal distribution and the plotted curve gives the extent of inequality in the items. If curves of various distributions are shown on the same Lorenz presentation, the curve that is farthest from the diagonal line represents greatest inequality.

Illustration 15

Draw the Lorenz Curve for the data relating to the profit of 50 business firms and. show the extent of inequality present in the profits. .

Profits (Rs. '000)	:	10-20	20-30	30-40	40-50	50-60
No. of Firms	:	5	13	18	10	4

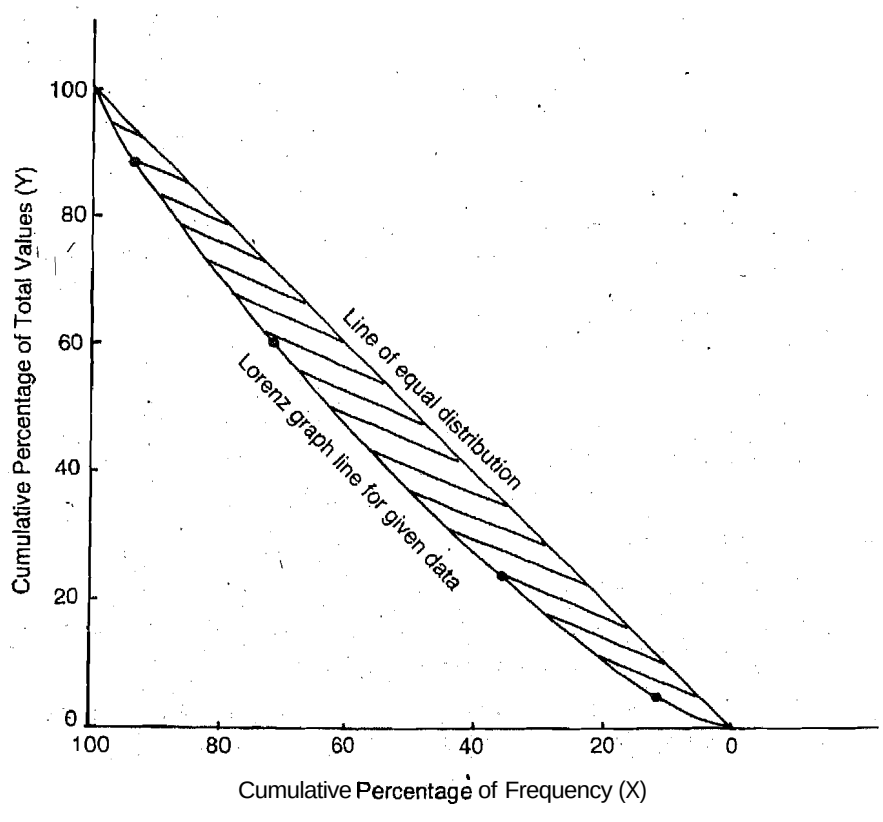
Solution

Computations for Lorenz Curve

Profit (Rs. '000)	Mid-Point (m)	Frequency (No. of Firms) (f)	Cum. Freq.	% Cum. Freq. (X)	Total Value (m x f)	Cum. Total Values	% Cum. Total Values (Y)
10–20	15	5	5	10	75	75	4.4
20–30	25	13	18	36	325	400	23.5
30–40	35	18	36	72	630	1,030	60.6
40–50	45	10	46	92	450	1,480	87.1
50–60	55	4	50	100	220	1,700	100.0

Now you follow the steps explained earlier one by one and draw the Lorenz Curve taking the percentage cumulative frequency (X) on X-axis and the percentage comulative total values (Y) on Y-axis. Look at Figure 15.1 carefully and study how it is drawn.

Figure 15.1



Lorenz Curve Showing the Inequalities in the Profit Distribution for 50 Business Firms

Explanation : The first points 10 on X and 4.4 on Y represent that 10% of items have a total value of 4.4%. Similarly, points 36 on X-axis and 23.5 on Y-axis represent that, if we count from lower values side, 36% of items have 23.5%. of the. total values. You can derive interpretation for other points in the same manner. When no item is taken, there will be no total value. So the point X = 0 and Y = 0 also lies on the Lorenz Curve. Therefore, along with the calculated values of X and Y point (0, 0) is also plotted and then the various points are joined.

As stated earlier, the line joining (100, 100) to (0, 0) is called line of equal distribution. Why is this? This line represents points like (0,0), (20,20), (70,70), etc These points mean that the percentages of items are same as the percentages of total values or total values increaseexactly in the same proportion as items. This is possible only when all items have equal values. Hence the line is called line of equal distribution.

The two axis (X-axis and Y-axis) taken together represents Lorenz Curve Graph for maximum inequality. Maximum inequality means all items except the last, has zero value. The entire value (wealth in the data which Dr. Lorenz took) is concentrated in the last item. So to draw the graph of maximum inequality we have to take points : 0% item with 0% value, 20% items with 0% value, 80% items with 0% value, 99% items with 0% value, etc. And as soon as last item is taken 100% items 100% value, all these points give rise to two axis. Thus, two axis together represents Lorenz Graph of maximum unequality. Graph of any given data, therefore, will be away from the line of equal distribution and will be towards the two axis. Thus the area between the line of equal distribution and the graph of given data (shown by shaded area in the diagram) gives the extent of inequality in the data.

Illustration 16

Draw the Lorenz Graph for the two sets of five workers each whose weekly income figures are given below. Point out which set has greater inequalities in income.

Income of Group A (Rs.) :	96	104	103	99	98
Income of Group B (Rs.) :	100	270	580	620	430

Solution

In this case frequencies are not given. In fact each income figure represents the income of one worker. So the frequencies for various income figures are one each in both Group A and B. Arranging the data in ascending order, the computations for two sets are as follows :

Calculations for Lorenz Graph.

Group A			Group B			For Both Sets		
Income Rs.	Cum. Income	% Cum. (Income) (YA)	Income Rs.	Cum. Income	% Cum. Income (YB)	Freq.	Cum. Freq.	% Cum. Freq. (X)
96	96	19.2	100	100	5.0	1	1	20
98	194	38 8	270	370	18.5	1	2	40
99	293	58.6	430	800	40.0	1	3	60
103	396	79.2	580	1,380	69.0	1	4	80
104	500	100.0	620	2,000	100.0	1	5	100

Now, follow the steps explained earlier, and draw the Lorenz Curves for both the groups. Look at Figure 15.2 carefully and study how the diagram is drawn.

Figure 15.2 shows that, the line for Group B is at a greater distance from line of equal distribution. Compared with the line for workers in Group A, it implies that the incomes .of the workers in Group B have greater inequalities.

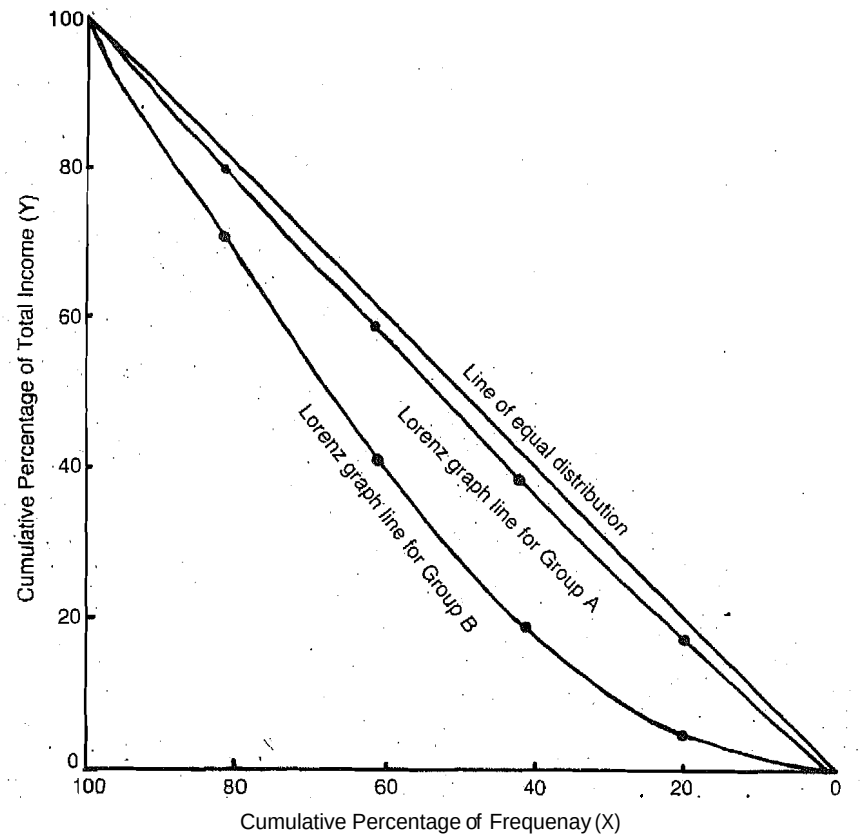


Figure 15.2 Lorenz Curve Showing the Income Inequalities between Two Groups of Workers

15.6 COMPARISON OF MEASURES OF DISPERSION

In this unit and the previous unit, we have discussed the meaning, computation, merits and limitations of various measures of dispersion viz., range, quartile deviation, mean deviation and standard deviation; But in a given situation, you should be able to select the appropriate measure of dispersion. To enable you to make a right choice, it is necessary for you to know the relative features of these measures. Therefore, let us now compare these measures with one another so that we know the relative merits and limitations of these measures of dispersion.

