School of Management Studies

BCOC-134
Business Mathematics and Statistics

PART B : BUSINESS STATISTICS

UNI-VARIATE ANALYSIS

BI-VARIATE ANALYSIS

TIME-BASED DATA ANALYSIS

COURSE DESIGN AND PREPARATION TEAM

Prof. Madhu Tyagi
Former Director
SOMS, IGNOU

Dr. H. K. Dogi
University of Delhi, Delhi
Units 10 to 11

Dr. Vidya Ratan
Sriram College of Commerce
University of Delhi, Delhi

Prof. Brahmha Bhatt
School of Commerce
Gujarat University Ahemdabad

Prof. M.S. Senam Raju
SOMS, IGNOU, New Delhi
(Unit 15,16,17 & 18)

Prof. G.P. Singh (Retd.)
University of Swarashtra,
Gujarat (Units 1 to 4)

Dr. Sarabjit Singh Kaur
Dayal Singh College
University of Delhi, Delhi
Units 5 to 9

Dr. O.P. Gupta
University of Delhi, Delhi

Dr. C.R. Kothari
Rajsthan University,
Jaipur

Prof. (Mrs.) Sarla Achuthan
Gujarat University, Ahemdabad

Faculty Members
SOMS, IGNOU
Prof. N V Narasimham
Prof. Nawal Kishor
Prof. M.S.S. Raju
Dr. Sunil Kumar
Dr. Subodh Kesharwani
Dr. Rashmi Bansal
Dr. Madhulika P Sarkar
Dr. Anupriya Pandey

Course Coordinators and Editors
Prof. M.S. Senam Raju
SOMS, IGNOU, New Delhi
Dr. Anupriya Pandey
SOMS, IGNOU, New Delhi

Content Editing (Part-A)
Prof. Gopinath Pradhan (Retd.)
SOSS, IGNOU, New Delhi

Part-B adopted from ECO-07 & MCO-03

MATERIAL PRODUCTION

Mr. Y.N. Sharma
Assistant Registrar (Publication)
MPDD, IGNOU, New Delhi

Mr. Sudhir Kumar
Section Officer (Pub.)
MPDD, IGNOU, New Delhi

January, 2020

ISBN:

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by the Registrar, MPDD, IGNOU.

Laser typeset by Tessa Media & Computers, C-206, A.F.E-II, Jamia Nagar, New Delhi-110025

Printed at:

COURSE INTRODUCTION

This is one of the core courses in B.Com programme under CBCS scheme. The main objective of this course is to familiarize the students with the application of Mathematics and Statistical techniques which will facilitate in business decision making. This course consists of two parts, viz., **PART- A: Business Mathematics** comprising of 11 units and **PART-B: Business Statistics** comprising of 7 units (unit 12 to unit 18). The brief introduction of **Part-B** is as follows:

PART B : BUSINESS STATISTICS

This Part-B dealt with comprises of 7 Units (Unit 12 to Unit 18). i) Uni-variate Analysis covering measures of central tendency and measures of dispersion with brief introduction to statistics, ii) Bi-variate Analysis covering simple correlation and regression, finally, iii) Time-Based Data Analysis pertaining to index numbers and time series. The brief description of each unit is given below:

UNI-VARIATE ANALYSIS

Unit 12: Introduction to Statistics deals with meaning, definitions, functions, importance, scope and limitations of statistics.

Unit 13 : Measure of Central Tendency describes the meaning of central tendency, list out its various measures such as Arithmetic Mean, Geomantic Mean, Harmonic Mean, Median, Partition Values and Mode. It also discusses the concept, method of computation, properties, uses and limitations of these measures.

Unit 14 : Measures of Dispersion discuses the need of dispersion. It further explains the meaning computation and uses of three measures of dispersion viz., Range, Quartile Deviation, Mean Deviation, Standard Deviation, and Co efficient of Variation.

BI-VARIATE ANALYSIS

Unit 15: Simple Linier Correlation introduces the concepts of co relation, its calculations, their merits and limitations.

Unit 16: Simple Linier Regression introduces the concept of regression, with its computation and applications thereof.

TIME-BASED DATA ANALYSIS

Unit 17 : Index numbers discusses meaning, concept, uses, and issues in construction of index numbers along with the methods of constructing it.

Unit 18: Time Series Analysis explain the basic concepts, utility, components of time series, along with the methods of measurement of trend to forecast the future from the historical time series data.

UNIT 12 INTRODUCTION TO STATISTICS

Structure

12.0 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, and
- explain the usages and importance of statics in business.

12.1 INTRODUCTION

So far, we have discussed business mathematics. In this unit, we will discuss business statistics and its usage and importance in business. 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.

12.2 MEANING OF STATISTICS

We have come across the statistics all the time for instance,

- the inflation rate has gone up 20% since last year.
- the crime rate has reduced by 5% thane that of last year.

All the above statements are statistical conclusions. These statistical statements are very convenient type of communication to understand the readers and also helps in formulating specific policies pertaining to that area.

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 – there are 932 females per 1,000 males in India, the per capita national product at current prices has increased from Rs. 246 in 1950-51to Rs. 2,596 in 1985-86 here 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.

12.2.1 Statistics Defined in Plural Sense

In its plural connotation, statistics means data or numerical figures pertaining to any given situation or a phenomenon. These may be quantitative or qualitative data.

Quantitative data may represent the numerical observations in relation to a continuous variable. A continuous variable is the one which can assume any value between any two points on a line segment. All characteristics such as weight, length, height, thickness, velocity, temperature, and the like are all continuous variables. Discrete data, on the other hand, refer to values assumed by a discrete variable. A discrete variable is represented by fixed values, generally integers such as 1,2,3,…… . These are count data collected by making a count of the number of items possessing or not possessing a certain characteristic. For example, the number of incoming flights at an airport, or the number of defective items in a consignment received for sale.

Qualitative data may be nominal or ranked. The *Nominal data* arise due to classification into two or more categories of a certain number of items according to some quality characteristic. For example, classification of students according to sex. (as males and females) or according to the level of education (as matriculates, undergraduates, and postgraduates). Such data are also count data. *The ranked data*, on the other hand, are the result of assigning ranks according to the level of performance in any competitive test, contest, or interview. Candidates appearing in an interview, for instance, may be assigned ranks in integers ranging from 1 to *n*, depending on their performance in the interview. The ranks so assigned may be viewed as continuous values of a variable which may be any quality characteristic under observation.

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 cay be stated in numbers or any tabular or classified arrangement." To Bowley statistics "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 Dr 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 t o 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 than 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 manner 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 comparableif 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.

12.2.2 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.

Croxton 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.

12.3 DESCRIPTIVE AND INFERENTIAL STATISTICS

Statistics as a subject is very wide. If consists of methods of handling massive data in a variety of problem situation.

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 characteristic features 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 not 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 not worthy 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.

12.4 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 statementsin 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 somecommon 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 of 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 2005, the per-capita income in India has increased by 10% in 2005. 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 2006.

4) **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.

5) **To forecast future values**: Some of the statistical techniques are used for forecasting future values of a variable. O n 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.

6) **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.

7) **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.

8) **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.

9) **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.

12.5 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-product of 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 world over 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 now 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.

12.6 LIMITATIONS OF STATISTICS

In spite of its important functions, statistics has its limitations too. These limitations should be kept in mind while using the various statistical methods. Now, we shall discuss some of the limitations of statistics.

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.Qs).

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 or 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 [or 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 onc o l 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 o r 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.

12.7 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 d o 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 avalid 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.

12.8 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 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 variables used, there are three categories of frequency distribution : 1) uni-variate frequency distribution, 2) bi-variate frequency distribution, and 3). multivariate frequency distribution. In this unit we will discuss uni-variate analysis and bi-variate analysis only.

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.

2) **Bi-variate Frequency Distribution**: The frequency distribution with two variable is called bi-variate frequency distribution. If a frequency distribution shows two variables i.e., marks in statistics and age, it is known as bi-variate frequency distribution.

12.9 LET US SUM UP

The word statistics can be used either plural sense or in singular sense. 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 a reasonable 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, and5) 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.

12.10 KEY WORDS

Data: A collection of measurements or reservations on one or more variables.

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.

12.11 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.

12.12 TERMINAL QUESTIONS

1) Why it is necessary to have knowledge on statistics?

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

3) What do you mean by statistics? Explain its importance to Economics and Business.

4) Define statistics and discuss the various functions of statistics.

5) Discuss the usefulness of statistics and explain the limitations of statistics.

6) 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 13 MEASURES OF CENTRAL TENDENCY

Structure

13.0 OBJECTIVES

After studying this unit, you should be able to:

- understand concept of central tendency,
- appreciate the purpose of calculating averages
- enumerate the qualities of ideal average
- define and compute the arithmetic mean, geometric mean, harmonic mean, median, partition values and mode for different types of data
- explain the properties, merits and limitations of different measures of central tendency or averages, and
- identify the suitable average for a given purpose.

13.1 INTRODUCTION

We have discussed why it is important and relevant to study statistics as a commerce students. Statistics broadly means data or numerical figures pertaining to any given situation or a phenomenon. To make life easier we need to present data properly and understand its characteristics behaviour and treatment. 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 Averages or Central Tendency for obtaining a single value that represent the entire data, which gives a bird's-eye view of the entire data.

In this unit you will study the purpose of calculating averages, the essentials of an ideal average, and identify different measures of average. You will further learn in detail the calculations, properties, merits, and limitations of measures of averages, viz. Arithmetic Mean, Weighted Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Partition Values (Quartile, Deciles and percentiles) and Mode.

13.2 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 valuables 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. Let us discuss the first set of measures which are called **Averages or Measures of Central Tendency or Measures of Location**. We discuss about the other set of measure (i.e., measures of dispersion) in next unit of this Block.

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.

Let us see some definitions of central tendency:

Clark defined it as "Average is an attempt to find one single figure to describe whole of figures". Croxtan and Cowden defined as "An average value is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is something called a measure of central value."

The above definitions explain us that the average or central value is a single value which represents the entire complex mass of data. Therefore, central value lies somewhere in between the highest value and the lowest value of the given data. Thus an average of a given data is frequently referred to as a measure of central tendency.

13.3 OBJECTIVES OF AVERAGES OR CENTRAL TENDENCY

You have studied the concept of central tendency. Now let us discuss the major objectives of computing averages. The following are the main objectives:

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 sale during the previous year and the current year's average. It is important to note that 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.

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 a plant, representative sales volume, overall productivity index, price index, etc. These all are the connotations of an average.

13.4 ESSENTIALS OF AN IDEAL AVERAGE

Keeping in view of the objectives of averages, let us try to understand the requisites 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.

13.4.1 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.

2) Averages of Position

   i) Median; ii) Partition Values – quartiles, deciles and percentiles; iii) Mode

13.5 ARITHMETIC MEAN

The word average, we use every frequently in day-to-day expressions. Such as average price, average income, average weight etc. In these expressions the word average is nothing but arithmetic mean. Generally a layman call an average but a statistician call the 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 obtained 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, production, etc. The arithmetic mean is also called as **simple Arithmetic Mean**.

13.5.1 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 $\overline{X}$ which is read as 'X bar'**

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:

$$\overline{X} = \frac{X_1 + X_2 + \ldots \ldots + X_n}{n}$$

Where $\overline{X}$ (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:

$$\overline{X} = \frac{\sum X_1}{n}$$

Where, the $\sum$ (read it as sigma) is the Greek symbol denoting the summation over all values of x.

$\sum x$ is sum of the value of observations; n is the number of the observations.

Steps to compute

1) Add all values of the given observations ($\sum$ x); 2) Obtain the total number of observations (n); 3) Apply the formula.

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; and Product 5 - Rs. 9

Solution: Average profit can be computed as follows:

$$\overline{X} = \frac{\sum x_i}{n}$$

$$= \frac{4+9+6+2+9}{5} = \frac{30}{5} = \text{Rs.}\, 6.00$$

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 ($\sum$d called sigma d). 4) Compute the arithmetic mean by using the following formula:

$$\overline{X} = \text{A} + \frac{\sum \text{d}}{n}$$

Where, $\overline{X}$ is the arithmetic mean of the variable x; A is the assumed mean; $\sum$d 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		**$\sum$d = – 24**

Assumed mean (A) = 25; $\sum$d = – 24 ; n = 10

$$\overline{X} = \mathrm{A} + \frac{\sum \mathrm{d}}{n} = 25 + \frac{-\ 24}{10} = 25 -\ 2.4\ = 22.6\ (Average\ scooter\ sole)$$

Grouped Data: 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: It is also called Direct Method. Under this method the mean for grouped data can be obtained by using the following formula:

$$\overline{X} = \frac{f_1 x_1 +\ f_2 x_2 + \ldots f_n x_n}{f_1 +\ f_2 + \ldots f_n}$$

where, x,, x_2, x_3 etc., refer to the values of the variable in classes 1 , 2 , 3etc., respectively. Similarly, f,, 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). $f_2 x_2$, $f_3 x_3$....f x indicate the same meaning.

This formula can be simplified as : $\overline{X} = \frac{\sum \mathrm{fx}}{\sum \mathrm{f}}\ or\ \frac{\sum \mathrm{fx}}{\mathrm{n}}$

Where, f is the frequency; x is the value of the variable.

Steps to compute:

1) Multiply the frequency of each row with the value of variable and obtain the total i.e. $\sum$fx; 2) Obtain the sum of frequency ($\sum$f); It is also termed as the number of observations (n); 3) Apply the formula.

Method 2: It is also called short-cut method. 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:

1) Take an assumed mean A. 2) 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. 3) Obtain ∑fd by multiplying deviations (d) with their respective class frequencies (f) and summing it. 4) Obtain the number of observations (n) i.e., total frequency (∑f) 5) Compute the mean by applying the following formula:

$$\overline{X} = \mathrm{A} + \frac{\sum \mathrm{fd}}{\sum \mathrm{f}} \; or \; \overline{X} = A + \frac{\sum \mathrm{fx}}{\mathrm{n}}$$

Where, A is the assumed mean; ∑f denotes the total number of items, it can also be denoted by 'n'; ∑fd is the sum total of the deviations (d = x – A) multiplied with their respective class frequencies.