1) **Type** : Range and quartile deviation are measures of dispersion which give the spread of the data, while mean deviation and standard deviation are measures of dispersions which give the average of the deviations of items from some central tendency measure.

2) Calculation : Range is the difference between the values of the upper limit and the lower limit of a series. Quartile deviation is the difference between the values of the upper quartile and lower quartile of a series divided by two. Mean deviation is the sum of all the deviations in absolute terms taken from an average divided by the number of items in a series. Standard deviation is the square root of the arithmetic mean of the squares of all the deviations taken from the arithmetic mean.

3) **Result** : Range is simple and easy to understand. Quartile deviation, mean deviation, and standard deviation are not so. Quartile deviation and mean deviation are to some extent understandable: But standard deviation is comparatively complicated and abstract.

3) **Items** : Range and quartile deviation calculations do not take into account all the items in a series. All the items in a series are taken into consideration when mean deviation and standard deviation are being calculated.

5) **Treatment** : Range, quartile deviation and mean deviation are not capable of mathematical treatment. Standard deviation is capable of mathematical treatment.

6) Extreme **Values** : Quartile deviation is not affected by the extreme or abnormal values of the items in a series. Between mean deviation and standard deviation, mean deviation is less affected by the extreme values. Range depends only on extreme items.

7) **Open-end Class** : Range, mean deviation, and standard deviation cannot be calculated in case of a frequency distribution with open end class intervals. Quartile may be calculated for such a distribution.

8) **Reliability** : Standard deviation is considered to be the most reliable and dependable measure of dispersion. Range or quartile deviation or mean deviation is not such a reliable and dependable measure of dispersion. In fact standard deviation is least affected by sampling errors.

9) **Use** : Standard deviation is considered to be the best measure of dispersion. It possesses all the qualities and properties of **a** good and reliable measure of dispersion. Hence it is widely used in statistical analysis and treatment. Range, quartile deviation and mean deviation are not so popular and are used only in limited but appropriate cases.

15.7 LET US SUM UP

While calculating mean deviation, the signs of the deviations are ignored. This introduces some limitations in the measure. To overcome such limitations, a new measure called Root Mean Square Deviation is defined to measure dispersion. It is the square root of the mean of the deviations of items from central tendency.

Root mean square deviation about arithmetic mean is the least and is given the name standard deviation. For computing standard deviation, there are two methods : 1) direct method and 2) short-cut method. Short cut method, using step deviations, is most common in use. The formula for it is : Standard Deviation $(\sigma) = C \times \sqrt{\frac{\Sigma fd'^2}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$. Standard deviation is rigidly defined and based on all items. It is amenable to algebric adjustments and is least affected by sampling fluctuations. But it is affected by extreme items much more than mean deviation. It has some mathematical properties also. It is independent of change of origin but is affected by change of scale to the same extent as items. Its value is never less than mean deviation. For normal type data, mean deviation is about 4/5 standard deviation and quartile deviation is about 213 standard deviation. There are 68% items in the range AM ± SD and about 95% items in the range AM ± 2SD.

The graphic method of determining dispersion is Lorenz Curve. This is devised by Max O. Lorenz. This is a double cumulative percentage curve. On X-axis cumulative percentage of frequencies are taken and the scale starts from 100 and goes to 0 as we move to right., Cumulative percentage of total values of items are plotted on Y-axis with scale starting from 0 and going upto 100, as we move up. The line joining the points (0, 0) to (100, 100) is called line of equal distribution and the two axes together are Lorenz Curve for the maximum inequality, So the area between the line of equal distribution and graph of any data represents the extent of inequalities present in the data.

15.8 KEY WORDS AND SYMBOLS

Coefficient of Variation : Standard deviation divided by arithmetic mean expressed **as** a percentage.

Lorenz Curve : **A** double cumulative percentage graph used in determining the extent of inequalities of items.

Root Mean Square Deviation : The square root of the mean of the squares of deviation ***of*** items from central tendency.

Standard Deviation : The root mean square deviation about arithmetic mean.

List of Symbols

Coefficient of Variation C.V.

Combined Standard Deviation σ_{12}, σ_c, SD_c

Difference between Combined Mean and given Mean $\bar{X}_{12} - \bar{X}_1$, d_1, etc.

Standard Deviation S.D., σ, s. Generally σ is used to denote population S.D. and 's' the sample S.D.

15.9 ANSWERS TO CHECK YOUR PROGRESS

A) 3) 49.1

4) 156.37

5) i) True ii) False iii) True iv) False v) False

B) 1) i) True ii) False iii) True iv) False v) False

2) i) less ii) 68 iii) 6.4 iv) non v) 0

3) a) ii b) iv c) i d) i e) i f) iv

15.10 TERMINAL QUESTIONS/EXERCISES

Questions

1) What is standard deviation? Explain its superiority over other measures of dispersion.

2) What is coefficient of variation? What is its role as a measure of variation? How does it differ from variance.

3) Define various measures of dispersion and explain their relative merits and limitations.

Exercises

1) The students of the B.Com. class of a college have obtained the following marks in statistics out of 100 marks. Calculate the standard deviation of marks obtained

Student	:	A	B	C	D	E	F	G	H	I	J
Marks	:	5	10	20	25	40	42	45	48	70	80

(Answer : 23.06)

2) Calculate standard deviation from the following data :

Mid-points	Frequency
1	2
2	60
3	101
4	152
5	205
6	155
7	79
8	40
9	1

(Answer : = 1.57)

3) Compute standard deviation for the following data which relate to the profits of 100 companies:

Profit (Rs. in lakhs)	No. of Companies
8–10	8
10–12	12
12–14	20
14–16	30
16–18	20
18–20	10

(Answer : $\sigma = 2.77$)

4) An analysis of production rejects resulted in the following figures. Calculate mean and standard deviation..

No. of Rejects per Operator	No. of Operators
21–25	5
26–30	15
31–35	28
36–40	42
41–45	15
46–50	12
51–55	3

(Answer : $\bar{X} = 36.96$; $\sigma = 6.735$)

5) Two samples of size 40 and 50 have the same mean 53 but different standard deviations 19 and 8 respectively. Find the standard deviation of the combined sample of size 90.

(Answer : $\sigma_{12} = 14$)

6) Find the standard deviation and the coefficient of variation from the following data:

Marks	No. of Students
Less than 10	12
Less than 20	30
Less than 30	65
Less than 40	107
Less than 50	202
Less than 60	222
Less than 70	230

(Answer : $\sigma = 13.9$, C.V. = 37.3%)

7) You are given the data pertaining to kilowatt hours of electricity consumed by 100 persons in a certain city:

Consumption K. Watt. Hours	No. of Users
0 but less than 10	6
10 but less than 20	25
20 but less than 30	36
30 but less than 40	20
40 but less than 50	13

Calculate i) mean,. ii) standard deviation, and iii) range within which middle 50% of the consumers fall.

(Answer : i) 25.9 ii) 10.96 iii) 34 to 17.6) .

8) In a small town, a survey was conducted in respect of profits made by retail shops. The following results were obtained :

Profit (+)/Loss (−) (In '000 Rs.)	No. of Shops
−4 to −3	4
−3 to −2	10
−2 to −1	22
−1 to 0	28
0 to 1	38
1 to 2	56
2 to 3	40
3 to 4	24
4 to 5	18
5 to 6	10

Calculate i) the average profit made by a retail shop, ii) total profit made by all shops, and iii) the coefficient of variation of earnings.

(Answer : i) 1348 ii) 3,37,000 iii) 152.8%)

9) A factory produces two types of electric lamps A and B. In an experiment relating to their life, the following results were obtained :

Length of Life (in hours)	No. of Lamps A	No. of Lamps B
500–700	5	4
700–900	11	30
900–1100	26	12
1100–1300	10	9
1300–1500	8	6

Compare the variability of the life of the two varieties using coefficient of variation.

(Answer : C.V. (A) = 21.64%, C.V. (B) = 23.41%)

10) In two factories A and B, engaged in the same activity, the average weekly wage and standard deviation are as follows:

Factory	Average Weekly Wages (Rs.)	S.D. of Wages (Rs.)	No. of Wage Earners
A	460	50	100
B	490	40	80

i) Which factory pays larger amount as weekly wages?

ii) Which factory shows greater variability in the distribution of wages?

iii) What is the mean and standard deviation of all the workers in these two factories taken together.

Answer : i) Factory A

ii) C.V. (A) = 10.87%, C.V. (B) = 8.16%

iii) $\overline{X}_{12}$ = Rs. 473.33, σ_{12} = 49.19

11) Draw Lorenz Curves from the following data :

Wages (Rs.)	300–400	400-500	500–600	600–700	700–800
No. of Workers					
Factory A	40	30	40	60	40
Factory B	300	200	180	220	100

12) The arithmetic mean and standard deviation of **20** items were found as **20** and 5 respectively. But while calculating an item **13** was misread as **30.** Find correct arithmetic mean and standard deviation.

(Answer : AM = **19.15; σ = 4.66)**

Note : These questions and exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the University. These are for your practice only.. .

UNIT 16 MEASURES OF SKEWNESS

Structure

16.0 OBJECTIVES

After studying this unit, you should be able to :

- distinguish between skewness and dispersion
- differentiate between symmetrical, positively skewed and negatively skewed data
- calculate skewness by different methods
- decide which of the methods of computing is suitable in a given situation
- appreciate the role of normal curve in the analysis of data and discuss its properties.

6 1 INTRODUCTION

As you know, to analyse any numerical data there are three main characteristics : **1)** central tendency **i.e., a** value around which many other items of the data congregate, 2) dispersion i.e., how much the items deviate from central tendency, and **3)** skewness i.e., how the items are distributed about the central tendency. In this unit, you will learn about the third characteristic i.e. skewness.

In Unit **10** to **13** you have studied the measures of central tendency viz., arithmetic mean, median, mode geometric mean, harmonic mean and moving average. In units 14 **and 15** you have studied the measures of dispersion viz. range, quartile deviation, mean deviation, standard deviation and Lorenz curve. In this unit you will learn about third characteristic i.e. skewness. You will study the meaning, purpose and methods of computing skewness. You will also study the role and properties of normal curve in analysis of data. In fact, there is one more characteristic called kurtosis i.e., concentration of frequencies in the central part of the data, which is not within the scope of this course.

16.2 MEANING OF SKEWNESS

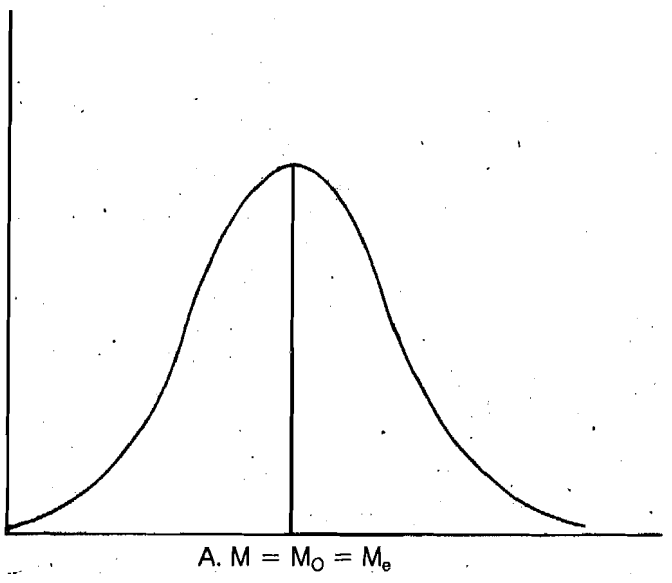
A frequency distribution is said to be 'symmetrical', if the frequencies are symmetrically distributed about central value, i.e., when values of the variable which are at an equal distance from middle have equal frequencies. Study the following two sets of distributions.

A) **X** : **10** **15** **20** **25** **30**
f : **5** 8 **26** **8** **5**
Here **X** = 20 is the group of middle items.

B) **X** **5–9** **9–13** **13–17** **17–21** **21–24**
f **7** **18** **25** **18** **7**
Here middle group is 13–17.

You can easily understand that they are symmetrical distributions. You should also note (can verify by calculation) that for each set the values of mean, median and mode are the same values. In fact, for any symmetrical distribution in which frequencies steadily rise and then steadily fall (i.e., bell shaped), mean, median and mode are equal. Study Figure 16.1 for the shape of such data on graph paper.

Figure 16.1



Symmetrical **Distribution**

If the graph of a perfectly symmetrical data is folded at the line passing through mean, one side of the curve perfectly coincides with the others side. You can say one side is the mirror image of the other side.

In general, however, frequency distributions are not perfectly symmetrical; some may be slightly asymmetrical and some others may be highly asymmetrical. Consider the following two asymmetrical (of Skewed) distributions :

A)

X	:	5–9	9–13	13–17	17–21	21–24
f	:	7	18	25	15	7

B)

X	5–9	9–13	13–17	17–21	21–24
f	7	28	15	10	2

Here the frequencies are not symmetrically distributed about the middle. In distribution A the extent of asymmetry is small while in distribution B it is comparatively larger.

The word 'skewness' is used to denote the 'extent of asymmetry' in the data. When the frequency distribution is not symmetrical, it is said to be ''skewed''. The word 'skewness' literally denoted 'asymmetry', or 'lack of symmetry' and the word 'skewed' denoted 'asymmetrical'. A symmetrical distribution has therefore zero skewness.

A distribution can be symmetrical even if frequencies first steadily fall and then steadily rise. Consider the following distributions:

Size of Items	:	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency	:	40	27	15	10	15	27	40

This is also a case of symmetrical distribution. But in this case there will be two values of the mode and both of them will be different from arithmetic mean and median which will be in the middle group, You may notice in such symmetrical distributions, which are called **bimodel** or u-shaped, only mean and median are equal. Look at figure 16.2 and study the shape of such data on graph paper.

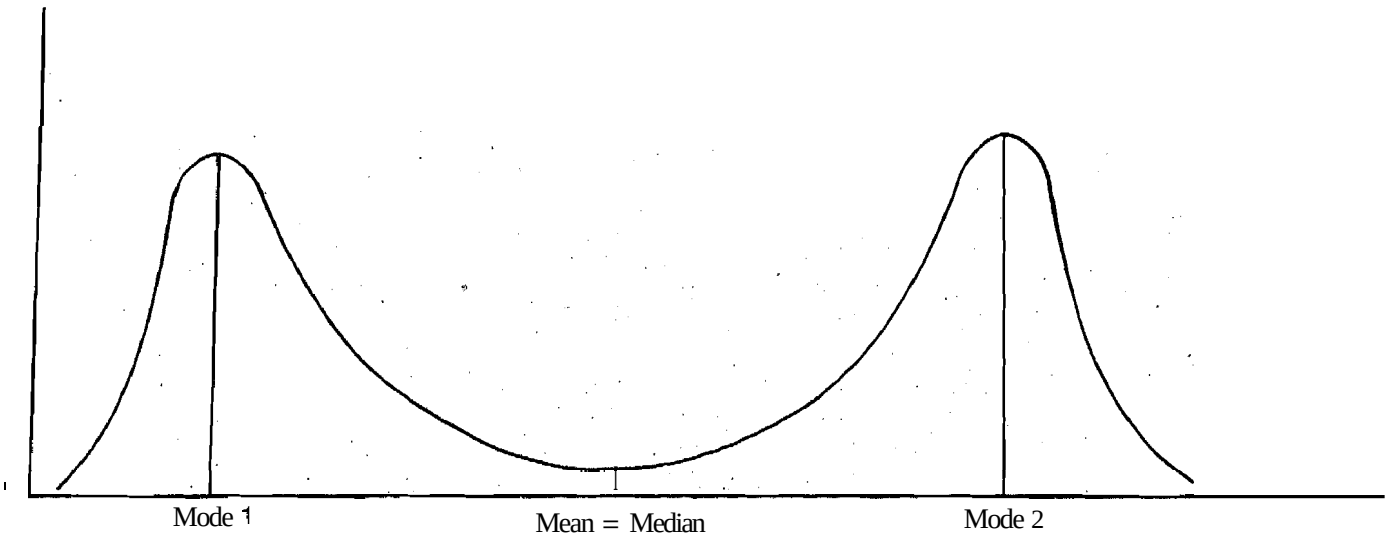


Figure 16.2 Bimodal or **U** Shaped Distribution

A bimodel distribution can also be a skewed distribution as in the following example :

Size of Items.	:	10–15	15–20	20–25	25–30	30-35	**35-40**	40–45
Frequency	:	27	18	**10**	**5**	17	**17**	**30**

Here also the distribution of items around the middle group or central value is not same on both sides. Thus, we can also say that study of skewness is the study of distribution of items around the central tendency.