Now let 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 fo 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	360
70	5	30	150	350
80	2	40	80	160
Total	**∑f = 100**		**∑fd = –330**	**∑fx = 3,670**

In this case assumed mean (A) is 40 marks.

Method 1: $\overline{X} = \frac{\sum \mathrm{fx}}{\mathrm{n}} = \frac{3,670}{100} = 36.70$ marks

Method 2: $\overline{X} = A + \frac{\sum fd}{n} = 40 + \frac{-330}{100} = 40 - 3.30 = 36.70$ marks

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: This method is also called direct method. While computing the mean, it is to be kept in mind that it is not necessary to convert inclusive classes into exclusive classes and there is no need to exchange the unequal classes into equal classes. Under this method the arithmetic mean is obtained by using the following formula:

$$\overline{X} = \frac{\sum fm}{\sum f} \; or \; \overline{X} = \frac{\sum fm}{n}$$

Where, $\overline{X}$ is the arithmetic mean; $\sum f$ or n is the total frequency or total number of items; m is the mid-value of the class.

Steps to compute:

1) Get the mid value of each class and denote it by m (i.e.,) m-class interval÷ 2
2) Multiply these mid-values by its respective frequency of each class and obtain the total i.e., $\sum fm$
3) Obtain the total frequency i.e. $\sum f$ or n
4) Apply the formula

Method 2: This is also known as short-cut method or deviation method. The same formula as used for discrete series can be used here also, with a slight change in obtaining 'd'. Here, deviation of mid-values from assured mean are obtained (i.e., d = m −A).

$$\overline{X} = A + \frac{\sum fd}{\sum f} \; or \; \overline{X} = A + \frac{\sum fd}{n}$$

Where, $\overline{X}$ is the assumed mean; f is the frequency; d is the deviation.

Steps to compute:

1) Obtain the mid value of each class (m). 2) Choose any mid-value as assumed mean (A). You are advised to choose the balanced value from the two exreems as assumed mean. 3) Subtract the assumed mean from each mid-value (m −A) i.e., d. 4) Obtain the sum of frequency i.e. $\sum f$ or n 5) Apply the formula

Method 3: This is known as Step Deviation Method. The formula of previous two methods can be used conveniently if the value of variable (x) and values of frequencies (f) are small. If the values of x and f are large computation of mean by using the above methods are quite tedious and time consuming. In such a situation the calculations can be reduced to a greater extent by using step-deviation method. If the deviations from assumed mean have some common factor. Common factor is the highest value which can divide all the deviations (d) without remainder a further reduction in the size of deviation is possible by dividing deviations by the common factor (c) and denoting these step deviations by d^1 i.e. d^1 = (m −A) ÷ c. The symbol d^1has been introduced to differentiate it from d i.e., (m −A) = d. It is to be noted that this method is applicable in discrete and continuous series, secondly, if all the class intervals are equal than, the class interval will be the common factor. Under this method, arithmetic mean will be computed by the following formula.

$$\overline{X} = \mathrm{A} + \frac{\sum \mathrm{f}d^1}{\sum \mathrm{f}} \times c \; or \; \overline{X} = A + \frac{\sum \mathrm{f}d^1}{\mathrm{n}} \times c$$

where A is assumed mean; f is the frequency, d^1 is the reduced deviation by dividing the deviation (d = m −A) with common factor (c); ∑f is the sum of frequencies or total number of observations (n).

Steps to compute the mean:

1) find the mid value of each class (m) and select assumed mean (A) from any mid-value.
2) Find the deviations by subtracting the assumed mean (A) from each mid-value i.e. (m −A) = d.
3) Find the common factor (c) and divide the above deviation (d) by the common factor and denote by d^1.
4) Multiply the reduced deviation (d^1) in step 3, with their corresponding frequencies (f) and obtain the total i.e. $\sum \mathrm{f}d^1$.
5) Obtain the total number of observation (sum of frequencies) i.e. ∑f or n
6) Apply the formula of step deviation method.

Illustration 4: Weekly sales of 50 salesmen of a company are given below. Calculate the arithmetic mean by following the direct method, short-cut method and 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 Sales men (f)	Mid point (m)	Deviation (m-17.5) (d)	Step deviation $d^1 = \frac{m - 17.5}{5}$	fm	fd	fd^1
0-5	3	2.5	– 15	– 3	7.5	– 45	– 9
5-10	12	7.5	– 10	– 2	90.0	– 125	– 24
10-25	25	17.5	0	0	437.5	0	0
25-50	10	37.5	20	4	375.0	200	40
Total	**∑f = 50**				**∑fm = 910**	**∑fd = 35**	**$\sum fd^1$ = 7**

Method 1:

$\overline{X} = \frac{\sum fm}{n}$, $\sum fm = 910$, $n = 50$

$\overline{X} = \frac{910}{50} = 18.2$ Mean of sales is Rs.18.2 thousand per week.

Method 2: It is apparent from deviation column that here assumed Mean (A) is 17.5.

Now, $\overline{X} = A + \frac{\sum fd}{n}$

A= 17.5, $\sum fd = 35$, n = 50

$\overline{X} = 17.5 + \frac{35}{50} = 18.2$ (Mean of sales is Rs. 18.2 thousand per week.)

Method 3: Step deviation method here the common factor (c) is 5.

Now, $\overline{X} = A + \frac{\sum fd^1}{n} \times c$

$\overline{X} = 17.5 + \frac{7}{50} \times 5$

$= 17.5 + 0.7 = 18.2$ (Arithmetic Mean of sales is Rs. 18.2 thousand per week.)

We understand from the above solution that the three different methods of calculation from arithmetic mean give us the same result. It is clear that the step-deviation method minimises the calculations. Thus, this method makes the calculation easier than the other two methods, though the method 1 is simplest. Method 3 is the suitable when mid-values and frequencies are very large.

Illustration 5: Find the average number of hours worked by the employees of the Yamto Machine Co. from the data given below:

Hours worked	**No. of employees**
36.0 - 37.8	6
37.8 - 39.6	7
39.6 - 41.4	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 clauses 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 $(\mathbf{m-42.3})$	$d^{1} = (m-A)/C$ $d^1 = \frac{m-42.3}{1.8}$	fd^1
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
Total		**n = 50**			$\sum fd^1 = 46$

$$\overline{X} = A + \frac{\sum fd^1}{n} \times c$$

$$\overline{X} = 42.3 + \frac{-46}{50} \times 1.8 = 42.3 + (-0.92) \times 1.8 = 42.3 - 1.656 = 40.644$$

(Arithmetic mean of the hours worked is 40.6 hours.)

You may notice when class intervals are all equal, d^1 values will be 1, 2, 3, and – 1, – 2, – 3, … etc. But when class intervals are not equal, the d^1 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^1 may be written directly.

It is important to note that when the classes are given in inclusive method it is not necessary to adjust the classes into exclusive method for calculation of arithmetic mean, geometric mean and harmonic mean because the mid-value remain the same. However, in case of positional averages, such as median and mode, adjustment is essential.

Check Your Progress A

1) 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: $\overline{X} = 33$, $\sum fd^1 = -20$, $\sum f = 100$ and $c = 10$; find the assumed mean A.

v) What is the major assumption we make while computing a mean from grouped data?

2) 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

3) 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?

4) 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

5) Find the mean from the following distribution by step deviation

Class Interval	:	15-25	25-35	35-45	45-55	55-65	65-75
Frequency	:	4	11	19	14	8	2

13.5.2 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. Thus the term weight refers the relative importance of the different items.

Computation: To compute weighted arithmetic mean, the formula is:

$$\overline{X}w = \frac{\sum \mathrm{wx}}{\sum \mathrm{w}}$$

Where, $\overline{X}$ is weighted arithmetic means; $\sum$wx is sum of the product of weights (w) multiplied with the respective variables (x); and $\sum$w is sum of the weights.

Steps: 1) If weights are not given assign arbitrary weights as per the situation; 2) Multiply the weights (w) with the respective variables (x) and obtain total i.e. $\sum$wx; 3) Obtain the sum of weights i.e., $\sum$w; 4) Apply the formula.

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.

Illustration 6: 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		**$\sum$w= 10**	**$\sum$wx = 510**

Weighted Arithmetic Mean $= \frac{\sum \mathrm{wx}}{\sum \mathrm{w}}$

$= \frac{510}{10} = 51\%$ (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.

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 x**	**Set 1**		**Set 2**	
		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	$\sum x = 190$	$\sum w_1 = 10$	$\sum xw_1 = 760$	$\sum w_2 = 30$	$\sum xw_2 = 1900$

1) Weighted Mean for set 1 $= \frac{\sum XW}{\sum w} = \frac{760}{10} = 76\%$

2) Weighted Mean for Set 2 $= \frac{\sum XW}{\sum w} = \frac{1900}{10} = 63.3\%$

3) Simple Mean $= \frac{\sum x}{n} = \frac{190}{3} = 63.3\%$

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.

13.5.3 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 denominators

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.

13.5.4 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., $\sum (x - \overline{X}) = 0$. This is explained in the following illustration.

x	(x - $\overline{X}$)
5	– 1
6	0
7	1
9	3
3	– 3
30	$\sum$ (x - $\overline{X}$) = 0

$\overline{X} = \sum \text{x}/\text{n} = 30/5 = 6$

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 kind; 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, $\sum(\text{x} - \overline{\text{X}})^2$ is always minimum. We can verify this for the illustration discussed above.

Squared Deviations taken from mean (X = 6)

x	(x – $\overline{X}$)	(x– $\overline{X}$)^2
5	– 1	1
6	0	0
7	1	1
9	3	9
3	– 3	9
		20

Squared deviations taken from any other values say 5

X	(x– 5)	(x – 5)^2
5	0	0
6	1	1
7	2	4
9	4	16
3	– 2	4
		25

It is clear that $\sum(\overline{X} - X)^2 < \sum(\overline{X} - 5)^2$

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., $\sum$X = n$\overline{X}$, 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 x 100 = Rs. 40,000.

4) If we add or delete an observation which is equal to mean, the arithmetic mean remains unaffected. For example, let us assume the arithmetic mean of 10 observations is 15 and the 11th observation value is 15. Now, the revised mean would be 15 i.e. (10×15)+15÷11.

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 '$\overline{X}$' 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) x 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 = 30/5 = 6

Mean of x+2 = 40/5 = 8 = 6 + 2 i.e., old mean +2

Mean of 3(x+2) = 120/5 = 24 or 8 × 3 or (6+2) × 3 i.e., (old mean +2) × 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:

$$\overline{X}_c = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{(n_1 + n_2)}$$

Where $\overline{X}_1$ and $\overline{X}_2$ are the arithmetic mean of group 1 and group 2 respectively, and n_1 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.

$\overline{X}_1$ = 400; $\overline{X}_2$ = 430; n_1= 8 (January to August – 8 months)

n_2 = 4 (September to December – 4 months)

The average for the whole year

$$\overline{X}_c = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{(n_1 + n_2)} = \frac{8 \times 400 + 4 \times 430}{8 + 4} = \frac{4920}{12}$$

= 410 tonnes per month.

The logic behind the formula is: $n_1 \overline{X}_1$ is the total value of all the items belonging to the first group and $n_2 \overline{X}_2$ is the total for the second group. Thus, $n_1 \overline{X}_1 + n_2 \overline{X}_2$ is the total of all the items in both the groups. In other words, the combined mean is the weighted average of the mean 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 different between the two.

Price per tonns (Rs.)	:	45.60	40.70	42.75
Tonnes purchased	:	135.00	40.00	25.00

3) From the results of two college 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 the student in written oral tests in subject A, B and C are as follow:

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 is weights.

13.5.5 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 computation of other statistical measures like standard deviation, coefficient of variation, co-efficient of skewness, etc.

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 and 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 distributions, 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 norms (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.

13.6 GEOMETRIC MEAN AND HARMONIC MEAN

You have already studied about arithmetic mean which belongs to the category of mathematical averages. Now, you will study about the two other mathematical averages viz, Geometric Mean and Harmonic Mean.

13.6.1 Computation of 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.

Computation: Like other averages, computation procedure of geometric mean is different for grouped data and ungrouped data. Let, us now, 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:

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. - $\sqrt[3]{4 \times 8 \times 16} = \sqrt[3]{518} = 8$

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.

Symbolically it can be expressed as:

$$\log G.M. = \frac{1}{n} \log(X_1, X_2 \ldots \ldots .. X_n)$$

$$= \frac{\log X_1 + \log X_2 + \ldots \ldots \ldots .. \log X_n}{n} = \frac{\sum \log X}{n}$$

Therefore, G.M.= $Antilog\ \frac{\sum \log X}{n}$

Steps to calculated GM: 1) Obtain the logarithm of the different values of the variable and take their total i.e., ∑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. **How to find logarithm and antilogarithm of a value is explained clearly and also provided logarithms and antilogarithms tables at the end of this unit**.