Analysing the skewness of data serves the following main purposes :

1) It helps in finding out the nature and the degree of concentration — whether it is in higher or the lower values.

2) The empirical relationship between mean, median and mode i.e., $M_o = 3 M_d - 2 \bar{X}$, is based on a moderately skewed distribution. The measure of skewness will reveal to what extent such empirical relationship holds goods.

3) It helps in knowing if the distribution is normal or not. You will learn about normal distribution later in this unit.

16.3 POSITIVE AND NEGATIVE SKEWNESS

Wherever data is skewed there can be two possibilities : 1) the skewness may be positive or 2) it may be negative. In a bell shaped data or a unimodel data, which is also most common in nature, it is quite easy to understand the concept of positive and nagative skewness i.e., direction of the skewness. Mode plays an important role in this connection. The spread of the data on either side of the mode helps in deciding the direction of skewness. Consider the two sets of data given below:

A) Size of Items	:	2–4	46	6–8	8–10	10–12	12–14	14-16
Frequency	:	5	12	27	10	8	3	1
B) Size of Items	:	10–15	15–20	20–25	25–30	30–35	35–40	40–45
Frequency	:	2	5	12	18	30	21	6

Like in Set B if there is a longer tail towards the lower value or left hand side i.e., larger spread on the lower side, of mode, the skewness is negative or left handed. In such a case Mean < Median < Mode. Look at Figure 16.3 to understand the data on graph paper.

Figure 16.3

Negatively Skewed Distribution

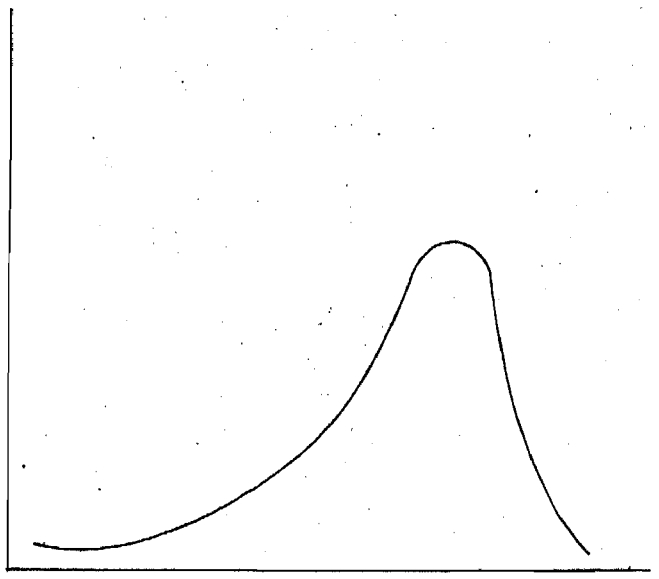
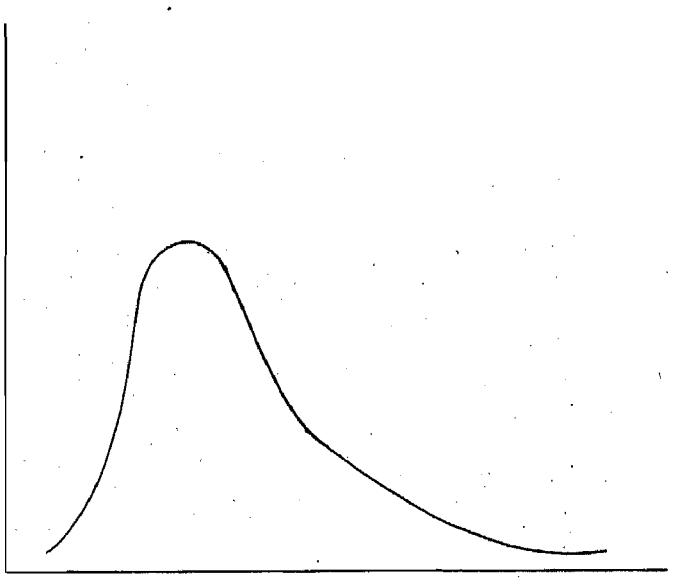


Figure 16.4

Positively Skewed Distribution



As in the case of Set A if there is a longer tail of the distribution towards the higher values or right hand side i.e., larger spread on higher side of mode, the skewness is positive or right handed. In this case Mean > Median > Mode. Shape of such a data on graph paper would be as shown in Figure 16.4.

Such data is termed 'elongated bell shaped data'. The case of extreme positive skewness would arise when frequencies are highest in the lowest values and then they steadily fall as the values increase. Similarly, the extreme negative skewness would arise when frequencies are lowest in the lower values and they steadily increase as values increase. the highest frequency representing the highest values: Such data is called 'J' shaped data. Consider the following two sets of data :

A)	Size of Items	:	10-12	12–14	14–16	1618	18–20
	Frequency	:	27	20	12	6	3
B)	Size of Items	:	10-12	12–14	14–18	1618	18–20
	Frequency	:	3	6	12	20	27

Set A shows very high positive skewness and Set B shows high negative skewness. Their shape on graph paper will be as shown in Figure 16.5A and 16.5B.

Figure 16.5A

J Shaped Positively Skewed Distribution

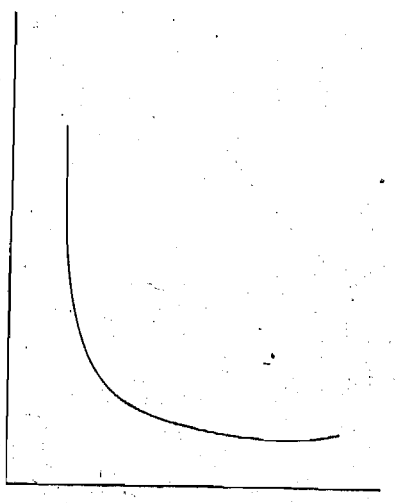
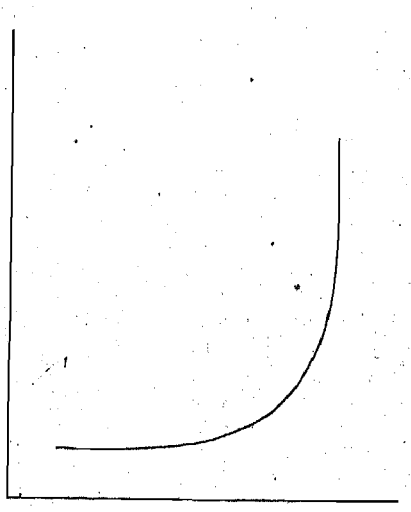


Figure 16.5B

J Shaped Negatively Slewed Distribution



Note : For the data to be skewed or symmetrical, deviations from central tendency must exist in the data.

Check Your Progress A

1) Distinguish between symmetrical data and skewed data.

.....

.....

.....

.....

2) Differentiate between positive skewness and negative skewness.

.....

.....

.....

.....

3) Distinguish between high skewness and moderate skewness.

.....

.....

.....

4) Differentiate between bell shaped and U-shaped data.

.....

.....

.....

5) State whether the following statements are True or False

i) All distributions can be classified as negative or positive skewed.

ii) Two halves of a symmetrical distribution are mirror images of each other.

iii) The sum of positive and negative deviations from median is always equal to zero in a symmetrical distribution.

iv) J-shaped distribution indicates moderate skewness.

v) It is possible that for some data Arithmetic Mean = Median = Mode, still. it is not perfectly symmetrical.

vi) Positive skewness implies that mean value is less than mode.'

vii) Median can never be equal to mean in a skewed distribution.

viii) Greater the difference 'between mean and mode, the more skewed is the distribution.

ix) U-shaped data has two modes.

x) A longer tail to right means data is negatively skewed.

xi) U-shaped distributions are always symmetrical,

xii) Highly skewed data is always positively skewed.

6) Comment on the nature of the following distribution:

i) 14, 14, 14, 14, 14

ii) 11, 12, 14, 16, 17

iii) 1, 3, 6, 18, 42

16.4 DIFFERENCE BETWEEN DISPERSION AND SKEWNESS

It has been explained in Units 14 and 15 that dispersion relates to the scatteredness or spread or the deviation of the items of a series from its central value. You also know that the measure of dispersion shows the degree of the scatteredness or average of deviation of the items of the central tendency. On the other hand, skewness relates to the departure of the items of a series from symmetry and the measure of skewness shows the degree of imbalance in the distribution of items around the central tendency. The distinguishing features are tabulated below :

Aspect	Dispersion	Skewness
1) Measureof	scatter of individual values	departure from symmetry of distribution
	how much it can deviate from central tendency	in what manner items are distributed about central tendency
2) Judges the extent of	representativeness of any of the three averages : Mean, Median, and Mode	difference between any two of the three averages : Mean, Median and Mode
3) For a symmetrical distribution	may have any value	zero
4) Useful to find	variability in data	concentration in higher or lower values

16.5 TESTS OF SKEWNESS

How can we say that a particular distribution is skewed or not? We can say skewness is present in a distribution if it has the following features:

1) Mean, median and mode should not coincide.
2) The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
3) Frequencies and their spread on either side of the mode are not equal.
4) Quartiles are not equidistant from the median i.e.. $(Q_3 - M_d)$ is not equal to $(M_d - Q_1)$.
5) When the observations in the series are plotted on a graph paper, they do not yield a symmetrical curve. This means when the graph is divided vertically through the median or mean and folded, the two halves of the curve do not coincide in a perfect manner.

16.6 MEASURES OF SKEWNESS

To study the extent of asymmetry and direction in a series, various measures of skewness are employed. These measures of skewness can be both absolute or relative.

Absolute Measures of Skewness.

Absolute measures tell us the extent of asymmetry and whether it is positive or negative.

The first absolute measure of skewness is based on the difference between mean and mode or mean and median. Symbolically i) Absolute Sk = Mean − Mode or ii) Absolute Sk = Mean − Median. If the value of mean is greater than the mode or median, skewness is positive, otherwise it is negative. It may be noted that for a positively skewed distribution; the value of the mean is the greatest and the value of mode is the least of the three measures. Likewise, for a negatively skewed distribution, mode has the maximum value and mean has the least value. In both the cases median is in between, the mean and mode.

.The second measure of skewness, based on quartiles depends upon the fact that normally for a symmetrical distribution Q_1 and Q_3 are equidistant from the median, i.e., $Q_3 - M_d = M_d - Q_1$. But if a distribution is asymmetrical, then one quartile lying on the longer tail side will be farther from the median than the other quartile. In such a case absolute measure of skewness can be, measured by the following formula:

Absolute Skewness $= (Q_3 - M_d) - (M_d - Q_1) = Q_3 + Q_1 - 2M_d$. The formula shows if $(Q_3 - M_d)$ is greater than $(M_d - Q_1)$, skewness is positive otherwise it is negative. This is true becasue $Q_3 - M_e > M_e - Q_1$ implies that the difference'between Q_3 and M_e is greater than the difference between M_e and Q_1. This in turn means that there is a longer tail on the Q, side or on right hand side i.e., skewness is right handed or positive.

Relative Measures of Skewness

In order to make comparison between the skewness in two or more distributions, coefficient of skewness is computed for the given series or distributions. The following are the two important methods of measuring relative skewness:

1) **Karl Pearson's** Coefficient of Skewness. This method is most frequently used for measuring skewness and is based on first absolute measure. The formula for measuring skewness is as follows:

$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$ or $\frac{\bar{X} - M_o}{\sigma}$ i.e., first absolute measure of skewness is divided by standard deviation. Thus, this value will be free of units of the data. The value of this coefficient would be zero in a symmetrical distribution. If mean is greater than mode, coefficient of skewness would be positive otherwise negative. In practice, the value of this coefficient usually lies between $\pm$ 3.

If the mode is ill-defined, then using the approximate relationship:

Mode = 3 Median − 2 Mean

The above formula reduces to

$Sk_p = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$ or $\frac{3(\bar{X} - M_d)}{\sigma}$

Note: As mean and standard deviations are calculated by using values of all the items of the data, Karl Pearson's method measures skewness utilising all the items of the data,

To understand the application of Karl Pearson method clearly, let us consider some illustrations.

Illustration 1

From the marks secured by 120 students in Sections A and B of a class of 120 students, the following measures are obtained:

Section A : $\bar{X} = 46.83$, $\sigma = 14.8$, Mode = 51.67
Section B : $\bar{X} = 47.83$, $\sigma = 14.8$, Mode = 47.07

Determine which distribution of marks is more skewed.

Solution

Section A

$$Sk_p = \frac{\bar{X} - \text{Mode}}{\sigma}$$

$$= \frac{46.83 - 51.67}{14.8}$$

$$= \frac{-4.84}{14.8}$$

$$= -0.327$$

Section B

$$Sk_p = \frac{\overline{X} - Mode}{\sigma}$$

$$= \frac{47.83 - 47.07}{14.8}$$

$$= \frac{0.76}{14.8}$$

$$= 0.051$$

Hence the distribution of marks in Section A is more skewed. The skewness for Section A is negative, while that of B is positive.

Illustration 2

Following statistical measures are given for a data set. Find out the value of standard deviation.

Coefficient of skewness is −0.375, Mean is 62 and Median is 6.

Solution

The coefficient of skewness that depends upon Mean, Median and Standard Deviation is Karl Pearson's coefficient of skewness.

$$Sk_p = \frac{3(\overline{X} - M_d)}{\sigma}$$, substituting the given values

$$-0.375 = \frac{3(62 - 65)}{\sigma}$$

$$\therefore \sigma = \frac{-3(62 - 65)}{0.375}$$

$$= \frac{-3(-3)}{0.375}$$

$$= \frac{9}{0.375}$$

$$= 24$$

Standard Deviation is 24.

2) **Bowley's Coefficient of Skewness:** This method is based on quartiles, i.e., second absolute measure of skewness. The formula for calculating skewness is :

$$Sk_B = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}$$

$$= \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

This method is particularly useful in case of open end distributions and where extreme values are present or when class-intervals are unequal. Skewness should be measured by this Bowley's method also when positional measures are called for.

If the value of this coefficient is zero, it is a symmetrical distribution. For positive value, it is a positively skewed distribution and for a negative value it is a negatively skewed distribution. The range of variation under this formula is ± 1. But the main drawback of this measure is that it is based on central 50% of the data and it ignores the remaining 50% of the data i.e., 25% of the data below Q_1, and 25% of the data above Q_3. To understand the application of Bowley's method clearly, study Illustrations 3 and 4.

Illustration 3

For a given data, $Q_1 = 58$, $M_d = 59$ and $Q_3 = 61$. Find coefficient of skewness.

Solution

$$Sk_B = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

$$= \frac{61+58-2\times59}{61-58}$$

$$= \frac{1}{3}$$

$$= 0.33.$$

Illustration 4

In a frequency distribution, the coefficient of skewness based upon quartiles is 0.6 If the sum of the upper and lower quartiles is 100 and the median is 38, find the value of the upper quartile.

Solution

Bowley's coefficient of skewness based on quartiles is given by:

$$Sk_B = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

Substituting the given values

$$0.6 = \frac{100 - 2 \times 38}{Q_3 - Q_1}$$

or $Q_3 - Q_1 = \frac{100 - 76}{0.6} = 40$ ... (i)

Also it is given, $Q_3 + Q_1 = 100$ ... (ii)

Adding (i) and (ii), we get

$$(Q_3 + Q_1) + (Q_3 - Q_1) = 100 + 40$$

$$2\,Q_3 = 140$$

$$Q_3 = \frac{140}{2} = 70$$

Hence the upper quartile is 70.

16.7 SOME ILLUSTRATIONS

Illustration 5

Calculate appropriate measure of skewness from the following data.

Payment of Commission	No. of Salesmen
1000 – 1200	7
1200 – 1400	15
1400 – 1600	18
1600 – 1800	20
1800 – 2000	25
2000 – 2200	10
2200 – 2400	5

Solution

Since the given distribution is not openended and also the mode can be determined, it is appropriate to apply Karl Pearson formula as given below :

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$$

Payment of Commission (Rs.)	Mid-point (x)	No. of Salesmen (f)	$d' = \frac{X - 1700}{200}$	fd′	fd′²
1000 – 1200	1100	7	−3	−21	63
1200 – 1400	1300	15	−2	−30	60
1400 – 1600	1500	18	−1	−18	18
1600 – 1800	1700	20	0	0	0
1800 – 2000	1900	25	+1	25	25
2000 – 2200	2100	10	+2	20	40
2200 – 2400	2300	5	+3	15	45
Total		n = 100		$\Sigma fd' = -9$	$\Sigma fd'^2 = 251$

$$\bar{X} = A + \frac{\Sigma fd'}{n} \times i = 1700 \frac{-9}{100} \times 200$$

$$= 1700 - 18 = 1682$$

$$\text{Mode} = L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

= Clearly the modal group is 1800 – 2000. Substituting the values.

$$= 1800 + \frac{25-20}{(25-20) + (25-10)} \times 200$$

$$= 1800 + \frac{5}{20} \times 200 = 1800 + 50 = 1850$$

Now calculating the standard deviation.

$$\sigma = i \times \sqrt{\frac{\Sigma fd'^2}{n} - \left(\frac{\Sigma fd'}{n}\right)^2}$$

$$= 200 \times \sqrt{\frac{251}{100} - \left(\frac{-9}{100}\right)^2}$$

$$= 200 \times \sqrt{2.51 - 0.0081}$$

$$= 1.582 \times 200 = 316.4$$

Now coefficient of skewness, $Sk_P = \frac{1682 - 1850}{316.4}$

$$= -0.531$$

This value of coefficient of skewness indicates that the distribution is negatively skewed and hence there is a greater concentration towards the higher commission.