For example, geometric mean of four numbers 20,65,83 and 135 will be:

G.M. = $Antilog \frac{\log 20 + \log 65 + \log 83 + \log 135}{4}$

$= Antilog \frac{1.3010 + 1.8129 + 1.9191 + 2.1303}{4} = Antilog\ 1.7908$

G.M. = 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%, 140% 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

Now, G.M. Antilog $= \frac{\sum \log X}{n} = Antilog \frac{\log 132 + \log 140 + \log 150}{3}$

$$= Antilog \frac{2.12.06 + 2.1461 + 2.1761}{3} = Antilog \frac{6.4428}{3} = Antilog\ 2.1476$$

G.M = 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:

$$\text{G.M.} = \sqrt[n]{X_1^{f_1} X_2^{f_2} \ldots \ldots . X_n^{f_n}}$$

where $X_1, X_2 \ldots \ldots . . X_n$, are the different values of the variate x with their respective frequencies $f_1, f_2 \ldots \ldots . . f_n$ and $n = f_1, f_2 \ldots \ldots . . f_n = \sum f$

$$\log \text{G.M.} = \frac{1}{n} (f_1 \log X_1 + f_2 \log X_2 + \ldots . . f_n \log X_n) = \frac{1}{n} \left(\frac{\sum f \log x}{n} \right)$$

The above expression can be simplified as: G.M. = Antilog $\left(\frac{\sum f \log x}{n} \right)$

Steps to calculate G.M.:

1) Find the logarithms of the given observations i.e. ∑log x; 2) Multiply these logarithms (logx) with the corresponding, frequency and obtain the total i.e., ∑f Logx; 3) Obtain the total of observations i.e., n or ∑f; 4) Apply the formula.

Let us, Now consider the following illustration:

Illustration 2: Calculate the geometric mean from the following distributions.

Marks :	5	15	25	35	45
No. of Students :	5	7	15	25	8

Solution: Calculation of G.M.

Marks (x)	No. of Students (f)	log x	f logx
5	5	0.6990	3.4950
15	7	1.1761	8.2327
25	15	1.3979	20.9685
35	25	1.5441	38.6025
45	8	1.6532	13.2256
	n = 60		**∑f logx = 84.5243**

G.M. $= \text{Anti log}\left(\frac{\sum f \log x}{n}\right)$

$= \text{Anti log}\left(\frac{84.5243}{60}\right) = \text{Anti log } 1.4087$

G.M. =25.63 marks

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. $= \text{Antilog}\left(\frac{\sum f \log m}{n}\right)$

Steps to be followed: 1) Find the mid-value of the classes i.e., m; 2) Find the logarithms of the mid-values i.e., log.m; 3) Multiply the logarithms of mid-values with the respective frequencies and get the total i.e., ∑f log.m; 4) Apply the formula.

Illustration 3: 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: Calculation of geometric mean

Class interval	Mid-Point(m)	Log.m	f	flog.m
7.5-10.5	9	0.9542	5	4.7710
10.5-13.5	12	1.0797	9	9.71 28
13.5-16.5	I5	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
			n=68	∑f log.m = 81.9092

G.M. $= \text{Antilog}\left(\frac{\sum f \log m}{n}\right)$

$= \text{Antilog}\left(\frac{81.9092}{68}\right)$ = Antilog 1.2045

G.M. =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_0 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:

$P_n = P_0\ (1+r)^n$, where 'n' is the time-span

$$(1+r)^n = \frac{P_n}{P_0}$$

$$(1+r) = \sqrt[n]{\frac{P_n}{P_0}}$$

$$r = \sqrt[n]{\frac{P_n}{P_0}} - 1$$

Let us now take an illustration to understand the calculation of the percentage compound growth rate per annum.

Illustration 4: The population of a country was 300 millions in1951. It became 520 millions in 1969. Calculate the percentage compound rate of growth per annum.

Solution: Here P_0 is 300: P_n is 520 and n is 18. Let 'r' be the growth rate per annum.

Here, $1 + r = \sqrt[n]{\frac{P_n}{P_0}}$

$= \sqrt[18]{\frac{520}{300}}$ using logarithms

$$\log(1 + r) = \frac{\log 520 - log\,300}{18}$$

$$1 + r = Antilog\,\left(\frac{2.7160 - 2.477}{18}\right)$$

$$= Antilog\,\left(\frac{0.2389}{18}\right) = Antilog\,0.0133 = 1.031$$

$$r = 1.031 - 1 = 0.031$$

$$percentage\ compound\ grwoth\ rate\ is\ 100 \times r = 3.1\%$$

Weighted Geometric Mean

Like weighted arithmetic mean, we can also calculate the weighted geometric mean. The computational procedure is as follows:

Weighted G.M. = $\sqrt[\Sigma w]{X_{1.}^{w1}, X_2^{w2} \ldots\ldots. X_n^{wn}}$

Where $X_1, X_2 \ldots\ldots.. X_n$ are the values of the variate and $W_1, W_2 \ldots\ldots.. W_n$ are the corresponding weights

Taking logarithms, Log Weighted G.M. = $\frac{W_1 \log X_1, + W_2 \log X_2 + \ldots\ldots W_n \log X_n}{\Sigma W}$

Or, log Weighted G.M. = $\frac{\Sigma W \log X}{\Sigma W}$

The above expression can be simplified as: Weighted G.M. = Antilog $\left[\frac{\Sigma W \log X}{\Sigma W}\right]$

Steps for calculation: 1) Find the logarithms of value of variables i.e., logx; 2) Multiply the above logarithms with respective weights (w.log x) and obtain the total i.e., ∑w.logx; 3) Obtain the total of weights i.e. ∑W; 4) Apply the formula.

Let us consider an illustration to understand the calculation:

Illustration 5: 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

Solution: Calculation of weighted Geometric Mean

Group	Index No.	Weight	Logx	W.Log x
Food	300	40	2.4771	99.084
Fuel	200	10	2.3010	23.01
Cloth	250	10	2.3979	23.979
House Rent	150	15	2.1761	32.6415
		∑W=75		**∑Wlogx = 178.7145**

Weighted G.M. $= \text{Antilog}\left[\frac{\sum W \log X}{\sum W}\right]$

$= \text{Antilog}\left[\frac{178.7145}{75}\right]$

$= \text{Antilog } 2.3829 = 241.50$

Therefore, weighted geometric mean of index numbers is 241.50

13.6.1.1 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 or example, the geometric mean of the items.4, 8 and 16 is 8. Therefore 4 × 8 × 16 = 8 × 8 × 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:

   Combined G.M. = Antilog $\left[\frac{N_1 \log GM_1 + N_2 \log GM_2 + \ldots N_n \log GM_n}{N_1 + N_2 + \ldots N_n}\right]$

Where, GM_1 = Geometric Mean of the first group; GM_2= Geometric Mean of the second group; GM_nGeometric mean of the nth group.

For example, let 100 items have GM = 50 and 200 items have GM = 40. Then the combined geometric mean will be:

Combined G.M. $= \text{Antilog}\left[\frac{100 \log 50 + 200 \log 40}{100+200}\right]$

$= \text{Antilog}\left[\frac{100 \times 1.6990 + 200 \times 1.6021}{300}\right]$

$= \text{Antilog } 1.6344 = 43.09$

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.

Arithmetic mean $= \frac{2+3+5+10+100}{5} = 24$

Geometric mean $= \text{Antilog}\left[\frac{\log 2+\log 3+\log 5+\log 10+\log 100}{5}\right]$

$= \text{Antilog}\left[\frac{0.3010+0.4771+0.6990+1.0000+2.0000}{5}\right]$

$= \text{Antilog}\ \frac{4.4771}{5} = \text{Antilog}\ 0.8954$

G.M. = 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.

13.6.1.2 Uses and Limitations

Uses:

1) For computing the averages of ratio 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 are 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 the case of income distributions

Check Your Progress C

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?

13.6.2 Computation of 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.

Computation: The method of computing harmonic mean is different for ungrouped data and grouped data. Let, us now, study these methods separately.

Ungrouped Data

If there are 'n' values of variate x viz., $X_1, X_2 \dots\dots.. X_n$ their harmonic mean (HM) is calculated as follows:

Harmonic Mean = $\dfrac{n}{\frac{1}{X_1}+\frac{1}{X_2}+ \dots..+\frac{1}{X_n}} = \dfrac{n}{\Sigma\frac{1}{X}}$

In simplest expression the formula is as follows:

H.M. $= \dfrac{n}{\Sigma\frac{1}{X}}$ or H.M. = Reciprocal of $\left[\dfrac{\Sigma\frac{1}{X}}{n}\right]$

= Reciprocal of (arithmetic mean of reciprocals of n values $(X_1, X_2 .. X_n)$.

Therefore, **harmonic mean is the reciprocal of the arithmetic mean of reciprocals**. **Like logarithms, we can find the reciprocals of the given value by consulting the reciprocal tables provided at the end of this unit.**

The procedure for finding the reciprocal in the table is similar of finding logarithms. But you should keep in mind the value in the mean difference column must be subtracted.

In individual series first we have to find the reciprocals of the value of observations. Then apply the above formula to obtain the H.M. For example, harmonic mean of two values 12 and 15 cañ be computed as follows:

H.M. = $\frac{2}{\frac{1}{12}+\frac{1}{15}} = \frac{2}{\frac{5+4}{60}} = \frac{120}{9} = 13.34$

Illustration 6: A motorist travelled for three days at the rate of 480 kms. A day. On the first day. he traveller 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 s constant, and time and speed are variable. We are required to compute the average speed. Therefore, harmonic mean is the appropriate average. ’

H.M. = $\frac{3}{\frac{1}{48}+\frac{1}{40}+\frac{1}{32}} = \frac{3}{\frac{37}{480}} = \frac{3\times 480}{37} = 39\ kms.per\ hour\ (approximately).$

Here, now 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 = l0 + 12 + 15= 37

Therefore, The average speed= $\frac{3\times 480}{37}$ = 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:

H.M. = $\frac{\text{n}}{\Sigma\text{f(reciprocals of x)}} = \frac{n}{\Sigma f\frac{1}{X}}\ or\ Reciprocal\ \frac{\Sigma f\frac{1}{X}}{n}$

Where, symbols have their usual meaning.

Steps for calculations: 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.($\Sigma f\ \frac{1}{X}$); 3) Take a ratio of the total frequency (n) to $\Sigma f\ \frac{1}{X}$.

Illustration 7: 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 = Σf = 60**		$\Sigma f\frac{1}{X}$ **= 12.0**

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. The answer is that 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 l/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 quantity units, the correct average is the harmonic mean. In general, we can say, that while finding the 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, H.M. = $\frac{\text{n}}{\Sigma \text{f.m}}$ or Reciprocal of $\frac{\Sigma f\frac{1}{m}}{n}$

Illustration 8: Calculate harmonic mean for the following information:

Class Interval	f
0-10	5
10-20	8
20-30	10
30-40	12
40-50	7
50-60	6
60-70	3

Solution: Computation of Harmonic Mean

Class Interval	f	Mid-Value (m)	1/m	$f \times \frac{1}{m}$
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	12	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
	n=51			$\Sigma f\frac{1}{m} = 2.591$

H.M. = $\frac{n}{\Sigma f\frac{1}{m}} = \frac{51}{2.591} = 19.68$

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 weighty. The formula to be used here is:

Weighted H.M. $= \frac{\Sigma w}{\Sigma \frac{w}{X}}$ where, 'W' refers to weights

Alternatively, Weighted H.M. = Reciprocal of $\frac{\Sigma \frac{w}{X}}{\Sigma w}$

In the above example x : 4 3 2

w: 10 5 4

Weighted H.M. = $\frac{10+5+4}{\frac{10}{4}+\frac{5}{3}+\frac{4}{2}} = \frac{19}{2.5+1.67+2} = \frac{19}{6.17} = 3.08\ kms. per\ hour$

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.	3km.p.h.	5/3	1.67
Third	4 kms.	2km.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 9: 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

$$\text{H.M.} = \frac{\Sigma w}{\frac{\Sigma w}{X}} = \frac{4+1.5+0.5}{\frac{1}{40}\times 4+\frac{1}{10}\times 1.5+\frac{1}{4}\times 0.5} = \frac{6}{0.1+0.15+0.125} = \frac{6}{0.375} = 16$$

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 x 60/ 22.5).

13.6.2.1 Properties of Harmonic Mean

1) If each value of the narrate is replaced by harmonic mean, the total of reciprocals of values of the narrate 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.

$$\text{H.M.} = \frac{5}{\frac{1}{2}+\frac{1}{3}+\frac{1}{5}+\frac{1}{10}+\frac{1}{100}} = \frac{5}{0.50+0.33+0.20+0.10+0.01} = \frac{5}{1.14} = 4.39$$

As computed either 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.

13.6.2.2 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.

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 10: 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.

   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.

   Arithmetic Mean $= \frac{30+20}{2} = 25$ kms. 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 39 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 D

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?

13.7 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'. Canor defined the median as "The median is that value of the variable which divides the group in two equal parts, one part comprising all the values greater and other all values less than the median.

13.7.1 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.

Ungrouped Data: Having arranged the data in ascending order or descending order, the median is calculated as $\frac{n+1}{2}$ th item, where 'n' is the total number of items. This process is to be followed in the following both situations.

1) **When n is odd**: When the number of observations is an odd number, the procedure to find the median is as follows.

   For example takes 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 mean that when the given series is arranged in a ascending order, the fifth item will be the median. Now, we can arrange the data is 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. Let us take an example to understand the procedure to find the median in this situation.

For example, take the seies 8, 11, 13, 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 involves a fraction 0.5. You should not that there is no 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 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 series are 16 and 20 respectively.

Therefore, M_d is 18 $\left(i.e., \frac{16+20}{2}\right)$.

Even when n is an even number, median can be taken as $\frac{n+1}{2}$th item. But for this purpose you have 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 $\left(i.e., 16 + \frac{1}{2}\,(20 - 16)\right)$. 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 not 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 formulat helps us in giving a general definitions to median for ungrouped data.

Grouped Data: As you know, when the data is 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, the following steps to be followed to calculate median.

1) Arrange the value of observations (x) either in ascending order or in descending order along with their respective observations (frequencies).

2) Convert the frequency (f) into cumulative frequencuy (cf).

3) Apply the formula i.e., $(M_d) = \frac{n+1}{2}$th item.

4) Now, locate the value of $\frac{n+1}{2}$th item in the cumulative frequency and determine the value of the variable corresponding to that cumulative frequency locate as above.

5) This value of the variable is the median value.

Let us understand the computation of median by an illustration.