Illustration 6

Calculate the coefficient of skewness based on mean and median from the following distribution:

Class Interval	Frequency
0 – 10	6
10 – 20	12
20 – 30	22
30 – 40	48
40 – 50	56
50 – 60	32
60 – 70	18
70 – 80	6

Solution

Calculationslor Mean, Median and S.D.

Class Interval	Mid-point (X)	$d' = \frac{x - 35}{10}$	I	Id'	fd'²	Cum. Frequ.
0-10	5	−3	6	−18	54	6
10-20	15	−2	12	−24	48	18
20-30	25	−1	22	−22	22	40
30-40	35	0	48	0	0	88
40-50	45	1	56	56	56	144
50-60	55	2	32	64	128	176
60-70	65	3	18	54	162	194
70-80	75	4	6	24	96	200
Total			200,	134	566	

$$\overline{X} = A + \frac{\sum fd'}{n} \times i = 35 + \frac{134}{200} \times 10 = 41.7$$

Median has $\frac{N}{2}$ observations or 100 observations below it.

Therefore, median lies in the 40–50 class.

$$M_d = l + \frac{\frac{N}{2} - c}{f} \times i$$

$$= 40 + \frac{100 - 85}{56} \times 10$$

$$= 42.14$$

$$\sigma = i \times \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2}$$

$$= 10 \times \sqrt{\frac{566}{200} - \left(\frac{134}{200}\right)^2}$$

$$= 10 \times \sqrt{2.83 - 0.449}$$

$$= 1.543 \times 10 = 15.43$$

Karl Pearson's coefficient of skewness based on mean and median is given by:

$$Sk_P = \frac{3(\overline{X} - M_d)}{\sigma}$$

$$= \frac{3(41.7 - 42.14)}{15.43}$$

$$= \frac{3(-0.44)}{15.43}$$

$$= -0.085$$

Hence, the distribution is negatively skewed with very low degree of skewness.

Illustration 7

Calculate the coefficient of skewness based on quartiles from the following data:

Monthly Salary	No. of Employees
1000 – 1200	5
1200 – 1400	14
1400 – 1600	23

1600 – 1800	50
1800 – 2000	52
2000 – 2200	25
2200 – 2400	22
2400 – 2600	7
2600 – 2800	2
	200

Solution

Computation of Quartiles

Monthly Salary	Frequency	Comulative Frequency
1000 - 1200	5	5
12W - 1400	14	19
1400 – 1600	23	42
1600 – 1800	50	92
1800 – 2000	52	144
2000 – 2200	25	169
2200 – 2400	22	191
2400 – 2600	7	198
2600 – 2800	2	200

Q_1 has $\frac{N}{4}$ observations or 50 observations below it. It lies in the class 1600 – 1800.

$$Q_1 = l + \frac{\frac{N}{4} - c}{f} \times i$$

$$= 1600 + \frac{50 - 42}{50} \times 200$$

$$= 1632$$

Q_2 (= Median) has $\frac{N}{2}$ observations or 100 observations below it. So it lies in the class 1800 – 2000.

$$M_d = 1800 + \frac{100 - 92}{52} \times 200$$

$$= 1800 + 30.77 = 1830.77$$

Q_3 has $\frac{3N}{4}$ observations or 150 observations below it. So it lies in the class 2000 – 2200.

$$Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times i$$

$$= 2000 + \frac{150 - 144}{25} \times 200$$

$$= 2000 + 48 = 2048$$

$$\text{Coefficient of Sk} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

$$= \frac{2048 + 1632 - (2 \times 1830.77)}{2048 - 1632}$$

$$= \frac{18.46}{416} = 0.044$$

Illustration 8

The following table gives the distribution of monthly income of 500 workers in a factory:

Monthly Income (Rs.)	No. of Employees
Below Rs. 1000	10
1000 – 1500	25
1500 – 2000	145
2000 – 2500	220
2500 – 3000	70
3000 and above	30

i) Obtain the limits of income of central 50 per cent of the observed employees

ii) Calculate Bowley's coefficient of skewness.

Solution

i) For obtaining the limits of central 50% of the workers, calculate Q_1 and Q_3.

Calculations for Quartiles

Monthly Income (Rs.)		Comulative Frequency
Below Rs. 1000	10	10
1000 – 1500	25	35
1500 – 2000	145	180
2000 – 2500	220	400
2500 – 3000	70	470
3000 and above	30	500

Q_1 has $\frac{N}{4}$ or 125 observations below it. So it lies in the class 1500 – 2000.

$$Q_1 = l + \frac{\frac{N}{4} - c}{f} \times i$$

$$= 1500 + \frac{125 - 35}{145} \times 500$$

$$= 1500 + 3103 = 1810.3$$

Q_3 has $\frac{3N}{4}$ or 375 observations below it. So it lies in the class 2000 – 2500.

$$Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times i$$

$$= 2000 + \frac{375 - 180}{220} \times 500$$

$$= 2000 + 443.18 = 2443.18$$

Hence the incomes of central 50% of workers lies between Rs. 1'810.3 and Rs. 2443.18.

ii) Bowley's coefficient of skewness is given by:

$$Sk_B = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

M_d had $\frac{N}{2}$ or 250 observations below it. So it lies in 2000 – 2500 class.

$$M_d = l + \frac{\frac{N}{2} - c}{f} \times i$$

$$= 2000 + \frac{250 - 180}{220} \times 500$$

$$= 2000 + 159.1 = 2159.1$$

$$\therefore Sk_B = \frac{2443.18 + 1810.3 - (2 \times 2159.1)}{2443.18 - 1810.3}$$

$$= \frac{-64.74}{632.88} = -0.102$$

The negative coefficient (–0.102) indicates that distance between Q_3 and M_d is smaller than that between M_d and Q_1 i.e., the distribution is skewed to the left.

Illustration 9

Calculate Karl Pearson's coefficient of skewness from the following data:

Incomes (Rs. per day)	No. of Shops
Above 0	150
Above 100	140
Above 200	100
Above 300	80
Above 400	80
Above 500	70
Above 600	30
Above 700	14
Above 800	0

Solution

Converting the cumulative frequency distributions to ordinary frequency distribution, we have:

Income (Rs. per day)	No. of Shops
0 – 100	10
100 – 200	40
200 – 300	20
300 – 400	0
400 – 500	10
500 – 600	40
600 – 700	16
700 – 800	14

As it is a u-shaped distribution, skewness will be calculated by using Mean and Median.

Calculations for Coefficient of Skewness

Income (Rs. per day)	Mid-point X	f	$d' = \frac{X - 350}{100}$	fd'	fd'^2	Cum. Freq.
0 – 100	50	10	–3	–30	90	10
100 – 200	150	40	–2	–80	160	50
200 – 300	250	20	–1	–20	20	70
300 – 400	350	0	0	0	0	70
400 – 500	450	10	1	10	10	80
500 – 600	550	40	2	80	160	120
600 – 700	650	16	3	48	144	136
700 – 800	750	14	4	56	224	150
Total		150		64	808	

Calculation of Mean

$$\bar{X} = A + \frac{\sum fd'}{N} \times i$$

$$= 350 + \frac{64}{150} \times 100$$

$$= 350 + 42.67 = 392.67$$

Calculation of Median

M_d lies $\frac{n}{2}$ or 75 observations below it. So it lies in the class 400 – 500.

$$M_d = l + \frac{\frac{n}{2} - c}{f} \times i$$

$$= 400 + \frac{75 - 70}{10} \times 100 = 450$$

Calculation of Standard Deviation

$$\sigma = i \times \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2}$$

$$= 100 \times \sqrt{\frac{808}{150} - \left(\frac{64}{150}\right)^2}$$

$$= 100 \times \sqrt{5.205}$$

$$= 2.281 \times 100 = 228.1$$

$$Sk_P = \frac{3(\bar{X} - M_d)}{\sigma}$$

$$= \frac{3(392.67 - 450)}{228.1}$$

$$= \frac{-171.99}{228.1}$$

$$= 0.754$$

Illustration 10

Find standard deviation, rnode and median when mean = 50, coefficient of variation = 40%, Skewness = −0.4.