Illustration 1: Calculate the median marks from 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: Examine the solution carefully by referring the above explained steps. 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	30
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 value 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 the class by any of the following two methods:

Method 1: When cumulative frequency is formed in "less than" method.

$$M_d = l + \frac{\frac{N}{2} - C}{f} \times i$$

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

N/2 = half of the number of observations, it is also denoted 'm'

The above formula can also be expressed in the following manner:

$$M_d = l + \frac{U - L}{f} \times (m - c)$$

Where, l = lower limit of the median class; u= upper limit of the median class; f = frequency of the median class; m = $\frac{n}{2}$ th item; c = cumulative frequency of the class preceding the median class.

Steps to calculate the median in continuous series data:

1) If the classes are given in inclusive form they must be converted into exclusive method or it is enough to convert the median class only. The procedure will be explained latter in this unit. It is not necessary to convert unequal classes into equal classes.

2) Calculate less than cumulative frequency (cf).

3) Find the n/2th item and locate the value of that item where it lies in the cumulative frequency then find the corresponding class of the cumulative frequency. This class is the median class.

4) Interpolate the value of median from the median class by using any formula as presented above under the method.

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 circulated from higher values side. i.e., "more than" method the above formula can be slightly modified as:

$$M_d = U + \frac{\frac{N}{2} - C}{f} \times i$$

Where, U = lower limit of the median class; C= cumulative frequency of a class next to the median class; f = simple frequency of the median class; i = the class-interval of the median class

You should note that the procedure for computation of median under this method is same as the procedure explained under method 1, except in step No.2, i.e., in this method we have to calculate more than cumulative frequency instead of less than cumulative frequency.

These two methods produce exactly the same result. The assumption and the logic for interpolating median by these two 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 I 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. Therefore $\frac{N}{2} - C$ item are to be covered in the median class. There are 'f' items spread over a class intervals '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 intervals, 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. Let us take up an illustration to understand the calculation of median by using both the methods.

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 Accounts**
30-44	40
45-59	45
60-74	40
75-89	25
90-104	25
105-119	20
120-134	5

Solution: Let us examine carefully the calculations in both the methods to determine the median by referring the steps explained above.

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. Now, as per the first step, we have to convert the Inclusive form of the class into Exclusive form to obtain the real limits of the median class 60-74. The procedure for conversion is as follows: obtain the difference between the lower limit of a class and upper limit of the preceding class here it is 1 (one), divide the difference by 2 i.e., ½ = 0.5.

Now, subtract the result (0.5) from the lower limit of the median class i.e., 60-0.5 = 59.5 and add the same result to upper limit of the same class i.e., 74 + 0.5 = 74.5. Accordingly, the real limit of the median 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) × 15 = 59.5 + 225/40

= 59.5 + 2.625 = 65.125 (Median = 65.1 days)

You can obtain the median by using the alternative **formula expressed in method-1**.

$$M_d = l + \frac{U-l}{f} \times (m - c)$$

Where, l = 59.5; u = 74.5; f = 40; ,=N/2 = 200/2 = 100; C = 85

$$M_d = 59.5 + \frac{74.5-59.5}{40} \times (100 - 85)$$

= 59.5 + (15/40) × 15 = 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) × 15 = 74.5 – 375/40

= 74.5 – 9.375 = 65.125 (Median is 65.1 days).

You should not that the two methods discussed above 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
800 and above	2	1,210

Median has N/2th item below it which mean 1,210/2 = 605th items below it. Therefore, the median lies in the 200-300 class. Now applying the formula of interpolation.

$$M_d = l + \frac{\frac{N}{2} - C}{f} \times i$$

Where, l = 200; c= 550; f = 555; i = 100; N = 1,210.

$$M_d = 200 + \frac{605 - 550}{555} \times 100$$

= 200 + (55/555) × 100 = 200 + 9.91 = 209.91

Median Monthly Income is Rs. 209.91

Note: 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 the following data:

Wages More Than (Rs.)	No. of Workers
20	58
40	54
60	48
80	38
100	22
120	10
140	3
160	0

Solution: Computation of Median

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
160	0	0

Cumulative frequency (more than method) is given in this illustration. So,. We have calculated simple frequency. Now median has N/2th items i.e., 58/2 = 29th, 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:

$$M_d = U + \frac{\frac{N}{2} - C}{f} \times i$$

Where, u = 100; c=22; f = 16; i = 20

$$M_d = 100 + \frac{29-22}{16} \times 20$$

= 100 – (7/16) × 20 = 100 – 8.75 = 91.25. Median wage is Rs.91.25.

Finding the missing frequency: It is possible to find the missing frequencies with the help of the value of median and the total number of observations (N). Let us consider the following illustration.

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 is 375-400 is F. Now, the frequency of the class 425-450 become 1,000 – (525 –F) = 475 – F (525 being the total given frequencies).

Values	Frequency	c.f.
300-325	5	5
325-350	17	22
350-375	80	102
375-400	F	102 + F
400-425	326	425 +F
425-450	425 - F	903
450-475	88	991
475-500	9	1000

Since the median is given as 413.11, the formula must be in 400-425 class.

Now $M_d = l + \frac{\frac{N}{2} - C}{f} \times i$

Where, l = 400; f = 326; C= 102+F; i = 25; M_d = 413.11.

$413.11 = 400 + \frac{500-(102-F)}{326} \times 25$

$413.11 - 400 = \frac{398-F}{326} \times 25$

13.11 × 326 = (398-F) × 25

4,273.86 = 9,950 – 25F

25 F = 5,676.14; F = 227.04

As frequency should be an integral value F = 227. Therefore, frequency for the class 375-400 = 227 and the frequency for the class 425-450 is 475 – 227 = 248.

13.7.2 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., $\sum|x-M_d|$ is the minimum. This property entails the use of median in various practical situations. For example, take the item 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**. $\left(i.e., \frac{5+7+8+9+21}{5}\right) = 10$

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	**18**	**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 distributions, median is the more suitable average, For example, since the income distribution is an open-ended distributions, median income would be a more representative figure.

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, an 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 also be located graphically.

6) It is easily to compute the lucid to understand. In some cases it is obtained even by an inspection.

13.7.3 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 tope 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 observations(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 E

1) 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, 1/6, 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 distributions, 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. 61, 62, 62,, 63, 61, 63, 64, 64, 60, 65, 63, 64, 65, 66, 64. Calculate the median.

   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

13.8 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.

Note: You must keep in mind that the procedure for calculation of these partition values is same as the procedure of median. However, application is slightly differ. You should try to understand the procedure of expressions carefully by comparing the expressions of the median.

13.8.1 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) h 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 expressed in less than cumulative frequency i.e., assigned 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)**

$$Q_1 = l + \frac{J\left(\frac{N}{4}\right) - c}{f} \times i \qquad J = 1, 2, 3$$

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

j = Position of the partition value.

13.8.2 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 ..... 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_1 = Size of $j\frac{N+1}{10}th$ item. J = 1 to 9

ii) **Continuous Series (i.e. Data with Class Intervals)**

Dj = size of $l + \frac{J\left(\frac{N}{10}\right)-c}{f} \times i$ J = 1 to 9

where, C is the cumulated frequency preceding the Jth decile class, the other symbols have usual meaning.

13.8.3 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. 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_{30} = 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 $j\frac{N+1}{100}th$ item.

e.g., P_{45} = Size of $\frac{45\,(N+1)}{100}th$ item

ii) **Continuous Series (i.e. Data with Class Intervals)**

$$P_j = l + \frac{J\left(\frac{N}{100}\right)-c}{f} \times i \qquad J = 1\ to\ 99$$

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 in the illustration.

Q_1 = Size of $\frac{(N+1)}{4}$th item

= $\frac{16+1}{4}$th item = $4\frac{1}{4}$th item

Q_1 = 4th item + $\frac{1}{4}$ (5th item – 4th item)

= $7\frac{3}{4}$ = 7.75

P_{35} = Size of $\frac{35(N+1)}{100}$th item

= $\frac{35\,(16+1)}{100}$th item = $5\frac{95}{100}\,th$ item

P_{35} = 5th item + $\frac{95}{100}$ (6th item – 5th item)

$= 10 + \frac{95}{100}(10-10) = 10 + 0 = 10$

D_9 = Size of $\frac{9(N+1)}{10} = 15\frac{3}{10}\,th$ item

D_9 = 15th item + $\frac{3}{10}$ (16th item – 15th 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_5, P_{25}, P_{75}, Q_3 and Median.

b) Obtain the lirnits 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 × 600/10 = 120 items below it. Therefore, D_2 falls in the 75-150 class.

Now, $D_2 = 1 + \frac{\frac{2N}{10} - C}{f} \times i$

$= 75 + \frac{120-69}{167} \times 75 = 75 + \frac{51}{167} \times 75$

$= 75 + 22.9 = 97.9$ $\qquad D_2 = 97.90$

D_5 has 5N/10 items below it. It means 5 × 600/10 = 300 items below it. Therefore, D_5 falls in the 150-225 class.

Now, $D_5 = 1 + \frac{\frac{5N}{10} - C}{f} \times i$

$= 150 + \frac{300-236}{207} \times 75 = 75 + \frac{64}{207} \times 75$

$= 150 + 23.19 = 173.19$ $\qquad D_5 = 173.19$

P_{25} has 25N/100 items below it. It means 25 × 600/100 = 150 items below it. Therefore, P_{25} falls in the 75-150 class.

Now, $P_{25} = 1 + \frac{\frac{25N}{100} - C}{f} \times i$

$= 75 + \frac{150-69}{167} \times 75 = 75 + \frac{81}{167} \times 75$

$= 75 + 36.38 = 111.38$ $\qquad P_{25} = 111.38$

P_{75} has 75N/100 items below it. It means 75 × 600/100 = 450 items below it. Therefore, P_{75} falls in the 225-300 class.

Now, $P_{75} = 1 + \frac{\frac{75N}{100} - C}{f} \times i$

$= 225 + \frac{450-443}{65} \times 75$

$= 225 + 8.077 = 233.077$ $\qquad P_{75} = 233.078$

Q_3 has 3N/4 items below it, which means 3 × 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 intervals

Now, $M_d = l + \frac{\frac{N}{2}-C}{f} \times i$

$= 150 + \frac{300-236}{207} \times 75 = 150 + \frac{64}{207} \times 75$

$= 150 + 23.19 = 173.19$ which is same as D_5

Therefore, Median is Rs. 173.19

b) Central 50% of observations are given by an interval Q_1 to Q_3 as Q_1 has 25% of items below it and Q_3 has 25% of items above it.

Here Q_1= P_{25} = Rs. 111.38 and Q_3 = P_{75} Rs. 233.08. Required limits of income of central Median 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_5 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.

Check Your Progress F

1) Define partition values. Name the partition values used in statistics.

2) Write the formulas for finding different partition values.

3) From the following data calculate Q_1, Q_3, D_4, P_{63}, P_{90} how many students have obtained less than 12 marks and how many have more than 95 marks.

Marks	No. of Students
0-12	40
12-23	85
23-38	75
38-45	50
45-60	65
60-73	60
73-83	75
83-95	35
95-100	15

13.9 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 a positional measure similar to median. **It is commonly denoted by M_o.**

For example, take the case of a shopkeeper who sells shoes or garments. He is interested to know the sizes of shoes or garments which are commonly demanded. Here 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. Similarly mode is also useful and appropriate average in problems related to the expression of preferences in a situation where it is not possible to measure in quantity. Such as design of garments, preferences on different advertisements etc. In such situations, we can consider the model preferences only for decision making but not arithmetic mean and median.

13.9.1 Computation of Mode

The method of computing mode is different for grouped data and ungrouped data. Now, let us study those methods separately.

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, 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 which is called **bi-model**. Similarly, the distribution can be a **tri-modal or even multi-model**. For such distributions, the mode as a measure of central tendency has little significance. Mode has very limited use for ungrouped data.

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, not only highest frequency but also 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** of the distribution.

Grouping Table: A grouping table has six columns as explained below:

Column 1 : It is of class frequencies written against each class and the highest frequency is masked or circled.

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 grouped in two’s their tables are obtained 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 from the top, remaining frequencies 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 by considering the highest tables (observation) in each column. It is two fold :

1) **vertical** (i.e., stubs) where the column numbers, as obtained in a grouping table, are taken.

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	10		15		
60	6		11		21	
61	5	15				35
62	10		30	52		
63	20	42			66	
64	22		46			52
65	24	30		32		
66	6		8		9	
68	2	3				
70	1					

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		

Now, look at the analysis tables, highest total table is five. The value of variable 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: In the case of continuous series, (i.e. data with class intervals) which have equal class intervals throughout, there be 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 model class, the value of mode (M_o) is obtained by interpolation by using any of the following formulas:

Formula 1: $M_o = l \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$

Where l = lower limit of the model class; i = class interval ; $\Delta_1 = f_1 - f_0$; $\Delta_2 = f_1 - f_2$; f_1 = is the frequency of the model class; f_0 = is the frequency of the model class preceding the model class; f_2 = is the frequency of the model class succeeding the model class.

By substituting the value of Δ_1 and Δ_2 in the above formula it can also be presented in the following forms:

i) $M_o = l + \frac{f_1 - f_o}{(f_1 - f_0) + (f_1 - f_2)} \times i$

ii) $M_o = l + \frac{f_1 - f_o}{2f_1 - f_0 - f_2} \times i$

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:

Formula 2: $M_o = l + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times i$

Where, $|f_1 - f_0|$ and $|f_1 - f_2|$ mean absolute values of the difference i.e., difference neglecting signs.

Where, the modal class is other than the one containing the maximum frequency, the following formula is more suitable:

Formula 3: $M_o = l + \frac{f_2}{f_0 + f_2} \times i$

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, l8 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 classes or land splitting classes and then apply the usual formula. The procedure will be explained in illustration 11.

4) If the distribution is given in inclusive method, the model class should be converted into 'exclusive' method. The procedure of this conversion is explained in the previous unit.

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	15
160-180	8
180-200	7
	100

Solution: By inspection the model 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		26		
40-60	9		20		34	
60-80	11	25				45
50-100	14		34	49		
100-120	20	35			45	
120-140	15		25			33
140-160	10	18		25		
160-180	8		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				
3				I	I				
4				I	I				
5					I	I			
6			I	I	I				
Total			1	3	6	1			

The highest total being 6, the modal group is 100-120.

Applying the formula: $M_o = l + \frac{f_1 - f_2}{2f_1 - f_0 - f_2} \times i$

$= 100 + \frac{20-14}{2(20)-14-15} \times 20$

$= 100 + \frac{6}{11} \times 20 = 100 + 10.91 = 110.91$

$\therefore$ mode of monthly rent is Rs. 110.91

Illustration 3: Calculate the mode from the following data:

Size	0-9	10-19	20-29	30-39	40-49	50-59
Frequency	3	4	8	7	6	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					
		7				
10-19	4			15		
			12			
20-29	8				19	
		15				
30-39	7					21
			13			
40-49	6			16		
		9				
50-59	3					

Analysis Table

Col. No.	Marks 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	

From 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:

$$M_o = l + \frac{f_2}{f_0 + f_2} \times i$$

$$= 29.5 + \frac{6}{8+6} \times 10$$

$$= 29.5 + \frac{60}{14} \qquad = 29.5 + 4.29 = 33.79$$

Therefore, mode is 33.8. You may note that a different result will be obtained if mode is calculated by the following formula:

$M_o = l + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times i$

$= 29.5 + \frac{|7-8|}{|7-8| + |7-6|} \times 10$

$= 29.5 + \frac{1}{1+1} \times 10 \quad = 29.5 + 10/2 = 34.5$

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 $M_o = l + \frac{f_1 - f_2}{2f_1 - f_0 - f_2} \times i$ denominator will become zero and the numerator will be negative and therefore, this formula is not applicable. It is important to note at unlike arithmetic mean and median, the different methods of calculating mode can give different results.

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 etc. frequency are equal, the mode is the mid-point of 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	6	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 $f_0 = f_2$ for both f_0 and f_2 less than f_1 this will always happen. When $f_0 \neq f_2$ (i.e., the two frequencies neighbouring the modal 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:

x:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f:	500	610	740	748	745	690	500

Here the modal class is 30-40 corresponding to the highest frequency 748 (f_1). 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 medal class is 30-40, 'l' is 30 and u is 40.

$\therefore M_o \quad = \frac{30 \times 740 + 40 \times 745}{740 + 745}$

$$= \frac{52{,}000}{1{,}485} \qquad = 35.02$$

The result derived by this method will always be the same as obtained by using the formula : $M_o = l + \frac{f_2}{f_0 + f_2} \times i$. **You may verify it.**

Illustration 4: For 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

Age	Col.1	Col.2	Col.3	Col.4	Col.5	Col.6
20-25	50	120		200		×
25-30	70		150		330	
30-35	80	260				410
35-40	180		330	450		
40-45	150	270			340	
45-50	120		190			240
50-55	70	120				
55-60	50					

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 $M_o = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

$$M_o = 40 + \frac{150-180}{2\times150 - 180 - 120} \times 5 = 40 + \frac{-30}{0} \times 5$$

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:

$$M_o = l + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times i$$

$$= 40 + \frac{|150-180|}{|150-180| + |180-120|} \times 5$$

$$= 40 + \frac{30}{30+60} \times 5$$

$$= 40 + \frac{5}{3} \qquad = 40 + 1.67 = 41.67 \qquad \therefore \text{Modal age} = 41.67$$

Illustration 5: Find the mode from the following table:

Size of the Item	40-49	50-59	60-69	70-79	80-89	90-59	100-109
Frequency	7	9	10	6	13	10	12

Solution: By inspection, the modal class is not clear. Hence, we have to do grouping and analysis.

Grouping Table

Age	**Col.1**	**Col.2**	**Col.3**	**Col.4**	**Col.5**	**Col.6**
40-49	7	16				
50-59	9		19	26		
60-69	10	16			25	
70-79	6		19			29
80-89	13	23		29		
90-99	10		22		35	
100-109	12	19				29
110-119	7					

Analysis Table

Col. No.	Marks					
	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	I
Total	1	2	5	5	3	1

In the analysis table maximum total occurs twice. The mode, therefore, is ill-defined and is to be determined empirically by using the formula:

$M_o = 3M_d - 2\overline{X}$. You may check yourself that here Median = 83.84 and $\overline{X}$ = 80.14.

$\therefore M_o$ = 3 (83.84) – 2(80.14).

= 251.52 – 168.28 = 91.24 ∴Mode = 91.24.

Check Your Progress G

1) Define mode.

2) State the various formulas for the computation of mode

3) What is 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 mode.

5) Calculate mean, median and mode from the following data:

X	20-40	40-60	60-80	80-100	100-120	120-140	140-160	160-180	180-200
F	6	9	11	14	12	15	10	8	7

13.9.2 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 significantly by 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.

13.9.3 Some Illustration

Illustration 6: 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:

$$M_o = 3M_d - 2\overline{X}$$

Substituting the value of M_o, and M_d

$$15.3 = 3 \times 14.2 - 2\overline{X}$$

$$2\overline{X} = 42.6 - 15.3$$

$$2\overline{X} = 27.3$$

$$\overline{X} = \frac{27.3}{2} = 13.65$$

Illustration 7: With the help of empirical relation between M_o, M_d and $\overline{X}$ shows that

i) $M_d = M_o + \frac{2}{3}(\overline{X} - M_o)$

ii) $\overline{X} = M_d + \frac{1}{2}(M_d - M_o)$

The empirical relation between mean, median and mode is:

i) $M_o = 3M_d - 2\overline{X}$

$$M_o - 2\overline{X} = 3M_d$$

$$\frac{1}{2}(M_o - 2\overline{X}) = 3M_d$$

$$\frac{1}{2}(M_o - 2\overline{X}) = 3M_d$$

$$M_d = \frac{1}{3}M_o + \frac{2}{3}\overline{X}$$

$$= M_o - \frac{2}{3}M_o + \frac{2}{3}\overline{X}$$

$$= M_o + \frac{2}{3}(M_o - \overline{X})$$

$$= M_o + \frac{2}{3}(\overline{X} - M_o)$$

$$\therefore M_d = M_o + \frac{2}{3}(\overline{X} - M_o)$$

$$Median = Mode + \frac{2}{3}(\text{Mean} - Mode)$$

ii) $M_o - 2\overline{X} = 3M_d$

$$2\overline{X} = 3M_d - M_o$$

$$\overline{X} = \frac{3}{2}M_d - \frac{1}{2}M_o = M_d + \frac{1}{2}(M_d - M_o)$$

$$Mean = Median + \frac{1}{2}(Median - Mode)$$

Illustration 8: Finding the missing frequency

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.

$$\text{Now, } M_o = l + \frac{f_1 - f_2}{2f_1 - f_0 - f_2} \times i$$

Where, l = 30, f_0= F, f_1= 18, f_2= 9, i = 5 and M_0= 32

Substituting the x values:

$$32 = 30 + \frac{18 - F}{2 \times 18 - F - 9} \times 5$$

$$2 = \frac{18 - F}{27 - F} \times 5$$

$$54 - 2F = 90 - 5F$$

$$3F = 36$$

$$F = 12 \therefore \text{Missing frequency is 12.}$$

Illustration 9: Unequal class interval

Calculate mode from the data given below:

Profit (Rs. in lakhs)	0-5	5-10	10-20	30-40	40-50
No. of Companies	4	6	15	18	20

Solution: Here the class interval 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

$$\text{Now, mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Substituting the values of l, f_0, f_1, f_2, i

$$M_0 = 20 + \frac{18-15}{2 \times 18 - 15 - 10} \times 10$$

$$= 20 + \frac{3}{11} \times 10 \qquad = 20 + 2.7 - 22.7$$

∴ 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

$$= 3 \times 23.6 - 2 \times 24.3$$

$$= 70.8 - 48.6 \quad = 22.2$$

∴ Mode of profit is Rs. 22.2 lakhs.

Illustration 10: 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 x 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,

13.10 CHOICE OF SUITABLE AVERAGE

Starting from Unit 13, we have discussed various averages viz., mean, geometric mean, median, partition values and mode 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 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 arithmetic mean will be an appropriate 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 if it. When the ratios or percentages are to be averaged, geometric mean is the most appropriate measures. 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 proper average.

5) **Special Purpose**: For calculating trend in time-series analysis, the moving average would be 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 give the higher value. Since we can select the measures of central tendency to sit our fancy, there is a possibility of selecting the average which produces the result we want. When use unscrupulously or incompetently, the user is at fault not the tool.

13.11 LET US SUM UP

The main characteristics of the data are represented by a single figure known as 'an average' or 'a main'. It is the point of location around which individual values cluster. An ideal average must satisfy certain properties such as case of calculation, rigidity in its dyination, should be based on all items, should remain unaffected by extreme items, should be capable of further algebraic treatment and should have sampling stability. On 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 mean for ungrouped and grouped data. When value in the data set are unequal importance, a weighted arithmetic mean will be a truly representative average.

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 use. Geometric mean is computed for both ungrouped and grouped data (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.

The data set 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 rations involving speed, time and distance. It is the reciprocal of the arithmetic mean of reciprocals of the individual observations. It can be computed fro ungrouped and grouped data. Like weighted geometric mean weighted harmonic mean also can be calculated.

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. Like median, there are other optional 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. Almost similar procedures is followed in the computation of the partition values, as prescribed for median.

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 interpolative formulas.

The choice of a suitable measure of central tendency depends on the purpose that an average designed to serve as the nature and the form of the data set, its amenability to further algebraic central tendency, however, is to be made cautiously and competently.

13.12 KEY WORDS

Key Words:

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.

Central Tendency: A single value that has a tendency to be somewhere at the centre and within the range of all values.

Deciles: The values of the variate that divide the series or distribution into ten equal parts.

Empirical Relationship of averages: The relationship that exists between average in a moderately skewed distribution viz, $M_o = 3Md - 2\bar{X}$

Extreme Values: The items that are too big or too small in comparison with the other terms of data. The unduly influence the mean.

Geometric Mean: If there are N items in the series, the geometric mean is the Nth root of their product.

Grouping Table: The table which has six columns, used for ascertaining a modal class.

Harmonic Mean: The reciprocal of the arithmetic mean of reciprocal of the individual observations.

Mean: The value obtained by dividing the sum of value of act 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.

Median: The value of the variate that divides the series into two equal parts.

Mode: The value of the variate around which the other items tend to concentrate most heavily.

Partition Values: The values of the variate that divide the distribution into a fixed number of equal parts.

Percentiles: The value of the variate that divide the series or distribution into 100 equal parts.

Positional Average: An average based on the position of a given derivation 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.

Weighted Arithmetic Mean: An average whose component items are assigned weights according to their relative importance.

13.13 ANSWERS TO CHECK YOUR PROGRESS

A) 1) i) 11; ii) $\bar{X} = \frac{\Sigma fm}{\Sigma f}, \bar{X} = A + \frac{\Sigma fd}{n}, \bar{X} = A + \frac{\Sigma fd}{n} \times c$

iv) 31

2) Rs. 164.33

3) i) 68; ii) 0

4) Rs. 128.33 by both methods

5) 40.2

B) 2) Simple Average = 42.92; Weighted Average = 44.23

3) Both are equal to 73.7%

4) 34.47

C) 1) 12.3% approximately

2) GM = 25.3 marks, AM = 28.4 marks

4) 29%

D) 2) HM = 50.55

3) 13

4) Rs. 14.63

5) 75.45 km.p.h.

E) 1) (a) 16, (b) 0.18

2) $\mathrm{M} = \mathrm{U} - \frac{\frac{N}{2} - C}{f}$

3) Class interval of median class is considered.

4) First Case 63. Second Case 64.

5) 30

6) 210.5

F) 3) 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.

G) 4) 30.1

5) $\bar{X}$ = 110, Median = 110, Mode = 110.9

13.14 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 averageis preferable to a simple average?
4) Compare arithmetic mean, geometric mean and harmonic mean in point out their relative merit and limitations.
5) How do you make a choice of suitable measure of central tendency?
6) What is median? Explain its merits and limitations.
7) Explain the methods of computing median.
8) Compare the arithmetic mean and median as measures of average?
9) Compare and contrast between Quartiles, Deciles and Percentiles?
10) 'Arithmetic Mean Median and Mode all try to give one main characteristic of the data but in their own way'. Discuss.
11) What is mode? Explain its limitations and uses as a measure of 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.

Labour	Bombay		Kolkatta	
	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.: Rs. 28.3)

5) The following table shows the number of workers in various trade categories who workedfrom Monday to Friday in a week for varyingnumber 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 workers for the whole week for all categories together.

Categories	**Number of Wrokers**				
	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 as 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) 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?
(Ans.: No. 1.035%)

8) 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.
(Ans.: 5.9 per annum)

9) Geometric mean of 8 items is 3 and geometric mean of 12 items is 11. What will be the geometric mean of all 20 items?
(Ans.: Rs. 6.54)

10) 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

(Ans.: i) 3.184; ii) 4.505)

11) You take a trip which entails travelling 900 miles by train at an average speed of 60 km.p.h.; 3000 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?
(Ans.: Rs. 31.6 km.p.h.)

12) 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?
(Ans.: Median 97.5)

13) 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

(Ans.: Approximately Rs. 402)

14) 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

(Ans.: Rs. 50)

15) Calculate: i) median from the following data and ii) obtain the range of marks obtained by middle 80%of the students.

Marks	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

(Ans. I) Rs. 27.5 ii) 11.7 to 47.1)

16) Find the missing frequency if median is 25.

Marks	0-10	10-20	20-30	30-40	40-50
No. of Students	14	-	27	-	15

(Ans.: 23, 21)

17) A laundary 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 the same price, which brand should be purchased by the laundry.