Solution

Substituting the values of mean and C.V. in the formula

$$C.V. = \frac{S.D.}{Mean} \times 100, \text{ we get}$$

$$40 = \frac{S.D.}{50} \times 100$$

$$S.D. = \frac{50 \times 40}{100} = 20$$

Again using Karl Pearson's formula

$$Sk_P = \frac{Mean - Mode}{S.D.}$$

$$-0.4 = \frac{50 - \text{Mode}}{20}$$

$$\text{Mode} = 50 + 20 \times 0.4$$

$$= 58$$

Using the empirical relationship, we obtain

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$

$$50 - 58 = 3\,(50 - \text{Median})$$

$$-8 = 150 - 3\,\text{Median}$$

$$3\,\text{Median} = 150 + 8$$

$$\text{Median} = \frac{158}{3} = 52.67$$

Illustration 11

Find the appropriate measure of skewness from the following data :

Sales (Rs. in Lakhs)	No. of Companies	Cumulative Frequency
Below 50	8	8
50 – 60	12	20
60 – 80	20	40
80 – 100	25	65
100 and Above	15	80

Solution

Here class intervals are unequal and open. So the appropriate method of determining skewness is Bowley's method.

$$Sk_B = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

Now Q_1 has $\frac{N}{4}$ observations or 20 observations below it. So it lies in the class 50 – 60

$$Q_1 = l + \frac{\frac{n}{4} - c}{f} \times i$$

$$= 50 + \frac{20 - 8}{12} \times 10 = 60$$

Q_2 (= median) has $\frac{N}{2}$ observations or $\frac{80}{2}$ or 40 observations below it. So it lies in the class 60 – 80.

$$M_d = l + \frac{\frac{n}{2} - c}{f} \times i$$

$$= 60 + \frac{40 - 20}{20} \times 20 = 80$$

Q_3 has $\frac{3n}{4}$ or 60 observations below it. So it lies in the class 80 – 100

$$Q_3 = l + \frac{\frac{3n}{4} - c}{f} \times i$$

$$= 80 + \frac{60 - 40}{25} \times 20$$

$$= 96$$

$$Sk_B = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

$$= \frac{96 + 60 - (2 \times 80)}{96 - 60}$$

$$= -0.11$$

This value of coefficient of skewness indicates that the distribution is slightly skewed to the left and, therefore, there is a greater concentration of the sales at the higher values than the lower values of the distribution.

Illustration 12

The following facts were gathered from a firm before and after an industrial dispute:

	Before Dispute	After Dispute
Mean Wages (Rs.)	850	900
Median Wages (Rs.)	820	800
Modal Wages (Rs.)	760	600
Quartiles (Rs.)	750 & 920	750 & 950
S.D. (Rs.)	30	110
Number Employed	600	550

By making use of the above data, compare the position of the firm before and after the dispute as fully as possible.

Solution

a) Number of workers has decreased by 50, from 600 to 550 as a result of the dispute.

b) Although the mean wage has slightly increased, the firm saves Rs. 15,000 (after dispute) in respect of the monthly salary bill:

Total Wages before Dispute (600 × 850)	= Rs. 5,10,000
Total Wages after Dispute (550 × 900)	= Rs. 4,95,000
Difference	15,000

c) The median and modal wages have decreased. Before the dispute, 50% of the workers used to get Rs. 820 and above. But after the dispute, workers in this category are less than 50%. Similarly, most of the workers are being paid around Rs. 600 (after dispute) as against Rs. 760 (before dispute).

d) The first quartile Q_1 has not changed. The second quartile Q_2 (i.e., Median) has decreased slightly, but the third quartile Q_3 has increased. The significance of the information is as shown below :

Wages (Rs.)

Category of Workers	Before Dispute	After Dispute
A. Lowest Paid 25%	upto 750	Upto 750
B. Next Higher Group of 25%	750 – 820	750 – 800
C. Next Higher Group of 25%	820 – 920	800 – 950
D. Highest Paid 25%	Above 920	Above 950

Category (A) workers are not affected. The next higher category (B) workers are now confined to a narrower range of salary. But the highest paid categories **(C)** and (D) are now generally paid more after the dispute.

e) Standard deviation has increased from Rs. 30 to Rs. 110 implying thereby that the variability in individual wages has increased after dispute. For proper comparison, we have :

C.V. (before dispute) $= \frac{30}{850} \times 100 = 3.53\%$

C.V. (after dispute) $= \frac{110}{900} \times 100 = 12.2\%$

The variability relative to mean has also increased.

f) Measure of skewness are:

	Before Dispute	After Dispute
Pearson's Measure	$\frac{850 - 760}{30} = 3$	$\frac{900 - 600}{100} = 2.73$
Bowley's Measure	$\frac{920 - 2(820) + 750}{920 - 750}$	$\frac{950 - 2(800) + 750}{950 - 750}$
	$= 0.18$	$= 0.50$

Pearson's measure of skewness. after dispute has decreased while the Bowley's measure has increased, both being positive. This means that for middle 50% of workers concentration in lower wages has increased. But when we consider all the workers, then the relative concentration of frequencies on lower values side is lower.

Note: There is nothing wrong if one formula gives result indicating increase in skewness while the other gives decrease in skewness. In fact, these can be situations when one formula gives positive skewness while the other may give negative skewness. This is because Bowley's me'thod is based on only middle 50% data while Pearson's method relates to entire data.

Check Your Progress B

1) State formalas of the Karl Pearson's and the Bowley's methods of measuring skewness.

.....

.....

.....

2) What is skrwncss?

.....

.....

.....

3) Differentiate between skewness and dispersion.,

.....

.....

.....

.....

4) State whether the following statements dre True or False.

i) Skewness judges the extent of representativeness of any average.

ii) For a positively skewed distribution. concentration of frequencies is on left.

iii) Only relative value of skewness is used'for comparison even though standard deviation is the same.

iv) Skewness cannot he calculated for open end class intervals.

v) Skewness does not exist in Bimodel distribution

vi) Two distributions having different coefficient of variations so they have different skewness.

5) Fill in the blanks:

i) If the mean and the mode of a given distribution are equal then its coefficient of skewness is .....

ii) Skewness is positive when mean is ..... mode.

iii) In a symmetrical distribution the mean, median and mode are .....

iv) Median can never be equal to ..... in case of skewed distribution.

v) If the mean, mode and standard deviation of a frequency distribution arc 41, 45, and 8 respectively, then its Pearson's coefficient of skewness is .....

vi) In a perfectly symmetrical distribution, 50% items are above 60 and 75% items are below 75. Therefore M_e = ...

Q_3 = ... Q_1 = ..., coefficient or quartile deviation is....., and coefficient of skewness is .....

16.8 PROPERTIES OF NORMAL CURVE

It has been observed that frequency distribution most of the phenomena that occur in nature such as measurements of human characteristics (height, weight, IQ, etc.,), measurements relating to industrial production and agricultural production, etc. are symmetrical in nature. Normally, they all have almost a fixed rate of rise and fall of frequencies from one group to another group. Their shape is like in Figure 16.1. Statisticians have tried to express these distributions by a single mathematical formula. As this formula describes most of the distributions which occur in nature, it has been called 'Normal Curve'. At this stage, it is not necessary for you to know the exact mathematical expression that gives the normal curve. But the properties that are exhibited by that formula are very useful in the analysis of data. Following are the main properties of the normal curve:

1) It is perfectly symmetrical about the mean and is bell shaped.

2) Mean = Median = Mode

3) It has only one mode, i.e., it is unimodel.

4) The quartiles Q_1 and Q_3 are equidistant from the median or mean and are given by

Q_1 = A.M. − 0.6745 S.D.

Q_3 = A.M. + 0.6745 S.D.

QD $= \frac{5}{6}$ M.D. Approximately.

$= \frac{2}{3}$ standard deviation (approximately)

5) The mean deviation about mean is $\frac{4}{5}$ x S.D.