(Ans.: Rs. 6.54)

18) Calculate Q_1, P_{30}, D_8 from the given data given below:

Size of collar worm	:	14”	14.5”	15:	15.5”	16”
No. of Students	:	20	37	43	26	14

(Ans.: Q_1 = 14.5”, P_{30} = 14.5”, D_8= 15.5”)

19) Calculate the values of D_6, Median, P_{20}, Q_1, and Q_3 from the following data.

Marks	**No. of Students**
Below 10	
10-20	25
20-30	40
30-40	70
40-50	90
50-60	40
60-70	20
Above 70	

(Ans.: D_6= 44.4, Median = 41.1, P_{20} = 27.5, Q_1, = 30.7; Q_3 = 49.4)

20) 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

(Ans.: 18 years)

21) The following tables gives the relative distribution of sales calls made on Amar Pharmaceuticals in the past months. Fin the modal calls.

No. of Sales Calls	0	1	2	3	4	5
Related Frequency	0.21	0.18	0.38	0.19	0.03	0.01

(Ans.: 2 sales calls)

22) 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

(Ans.: 71.34)

23) Estimate the median when arithmetic mean is 27.9 and mode is 25.2. Give the assumption, if any.?

(Ans.: 27)

24) Calculate Mode from the following distribution:

Class		10-20	20-30	20-24	24-30	30-50	50-52	52-60
Frequency		4	9	6	9	52	2	8

(Ans.: 40)

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. 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.** 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 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 the numbers which are less the characteristic is equal to one more than the number of zeros after the decimal point and before any significant digit. Thus, characteristic of 0.98 is −1,10.098 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 1 (the fourth digit in the given number) in the same row. The value is 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 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 the mantissa 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	1.3892
0.0245	12.3892
0.00245	13.3892

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. **Anti Logarithms**: As you know the logarithm tables give the value of mantissa in the logarithms of Whereas the antilog tables give 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 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, Look at the antilog tables at the row corresponding to .38 and column corresponding to number is 2449. Look at the mean 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 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 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.

FURTHER READINGS

Arora, P.N. Sumeet Arora and Arora. A., 2007, *Comprehensive Statistical Methods.* S. Chand and Company Ltd., New Delhi.

Beri, G.C., 2005, *Business Statistics,* Tata Mc Graw-Hill Publishing Company, Ltd., New Delhi.

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.

Simpson, G, and.Kafka, F. Basic Statistics, Oxford & IBH Publishing 1 New Delhi.

UNIT 14 MEASURES OF DISPERSION

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.
- define & compute standard deviation and narrate its properties, means and limitations
- define and compute variance and coefficient of variation for different kinds of data
- compare different measures of dispersion and use them at appropriate situations

14.1 INTRODUCTION

In Unit 13 you have studied about different measures of central tendency. In the unit you studied, the measures of central tendency give us one single value that represents the entire data. But central tendency alone is not sufficient to analyse different characteristics of the data unless all the observations are having the same value. For more meaningful analysis of the data, it is necessary to study dispersion i.e., the spread of the data or the extent to 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. Besides these, you will also learn about the month of computing standard deviation and its coefficient for different kind of data, their merits and uses.

14.2 CONCEPT OF DISPERSION

In order to understand the concept of dispersion, let us consider some important definition of dispersion.

- The Degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data (Spiegel)
- The measurement of scatterness of the mass of figures in a series about an average is called measures of variation or dispersion (Simpson and Kafka)
- Dispersion or spread is the degree of the scatter or variation of the variables about a central value (Brook and Dick)

From the above definitions, it is clear that the word, dispersion (**also termed as variation or spread or scatter**) 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. It gives an average of the differences of various observations from a central value (an average). Thus, the significance of an average is determined in the light of dispersion. In order to understand the concept of dispersion more clearly, study the following Illustration 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	22,000
60,000	56,500	18,000
60,000	60,500	27,000
60,000	68,250	2,10,000
$\overline{X}_A = 60{,}000$	$\overline{X}_B = 60{,}000$	$\overline{X}_C = 60{,}000$

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 of 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 of 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 mean sales will be smaller for Firm B compared to Firm C. In other words, Firm B has smaller dispersion than Firm C. In firm A, there is no dispersion.

14.3 SIGNIFICANCE OF MEASURING DISPERSION

Measures of dispersion (variations) are calculated to serve the following purposes:

1) Measuring variability determines the reliability of an average by pointing out to what extent the average is representative of the entire data. In Illustration 1 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, etc.

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 methods should produce the same answer. Different methods of computation leading to different answers are not proper.

3) It should be based on all items. Where 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., 16,000/60,000 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., $\frac{16,000}{52,250+68,250}$. 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 kilometre (1,00,000 cm) in measuring distance from Delhi to Mumbai is hardly of any significance. But a variation of 10 cm in measuring a piece o f cloth of 1.40 meters is of very great significance. So, whenever comparisons of variability in two sets of data are done, it is always done in terms of relative measures.

Check Your Progress A

1) What is the meaning of the term Dispersion?

2) Differentiate between absolute measures and relative measures of dispersion.

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 (This, however, is not a part of discussion in this course).

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 and Standard Deviation.**

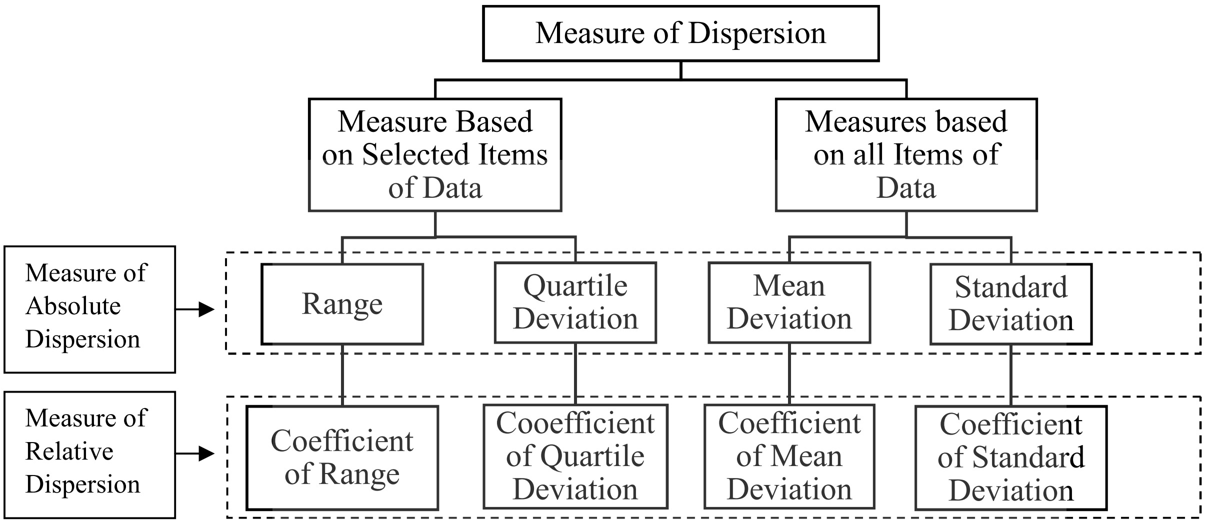


Figure 14.1: Classification of Measures of Dispersion

14.6.1 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 (section 14.2), 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 = $68{,}250 - 52{,}250 = 16000$; 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 cases of Firm B, and the variation is very large in case of Firm C.

For grouped data, the range may be determined, in discrete series, as the difference between the highest value and lowest value of the observation. In case of continuous series, 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:

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 Days
30 - 40	12
40 - 50	18
50 - 60	20
60 - 70	19
70 - 80	13
80 - 90	8

Solution: Range $= X_{max} - X_{min}$

X_{max} = upper limit of largest class interval

X_{min} = lowest limit of smallest class interval

$= 90 - 30$

$= 60$

Coefficient of Range $= \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$

$$= \frac{90 - 30}{90 + 30}$$

$$= \frac{60}{120}$$

= 0.5.

You should note that the frequency of the distribution should not be taken into account for computing range.

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. It fails to disclose the characteristics of the distribution and it is not applicable in case of open-end distribution.

Applicability: The concept of range is extensively used in statistical quality control. Range is helpful in studying variations in the prices of shares, debentures and agricultural commodities which are very sensitive to price changes. The range is a good indicator for weather forecast.

14.6.2 Quartile Deviation

Quartile deviation is defined as half the difference between the upper quartile and lower quartile. You have already studied the methods of computing Quartiles in Unit 13 at partition values.

Quartile Deviation $= \frac{Q_1 - Q_3}{2}$

Where, Q_1 is the first quartile (lower quartile) and Q_3 is the third quartile (upper quartile).

To understand the procedure for computing Q_1 and Q_3, you are advised to refer unit 13 once again where we discussed the methods of computing quartiles under the section Partition Values.

As the difference between Q_1, 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 quartiles, and does not 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:

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 from ungrouped and grouped data.

Ungrouped Data (Individual Observations):

Illustration 3: Calculate the value of quartile deviation and its coefficient from the following data relate the marks obtained by 7 students.

Marks :	40	10	26	32	15	49	25

Solutions : As we discussed in unit 13 (Quartiles) we have to arrange the value of variables either in ascending or in descending order. Here, the marks are arranged in ascending order as follows:

Marks :	10	15	25	26	32	40	49

Q_1= size of $\frac{N+1}{4}$ th observation = size of $\frac{7+1}{4}$ = 2^{nd} observation

The size of 2^{nd} observation = 15 marks; Q_1= 15.

Q_3= size of $3\left(\frac{N+1}{4}\right)^{th}$ observation = size of $3\left(\frac{7+1}{4}\right)$ = 6^{th} observation

The size of 6^{th} observation = 40 marks; Q_3= 40.

Quartile Daviation (Q.D) = $\frac{Q_3 - Q_1}{2} = \frac{40-15}{2}$ = 12.5 Marks

Coefficient of Q.D = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40-15}{40+15}$= 0.45

Illustration 4: 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 $\left(\frac{N+1}{4}\right)^{th}$ = 8^{th} observation

= 62 kgs. (because 5^{th} observation falls in this category as it lies in 9 cumulative frequency)

Q_3= size of 3 $\left(\frac{N+1}{4}\right)^{th}$ observation = 24^{th} observation

= 65 kgs. (because 24^{th} observation falls in this category as it lies in 26 cumulative frequency)

Quartile Deviation $= \frac{Q_3 - Q_1}{2} = \frac{65-62}{2} = 1.5$ kgs.

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{65-62}{65+62} = \frac{3}{127}$

$= 0.024.$

Continuous distribution

Illustration 5: 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 $\frac{N}{4}$ th observation i.e., $\frac{204}{4}$ = 51 th observation. 51th observation which lies in 54 cumulative frequency. 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 the lower quartile class; c = cumulative frequency of the class proceedings to the lower quartile class; f = simple frequency of the lower quartile class; i = class-interval of the lower quartile class

$$Q_1 \quad = 20 + \frac{51 - 29}{25} \times 10$$

$$= 20 + 8.8 = 28.8$$

Q_3 has $\frac{3N}{4}$ th observation i.e., $3 \times \frac{204}{4} = 153$ th observation. 153th observation which lies in 167 cumulative frequency. So, Q_3(upper quartile) class is 60-70 class.

$$Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times i$$

Here, the value of l, c, f and i are relate to the upper quartile (Q_3)

Thus, $Q_3 = 60 + \frac{153-145}{22} \times 10 = 63.64.$

Semi-inter Quartile Range or Quartile Deviation is given by:

Q.D $= \frac{Q_3 - Q_1}{2} = \frac{63.64 - 25.8}{2} = \frac{34.84}{2}$ = 17.42 Marks

The relative measure corresponding to quartile deviation, called the coefficient of quartile deviation, is calculated as follows:

Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63.64-28.8}{63.64+28.8}$ = 0.37 marks

Illustration 6: 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
	N = 300	

Q_1has $\frac{N}{4}$ th observation i.e., $\frac{300}{4}$ = 75 th observation, which lies in 67 cumulative frequency. So Q_1lies in the 950-1,000 class.

$$Q_3 = 1 + \frac{\frac{3N}{4} - c}{f} \times i$$

$$= 950 + \frac{\frac{300}{4} - 67}{56} \times 50$$

= Rs. 957.14.

Q_3 has $\frac{3N}{4}$ th observation i.e., $\frac{3 \times 300}{4} = 225$ th observation which lies in 260 cumulative frequency. So, Q_3 class in the class 1050-1100.

$$Q_3 = 1 + \frac{\frac{3N}{4} - c}{f} \times i = 1{,}050 + \frac{225 - 185}{75} \times 50 = \text{Rs. } 1{,}076.67.$$

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{1{,}076.67 - 957.14}{2} = \text{Rs.}59.760.$$

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, 57, 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

14.6.3 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. This is the reason these two methods of dispersion do not show the scatterness around of central value. 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. Hence, these deviation should be written as |D| (it is pronounced as 'Modules D'). Hence, |D| means deviation are taken from an average by ignoring their actual signs. It is, therefore, also called mean absolute deviation. **An important property of Mean Deviation (M.D.) 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 computing from median, the coefficient of mean deviation shall be obtained by dividing the mean deviation by the median.

Coefficient of M.D. about $\mathrm{M_d} = \frac{\text{M.D.about Median}}{\text{Median}}$

Similarly a coefficient of M.D. about Mean$(\overline{X}) = \frac{\text{M.D.about } \overline{X}}{\overline{X}}$

You should keep in mind that the procedure to compute mean deviation from ungrouped and grouped sets of data is different, but computing coefficient of mean deviation is the same.

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 algebraic signs and taking absolute deviations makes it difficult to be treated algebraically. Although, the average deviation is a good measure of variability, its use is limited. If one desires only to measures and compare variability among several sets of data, the average deviation may be used. The computation of mean deviation from different sets of data will become clear if you study the following illustration carefully.

Calculation of mean deviation – Ungrouped data

Formula = M.D. = $\frac{\sum|\mathrm{D}|}{n}$

Where, $\sum$|D| = Sum of the deviations (ignoring signs) from an average

n = Number of observations or items.

Procedure to compute: 1) Compute an average (mean or median or mode); 2) Find the absolute deviations of the value of observations from the chosen average (in step 1) i.e. |D| and obtain the total of |D|, i.e., $\sum$|D|; 3) Obtain the total number of observations (n); 4) Apply the formula.

Illustration 7: Calculate the mean deviation and its co-efficient from the following values about the Mean, median and mode:

18, 25, 63, 59, 29, 72, 17, 25, 105, 87.

Solution: Calculation of $\overline{X}$, M_d and M_o :

Mean $(\overline{X}) = \frac{\sum x}{n} = \frac{500}{10} = 50$

Median: Since there are ten observations which is an even number, the median is the average of the two middle most observations, when arranged in order of magnitude as follows:

17, 18; 25, 25, 29, 59, 63, 72, 87, 105

Median (M_d) = Size of $\left(\frac{N+1}{2}\right)^{th}$ item = $\left(\frac{10+1}{2}\right)$= 5.5th item

$$(M_d) = \left(\frac{29+59}{2}\right) = 44$$

Mode $(\mathbf{M_o})$ = 25, since it appears maximum number of times in the distribution.

Calculation of Mean Diviation about mean, median and mode

X	Deviation from Mean (50) |D|	Deviation from Median (44) |D|	Deviation from Mode (25) |D|
18	32	26	7
25	25	19	0
63	13	19	38
59	9	15	34
29	21	15	4
72	22	28	47
17	33	27	8
25	25	19	0
105	55	61	80
87	37	43	62
N = 10	**$\sum$|D| = 272**	**$\sum$|D| = 272**	**$\sum$|D| = 280**

Mean Deviation about mean $= \frac{\sum|D|}{n}$

$= \frac{272}{10} = 27.2$

Coefficient of M.D. $= \frac{M.D.}{\overline{X}} = \frac{27.2}{50} = 0.544$

Mean deviation about Median $= \frac{\sum|D|}{n} = \frac{272}{10} = 27.2$

Coefficient of M.D. $= \frac{M.D.}{Median} = \frac{27.2}{44} = 0.62$

M.D. about mode $= \frac{\sum|D|}{n} = \frac{280}{10} = 28$

Coefficient of M.D. $= \frac{\text{M.D.}}{\text{M}_o} = \frac{28}{25} = 1.12$

Calculation of Mean Deviation – Group Data (Discrete Series)

M.D. $= \frac{\sum f|D|}{n}$

Where, $\sum f|D|$ = sum of the products, which is obtained by multiplying the absolute deviations (ignoring signs) with its corresponding frequencies.

N = Number of items or total of the frequency

Steps to solve the problem: 1) Compute average ($\overline{X}$ or M_d or M_o); 2) Take the deviations of the value of items from the average ignoring algebraic (±) signs and denote them |D|; 3) Multiply these deviations with there respective frequencies and obtain the total i.e., $\sum f|D|$; 4) Obtain the total of frequency (n) 5) Apply the formula

Illustration 8: Find Mean Deviation about Mean and Median, and their coefficients.

Marks :	20	30	40	50	60	70
No. of students :	8	12	20	10	6	4

Solutions: Calculation of Mean deviation and its coefficient about $\overline{X}$ and M_d.

Marks X	No. of Students f	fx	cumulative frequency (c.f.)	deviations about Mean (41) |D|	f|D|	Deviation about Median (40)|D|	f|D|
20	8	160	8	21	168	20	160
30	12	360	20	11	132	10	120
40	20	800	40	1	20	0	0
50	10	500	50	9	90	10	100
60	6	360	56	19	114	20	120
70	4	280	60	29	116	30	120
	N=60	**∑fx=2,240**			**∑f|D|=640**		**∑fd=620**

Mean ($\overline{X}$) $= \frac{\sum fx}{n} = \frac{2{,}240}{60} = 41$marks

Median (M_d) = Size of $\left(\frac{N+1}{2}\right)^{th}$ item

$= \left(\frac{60+1}{2}\right) = 30.5^{th}$ item

Size of 30.5^{th} item lies in the 40^{th} item of the cumulative frequency, its corresponding value is 40 marks.

Hence, Median = 40 marks

Mean Deviation about mean $= \frac{\sum f|D|}{n} = \frac{640}{60} = 10.67$ marks

Coefficient of M.D. (about Mean) $= \frac{\text{M.D.}}{Mean} = \frac{10.67}{41} = 0.26$

Mean deviation about Median $= \frac{\sum f|D|}{n} = \frac{620}{60} = 10.33$ marks

Coefficient of M.D. (about Median) $= \frac{\text{M.D.}}{Median} = \frac{10.33}{40} = 0.26$

Here, you may notice that, as we discussed earlier, the mean deviation about median is least.

Illustration 9: From the following grouped data relating to the sales of 100 Companies, find the Coefficient of Mean Deviation by using mean ($\overline{X}$) .

Sales(Rs.'000)	40-50	50-60	60-70	70-80	80-90	90-100
No. of Companies	5	15	25	30	20	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**		**∑fX = 7,100**		**∑f|D|= 1,040**

($\overline{X} = \frac{\sum fx}{n} = \frac{7,100}{100} = 71$

Mean Deviation (about mean) $= \frac{\sum f|D|}{n}$

$= \frac{1,040}{100} = 10.40$ 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 10: 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 Mean Deviation from Median**

Age-Group (in Years)	Frequency (f)	Cumulative Frequency (Cf)	Class Mid-Point (M)	$\|X-M_d\|$ i.e., $\|X-31.5\|$ $\|D\|$	$f\|X-M_d\|$ $f\|D\|$
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**				**$\sum f\|d\|$ =582.0**

Median has $\frac{N}{2} = \frac{80}{2} = 40th$ observation. So it lies in the 56 cumulative frequency and its corresponding class interval is 31-35. Converting into exclusive class 30.5–35.5 Median = $l + \frac{\frac{N}{2}-c}{f} \times i = 30.5 + \frac{40-35}{20} \times 5$ = 31.5 years

Mean Deviation (About Median) = $\frac{\sum f|D|}{n} = \frac{582}{80} = 7.275$ years

Coefficient of Mean Deviation about median= $\frac{\text{M.D.about } M_d}{M_d}$

$= \frac{7.275}{31.5} = 0.23.$

14.6.4 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 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.

Thus, the standard deviation is defined as the position square root of the variance. This concept was introduced by Karl Pearson in 1893. It is widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer. As this measure is calculated by finding square root of the mean of the squares of the deviation of items from the arithmetic mean, it is also called Root mean Square Deviation. Standard Deviation is usually denoted by the Greek letter 'σ' (read as sigma). Now, let us study the meaning, method of computation, merits and limitations of standard deviation.

Computation

There are two methods of calculating standard deviation for ungrouped and grouped distributions. They are: 1) direct method and 2) short cut method. Let us study these two methods.

1. **Direct method**: under this method, standard deviation is calculated by taking deviations of the items from the actual arithmetic mean of the distributions.

2. **Short-Cut method**: Under this method, standard deviation is calculated by taking deviations of the items from the assumed mean.

Among the above two methods, short cut method is convenient when the size of items and their numbers are large or the arithmetic mean comes out in fractions. If mean is with fractional value, it is a time consuming process to find the deviations and its square deviations to compute the standard deviation.

Let us study the formulas and consider some illustrations to understand the procedure involved in direct and short –cut methods and computation of the standard deviation under ungrouped and grouped distributions.

Ungrouped data (individual distribution): ***Direct Method***

Formula: $\sigma = \sqrt{\frac{\sum d^2}{n}}$

Where, σ = Standard deviation; $\sum d^2$= Sum of the squared deviation from actual mean; N = Number of items.

Steps for computing standard deviation by direct method:

1) Calculate: the arithmetic mean of the data ($\overline{X}$); 2) Take the deviations by subtracting the arithmetic mean from each and every value of the items (X - $\overline{X}$). Denote it by 'd'; 3) Square the deviations (d^2) and obtain the total i.e., $\sum d^2$; 4) Obtain the number of items (n) and apply the formula.

Short-cut Method: Formula: $\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$

Where, σ = Standard deviation; $\sum d$ = Sum of the deviation from assumed mean; $\sum d^2$ = Sum of the squared deviation from assumed mean; n = Number of items.

Steps to compute standard deviation by short-cut method:

1) Take a balanced value from the given data as assumed mean, and calculate the deviation of the items from the assumed (X – Assumed mean). Denote these deviations by 'd' and obtaing the total i.e. $\sum d$; 2) Square the deviation (d^2) and obtain the total i.e. $\sum d^2$; 3)Take the number of items (n) and apply the formula

Examine Illustration-11 carefully to understand the procedure involved in the calculation of standard deviation by direct method as well as short-cut Method.

Illustration 11

Calculate standard deviation for the following distribution by direct method and short-cut method.

Serial No. of Workers :	1	2	3	4	5	6	7	8	9	10
Wage (Rs.) :	20	22	27	30	31	32	35	45	40	48

Solution : Direct Method : Calculation of Standard Deviation. Here deviations are taken from Actual Mean.

S.No. of workers	Wages (Rs.) X	$(X - \overline{X})$	$(X - \overline{X})^2$ d^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
n = 10	**$\sum X = 330$**		**$\sum d^2 = 762$**

Now, $\overline{X} = \frac{\sum X}{n} = \frac{330}{10} = 33$

$$\sigma = \sqrt{\frac{\sum d^2}{n}} = \sqrt{\frac{762}{10}} = \sqrt{76.2} = \text{Rs. } 8.73$$

Short-cut Method : calculation of Standard Deviation. Here, Rs. 32 is taken as assumed mean.

S.No. of workers	Wages (Rs.) 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
n = 10		**$\sum$d = 10**	**$\sum d^2$ = 772**

Now, σ $= \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$

$= \sqrt{\frac{772}{10} - \left(\frac{10}{10}\right)^2} = \sqrt{77.2 - 1} = \sqrt{76.2}$ = Rs. 8.73

You may note that the results obtained by both the methods are the same. You should also note that if you take any value as assumed mean, you would get the same result. To test this , you are advised to find the standard deviation by choosing different values of assumed mean say 27, 35, ….. so on.

Grouped Data-Discrete distribution: ***Direct method***

Formula for σ = $\sqrt{\frac{\sum f d^2}{n}}$

Where $\sum f d^2$is the sum of the products, which obtained by multiplication of the squared deviations from actual mean with it respect frequencies; and n is the number of the item (sum of the frequency).

Steps to compute S.D. for discrete series by direct method:

1) Calculate Arithmetic mean of the series; 2) Take the deviations of the items from the arithmetic mean (x); 3) Square the deviation (X^2); 4) Multiply the squared deviations with their corresponding frequencies (fX^2); 5) Obtain the total of the frequency (n) and apply the formula.

Short-cut Method:

Formula = σ = $\sqrt{\frac{\sum f d^2}{n} - \left(\frac{\sum f d}{n}\right)^2}$

Where, $\sum f d^2$is the sum of the products which obtained by the multiplication of squared deviation from assumed mean (d^2) by its respective frequency; $\sum f d$ is the sum of the products which obtained by the multiplication of the deviations from assumed mean to its respective frequencies; and n is the total of the frequency.

Steps to compute S.D. for discrete series by short-cut method:

1) Select a value as assumed mean and take deviations of the items from the assumed mean (X-A) denote these deviations by 'd'; 2) Square the above deviation (d^2); 3) Multiply the deviations with its corresponding frequency (fd) and obtain ($\sum$fd); 4) Multiply the squared deviation (d^2) with its corresponding frequency and denote it as fd^2; 5) Obtain the total (i.e., $\sum f d^2$); 6) Obtain the total of the frequency (n) and 7) apply the formula.

Study Illustration 12 carefully to understand the procedure clearly.

Illustration 12: Calculate the standard deviation from the following frequency distribution by direct and short-cut methods using 14 as assumed mean.

Daily wages (Rs.) :	10	12	14	16	18	20	22
No. of workers :	3	5	9	16	8	7	2

Solution: Calculation of Standard Deviation and Variance: **Direct Method**

Daily Wages (Rs.) X	**No. of Workers f**	**fX**	**d ($X - \overline{X}$) $\overline{X}$=16**	d^2	fd^2
10	3	30	-6	-36	108
12	5	60	-4	-16	80
14	9	126	-2	-4	36
16	16	256	0	0	0
18	8	144	2	4	32
20	7	140	4	16	112
22	2	44	6	36	72
	n = 50	**∑fX = 800**			$\sum fd^2 = 440$

Now, $\overline{X} = \frac{\sum fX}{n} = \frac{800}{50}$ = Rs. 16

$\sigma = \sqrt{\frac{\sum fd^2}{n}} = \sqrt{\frac{440}{50}} = \sqrt{8.8}$ = Rs. 2.97

Short-Cut Method: Taken Assumed Mean as 14.

Daily Wages (Rs.) X	**No. of Workers 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
	n = 50		**∑fd = 100**	$\sum fd^2 = 640$

Now, σ $= \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$

$= \sqrt{\frac{640}{50} - \left(\frac{100}{50}\right)^2}$

$= \sqrt{12.8 - 4} = \sqrt{8.8}$ = Rs. 2.97

You may note that when arithmetic mean is in whole numbers, there is not much simplification in calculations by short-cut method.

Continuous distribution: ***Direct Method***

Formula: $\sigma = \sqrt{\frac{\sum fd^2}{n}}$

Where, $\sum fd^2$ is the sum of the products, which obtained by multiplying the squared deviations (taken from actual mean to midvalues) with its respective frequencies; n = total of the item

Steps to compute S.D. for continuous series by direct method:

1) Find out the mid values; 2) Compute of the arithmetic mean; 3) Take the deviations of the mid values form the arithmetic mean (M-X) i.e., d; 4) Square the deviation i.e., d^2; 5) Multiply the squared deviations (d^2)with its respective frequencies (f) and obtain the total i.e., $\sum fd^2$; 6) Obtain the total of items (n) and apply the formula.