6) One of the most fundamental properties of the normal probability curve is the area property.

i) Mean ± 0.6745 SD covers 50% area, i.e., 25% on each side.

ii) Mean ± 2.5758 SD covers 99% area, i.e. 49.5% on each side.

iii) Mean ± 1.96 SD covers 95% area, i.e. 47.5% on each side.

iv) Mean ± 1 SD covers 68.37% area, i.e., 34.14% on each side.

v) Mean ± 2 SD covers 95.4% area, i.e., 47.7% on each side.

vi) Mean ± 3 SD covers 99.7% area, i.e., 49.85% on each side.

Let us take one example to point out the usefulness of these properties.

Illustration 13

Suppose mean height of 100 persons selected from a big group is 68 inches and standard deviation is 1.5 inches.

i) What is the range of height of middle 95% persons in the whole group?

ii) How much would be the expected value of mode, Q.D. and M.D. for the whole group?

Solution

i) Now 95% of items have values between the range Mean $\pm$ 1.96 SD. So the required range is 68 $\pm$ 1.96 $\times$ 1.5 or 65.06 inches to 70.94 inches.

ii) Mean = Mode. Therefore mode is also 68 inches

QD $= \frac{2}{3}$ SD approximately. So QD $= \frac{2}{3} \times 1.5 = 1$ inch approximately

MD $= \frac{4}{5}$ SD approximately. So MD $= \frac{4}{5} \times 1.5 = 1.2$ inch approximately

In fact normal curve is very much useful in drawing statistical inference. It is also used as a standard to find out the extent of concentration of frequencies in the central part of the given data. This is the fourth main characteristic in analysis of data, called Kurtoses, the details of which are out of scope of this course.

16.9 LET US SUM UP

The measures of central tendency and variation do not reveal all the characteristics of a data set. Two distributions may have the same mean and standard deviation, but may differ widely in the shape of their distribution. If the distribution of data is not symmetrical, it is called asymmetrical or skewed. Skewness refere to the lack of symmetry in distribution. Different methods of measuring skewness are as follows:

	Absolute Measure	Relative Measure	Limits on Range	Given by
1.	Mean − Mode	$\frac{\text{Mean} - \text{Mode}}{\text{SD}}$	± 3	Karl Pearson
2.	Mean − Median	$\frac{3(\text{Mean} - \text{Mode})}{\text{SD}}$	± 3	Karl Pearson
3.	$Q_3 + Q_1 - 2M_d$	$\frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$	± 1	Bowley

In highly skewed data highest frequency exists on one extreme of the data. A positively skewed distribution has a long tail on right hand side of the data and is also termed as right handed skew. A negatively skewed data has a long tail on left hand side of the data and is also termed as left handed skew. When the graph of a perfectly symmetrical data, bell shaped or U-shaped, folded at the line at mean, two sides of the curves perfectly coincide with one another.

Most of the data which occurs in nature resembles the normal distribution. Normal curve is a perfectly symmetrical data with bell shape. It has a fixed percentages of frequencies lying in different ranges from mean. These values of percentages help us in deciding whether the given data is normal or not.

16.10 KEY WORDS AND SYMBOLS

Bell Shaped Data : Frequencies steadily rise, reach a maximum and then steadily fall.

J Shaped Data : Start with highest and end with lowest frequency and has a steady rate of fall in between or vice-versa.

Skewness : Refers to the lack of symmetry

Symmetrical Data : When values of variable equidistant from middle have equal frequencies.

U-Shaped Data : Data has high frequencies in the beginning and end, and lowest frequencies in the middle.

List of Symbols

Coefficient of Skewness : Bowley's – Sk_B,

Coefficient of Skewness : Pearson's – Sk_P

Skewness – Absolute Measure Sk, J

16.11 ANSWERS TO CHECK YOUR PROGRESS

A) 5) i) False ii) True iii) True iv) False v) False vi) False vii) True viii) True ix) True x) False xi) False xii) False.

6) i) no variation ii) symmetrical iii) skewed

B) 4) i) True ii) True iii) True iv) False v) False vi) May or may not be true.

5) i) zero ii) greater than iii) equal iv) mean v) –0.5 vi) $M_e = 60$, $Q_3 = 75$, $Q_1 = 45$ coefficient of $QD = 0.25$, $Sk_B = 0$.

16;12 TERMINAL QUESTIONS/EXERCISES

Questions

1) Give the absolute and relative measures of skewness.

2) Central tendency, dispersion and skewness are three different measures to analyse numerical data, Comment.

Exercises

1) From the following frequency distribution of marks of students in an examination, calculate the value of Karl Pearson's coefficient of skewness:

Marks less than	10	20	30	40	50	60	70	80
No. of Students	5	15	30	50	80	100	120	125

(Answer: $\bar{X}$ 53, = 17.66, $Sk_P = 0.453$)

2) Calculate Pearson's Coefficient of Skewness from the table given below:

Life Time (In Hours)	No. of Tubes
300 – 400	14
400 – 500	46
500 – 600	58
600 – 700	76
700 – 800	68
800 – 900	62
900 – 1000	48
1000 – 1100	22
1100-1200	6

(Answer : $Sk_P = \frac{715.5 - 669.23}{190.2} = 0.243$)

3) The following data shows the daily sales at a petrol station. calculate the mean, median, standard deviation and coefficient of skewness.

Quantity sold (in Litres)	No. of Days
700 – 1000	12
1000 – 1300	18
1300 – 1600	20
1600 – 1900	25
1900 – 2200	18
2200 – 2500	5
2500 – 2800	2

(**Answer** : $\overline{X} = 1426$, $M_d = 1600 = 447.35$, SK = 1.167)

4) The following table gives the distribution of daily travelling allowance of salesmen in a company. Compute Bowley's Coefficient of Skewness and comment on its value.

Travelling Allowance (in Rs.)	No. of Salesmen
100 – 120	14
120 – 140	16
140 – 160	20
160 – 180	18
180 – 200	15
200 – 220	7
220 – 240	6
240 – 260	4

(Answer: $Sk_B = \frac{189.33 + 133.75 - (2 \times 160)}{189.33 - 133.75} = 0.145$)

5) Calculate an appropriate measure of skewness for the data given below:

Age (Years)	No. of Employees
Below 20	13
20 – 25	29
25 – 30	46
30 – 35	60
35 – 40	112
40 – 45	94
45 – 50	45
50 and above	21

(**Answer : $Sk_B = \frac{42.92 + 31.42 - (2 \times 37.77)}{42.92 - 31.42} = 0.104$**)

6) Find a suitable measure of skewness from the following distribution:

Annual Sales (Rs. in 0 0 0	:	0-20	20-50	50-100	100-250	250-500	500-1000
No. of Firms	:	20	50	69	30	22	19

(Answer: $Sk_B = \frac{203.75 + 39.95 - (2 \times 76.45)}{203.75 - 39.95} = 0.554$)

7) You are given below the details relating to the wages in respect of two factories. From this it is concluded that the skewness and variability are the same in both the factories. Point out the mistake or wrong inference in the above statement.

	Factory A (Rs.)	Factory B (Rs.)
Arithmetic Mean	50	45
Mode	45	50
Variance	100	100

8) Calculate Karl Pearson's coefficient of skewness based on the empirical relationship that exists between the central tendencies in a moderately asymmetrical distribution:

Mean = 23, Median = 24, Standard Deviation = 10.
Is this distribution negatively or positively skewed?

(Answer : = –0.3)

9) The following is the position in a factory before and after the settlement of an industrial dispute. Comment on the gains or losses from the point of view sf workers and that of management:

	Before	**After**
No. of Workers	3,000	2,900
Mean of Wages (Rs.)	220	230
Median of Wages (Rs.)	250	240
Standard Deviation	30	26

Note : These questions and exercises will help you to understand the unit better. Try to write answer for them. But do not submit your answer to the University. These are for your practice only.

SOME USEFUL BOOKS

Elhance, D.N. and Veena Elhance, 1988. *Fundamentals* of *Statistics*, Kitab Mahal : Allahabad. (Chapters 9, 10 & 18)

Gupta, C.B., An *Introduction to Statistical, Methods,* Vikas Publishing House : New Delhi. (Chapters 10, 11 & 17)

Gupta, S.P., 1989, *Elementary Statistical Methods,* Sultan Chand & Sons : New Delhi. (Chapters 8 & 9)

Sancheti, D.C., and Kapoor, V.K., 1989, *Statistics Theory Methods and Applications,* Sultan Chand & Sons : New Delhi. (Chapters 5, 7 & 16)

Shenoy, G.V., Srivastava V.K., and Sharma, S.C., 1989, *Business Statistics,* Wiley Eastern : New Delhi. (Chapters 5, 6 & 11)

Simpson, *G*, and Kafka, F. *Basic Statistics,* Oxford & IBH Publishing : New Delhi. (Chapters 13, 16 & 21)

NOTES