Short Cut Method

Formula : $\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$

Where $\sum fd^2$ = sum of products, which obtained by multiplying the squared deviations (taken from assumed mean to midvalues) corresponding to its frequencies; $\sum fd$ = Sum of the products obtained by multiplying the deviations (d) with its corresponding frequencies and n = total of the items

Steps to compute SD for continuous series by short-cut method:

1) Find out the mid-values; 2) Select any mid values as an assumed mean and find the deviation by deducting the assumed mean from the values (M-A) these are denoted by d; 3) Square the deviations, denoted by d^2; 4) Multiply the deviations with corresponding frequencies and obtain the total i.e., $\sum fd$; 5) Multiply the squared deviations with its corresponding frequencies and obtain the total i.e., $\sum f\, d^2$; 6) Obtain the total of the variables (n) and apply the formula.

Illustration 13: The profits (in Rs. Lakhs) earned by 100 companies during 1998-99 are shown below. Compute Standard Deviation by using Direct and Short-Cut methods.

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: Direct Method

Calculation of Standard Deviation

Classes (Profit Rs. in lakhs)	Mid Values X	No. of Companies f	fx	d (X - $\overline{X}$)	d^2	fd^2
20-30	25	4	100	-34.1	1162.81	4651.24
30-40	35	8	280	-24.1	580.81	4646.48
40-50	45	18	810	-14.1	198.81	3578.58
50-60	55	30	1650	-4.1	16.81	504.30
60-70	65	15	975	5.9	34.81	522.15
70-80	75	10	750	15.9	252-81	2525.10
80-90	85	8	680	25.9	670.81	5366.48
90-100	95	7	665	35.9	1288.81	9021.67
		n = 100	**∑fx = 5,910**			**$\sum fd^2$ = 30819.00**

$\overline{X} = \frac{\sum fX}{n} = \frac{5910}{100}$ = Rs. 59.10 lakhs

σ = $\sqrt{\frac{\sum fd^2}{n}} = \sqrt{\frac{30819}{100}} = \sqrt{308.19}$ = Rs. 17.56 lakhs

Short-cut Method: Here the assumed mean is 55

Classes (Profit Rs. in lakhs)	Mid Values X	No. of Companies f	$\frac{x-A}{d}$ (A-55)	d^2	fd	fd^2
20-30	25	4	-30	900	-120	3600
30-40	35	8	-20	400	-160	3200
40-50	45	18	-10	100	-180	1800
50-60	55	30	0	0	0	0
60-70	65	15	10	100	150	1500
70-80	75	10	12	400	200	4000
80-90	85	8	30	900	240	7200
90-100	95	7	40	1600	280	11200
		n = 100			**∑fd = 410**	**$\sum fd^2$ = 32,500**

σ $= \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} = \sqrt{\frac{32,500}{100} - \left(\frac{410}{100}\right)^2}$

$= \sqrt{325 - 16.81} = \sqrt{908.19}$ = Rs. 17.56 lakhs

1. Now, we will compare the procedure for calculation of standard deviation in discrete and continuous series. Formulas are same and the steps also same except one step in continuous series, i.e., finding mid values, thus, the only difference in procedure is that in case of continuous series is to find mid values of the various classes.

2. By comparing illustration 13 the computation of standard deviation by Direct and short-cut methods you could may notice the difficult and time consuming calculations in direct method if the arithmetic mean is in fraction. This difficulty can overcome in short-cut method.

The short-cut method is further simplified which is termed as step deviation method. Let us, now, study the importance and procedure of step deviation method to compute standard deviation.

Step Deviation Method: The formulas of direct and short-cut methods could be used conveniently, if the value of X and f are small. If the values of X and f are large, the calculation standard deviation through the above discussed methods are quite tedious and time consuming. In such a case, the calculation can be reduced to a greater extent by step deviation method. This method may be applied for grouped data. It is applicable when there is constant gap in between the values of items. In case of continuous series, if class intervals are equal then only it is applicable. Now, you study the procedure carefully to understand this method.

Formula: $\sigma = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times C$

Here, C is the common factor.

Steps to compute SD by step deviation method:

1) Find mid value of various classes; 2) Select a mid value as the assumed mean and take the deviation of the mid values from the assumed mean (M-A) and denote these deviations by 'd'; 3) Take the common factor of the deviations and divide the deviations by the common factor, denote these deviation by d'; 4) Square the deviations and denote by d'^2; 5) Multiply the respective frequencies with their deviations (d') obtained in step 3 and get the total i.e., $\sum fd'$; 6) Multiply the squared deviation (d'^2) with the respective frequencies and obtain the total i.e., $\sum fd'^2$; 7) Get the sum of the items (n) and apply the formula.

Note: Instead of squaring the deviations (in step 4) you may also multiply the fd'values with its respective deviations (d') to find fd'^2. The clarifications is that:

fd'^2 means f (d'^2); d'^2 = (d') (d'); Therefore fd'^2= f (d') (d') i.e. $fd'(d')$

Illustration 14: Find the standard deviation of the following distribution:

Income per month (Rs.) :	0-500	500-1000	1000-1500	1500-2000	2000-3000
No. of Employees :	90	218	86	41	15

Solutions : Calculation of Standard Deviation

Income per month (Rs.) x	No. of employees (f)	Mid-point (m)	(x-750) (d)	$d' = \frac{m-750}{250}$	fd'	$f(d'^2)$
0-500	90	250	-500	-2	-180	360
500-1000	218	750	0	0	0	0
1000-1500	86	1250	500	2	172	344
1500-2000	41	1750	1000	4	164	656
2000-3000	15	2500	1750	7	105	735
	N = 450	**-**	**-**	**-**	$\sum fd' = 261$	$\sum fd'^2 = 2095$

Here, assumed mean (A) is 750 and common factor (C) is 250.

S.d. $= \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times c$

$= \sqrt{\frac{2095}{450} - \left(\frac{261}{450}\right)^2} \times 250$

$= \sqrt{4.6556 - (0.58)^2} \times 250$

$= \sqrt{4.3192} \times 250 = 519.2$ 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 C

1) Define Standard deviation.

2) Write the formulae used and the procedure for computing standard deviation by direct, short-cut and step deviation methods.

3) Computing standard deviation by using direct method and short-cut method from the following set of observations.

   245, 322, 192. 310, 231

4) Calculate standard deviation by using direct, short-cut and step deviation methods from 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

14.6.4.1 Properties of Standard Deviation

You have learnt the meeting and methods of computing standard deviation. Let us, now, 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 = $\frac{\sum X}{n} = \frac{15}{5} = 3$

$$\sigma \text{ of } X = \sqrt{\frac{\sum(X - \overline{X})^2}{n}} = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.414$$

Arithmetic mean of Y = $\frac{\sum Y}{n} = \frac{15}{5} = 3$

$$\sigma \text{ of } Y = \sqrt{\frac{\sum(Y - \overline{Y})^2}{n}} = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.414$$

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 -$\overline{X}$	$(X - \overline{X})^2$	Let Y = 10X	(Y - $\overline{Y}$)	$(Y - \overline{Y})^2$
1	-2	4	10	-2	4
2	-1	1	20	-1	1
3	0	0	30	0	0
4	1	1	40	1	1
5	2	4	50	2	4
Total 15	0	10	150	0	10

$\overline{X} = \frac{\sum x}{n} = \frac{15}{5} = 3$

$\overline{Y} = \frac{\sum Y}{n} = \frac{150}{5} = 30$

$$\sigma \text{ of } Y = \sqrt{\frac{\sum(Y - \overline{Y})^2}{n}} = \sqrt{\frac{1000}{5}} = \sqrt{200} = 10\sqrt{2} = 14.14$$

σ 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) It two groups contain $n_1$and n_2 observations with means $\overline{X}_1$ and $\overline{X}_2$ and standard deviation $\sigma_1$and σ_1 respectively, then the standard deviation of the combined

$$\sigma_{12} \quad = \sqrt{\frac{(n_1\,\sigma_1^2 + n_2\,\sigma_2^2) + n_1\,d_1^2 + n_2\,d_2^2}{(n_1 + n_2}}$$

Where σ_{12} = combined standard deviation of the two groups

d_1 $= \overline{X}_{12} - \overline{X}_1$; $d_2 = \overline{X}_{12} - \overline{X}_2$

$\overline{X}_{12}$ = combined arithmetic mean of the two groups.

To understand the properties 3 and 4, study Illustrations 20 and 21 given under Section 14.8 (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 ($\overline{\boldsymbol{X}}$)which is 3.

X	:	1	2	3	4	5
X – 4	:	-3	-2	-1	0	1
$(\mathrm{X} - 4)^2$	:	9	4	1	0	1

Now, $\sum(\mathrm{X} - 4)^2 = 15$

Root Mean Square Deviation about 4 $= \sqrt{\frac{\sum(\mathrm{X}-4)^2}{5}}$

$$= \sqrt{\frac{15}{5}} = \sqrt{3} = 1.732$$

But standard deviation of X is$\sqrt{2}$ or 1.414. So, root mean square deviation about a value other than arithmetic mean is than standard deviation.

6) In an ordinary type data or normal type data the number of items between the range A. M. ± σ is about 68%, in the range A.M. ± 2 σ is about 95% and in range A.M. ± 3 σ is almost all the items of the data lie.

To explain it, let us consider the data of **Illustration 12**. For this data A.M. is 16 and σ is 2.97. So the range A.M. ± σ will be 16 ± 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. ± 2 σ will be 16 ± 2 x 2.97 o r 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. 23 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 12 is not perfectly normal but is quite close to it.

14.6.4.2 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 t o other measures of dispersion is that it is amenable to algebraic 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 limitations 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 o r 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 items and it gives less importance to items which are nearer to mean.

iii) The standard deviation cannot be computed for a distribution with open-and classes.

14.7 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:

Coefficient of Variation (c.v) = $\frac{Standard\ Deviation\ (\sigma)}{Arithmetic\ Mean\ (\bar{X})} \times 100$

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 indifferent 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 cms., 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 compared directly with 10 cms., deviation of the second case. As, 1,00,000 cms., is larger than 10 cms., 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% used in the second case the coefficient is $\frac{10}{1000} \times 100$ or 1%. So deviation in second case is relatively larger. **Thus, whenever comparisons of variations is to be done it must be done in terms of coefficient of variation only.**

Variance

In 1913 F.A. Fisher used the measure of variance to describe the square of the Standard deviation. Variance is defined as “the square of standard deviation”. This concept is useful in advanced work where it is possible to split the sum into several pasts each attributable to one of the factors causing variation in the original data set.

Variance = σ^2 or σ = $\sqrt{Variance}$

Thus, the formula can be present as follows:

In ungrouped data:

Variance (direct method) = $\sum x^2 / n$

Variance (sort-cut method) = $\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$

In group data: discrete series

Variance (direct method) = $\sum fx^2 / n$

Variance (sort-cut method) = $\sum fd^2 / n \; (\sum fd^2 / n\,)^2$

Continuous Series

The formulas presented in discrete series are same in continuous series also.

In step deviation method:

Variance: $\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2 x\ C^2$

Illustration 15: 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		**$\sum fd = 0$**	**$\sum fd^2 = 30$**

Arithmetic Mean of Team A: $= A + \frac{\sum fd}{n} \quad = 2 + \frac{0}{25} = 2$ goals

Standard Deviation of Team A: $= \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$

$= \sqrt{\frac{30}{25} - \left(\frac{0}{25}\right)^2}$

$= \sqrt{1.2 \; - 0} = \sqrt{1.2}$

$= 1.1$

Coefficient of Variation of Team A : $= \frac{S.D.}{\bar{x}} \times 100 \; = \frac{1.1}{2} \times 100 = 55\%$

Coefficient of Variation of Team B: $= \frac{S.D.}{\bar{x}} \times 100 \; = \frac{1.25}{2.5} \times 100 = 50\%$

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 16: 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) (x)	Mid-Value (m)	Frequency (f)	Step Deviation (d)	fd	Fd^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	83
60-70	65	18	3	54	162
		N = 140		**$\sum$fd= 0**	**$\sum$fd^2 = 30**

Here, assumed mean (A) is 35 and C is 10.

$\overline{x_A} = \text{A} + \frac{\sum \text{fx}}{n} + \text{C}$

$= 35 + \frac{100}{140} + 10 = 35 + 7.143 = 42.1$ approximately.

$\sigma_A = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times c$

$= \sqrt{\frac{348}{140} - \left(\frac{100}{140}\right)^2} \times 10$

$= \sqrt{2.486 - 0.510} \times 10$

$= 1.4057 \times 10 = 14.057$

C.V. (Series A) $= \frac{\sigma}{X} \times 100$

$= \frac{14.06}{42.1} \times 100 = 33.3\%$

Variance (Series A): $= \frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2 \times C^2$

$= \frac{348}{140} - \left(\frac{100}{140}\right)^2 \times 10^2$

$= 2.486 - (0.510)^2 \times 100$

$= 1.976 \times 10$

$= 197.6$

We can also compute the variance as follows:

Variance $= \sigma^2$

σ of A Series $= 14.057$

Variance (A) $= 14.057^2 = 197.6$

Computation of Arithmatic Mean and Standard Deviation of Series-B

Class-Interval (Variable) (x)	Mid-Value (m)	Frequency (f)	Step Deviation (d)	fd	Fd^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
		N = 140		**$\sum$fd= 40**	**$\sum$fd^2 = 288**

Here, assumed mean (A) is 35 and C is 10.

$$\overline{X}_B = \text{A} + \frac{\sum \text{fx}}{n} + \text{C}$$

$$= 35 + \frac{40}{140} + 10 \quad = 35 + 2.85 = 37.85 \text{ approximately.}$$

$$\sigma_B = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times c \quad = \sqrt{\frac{288}{140} - \left(\frac{40}{140}\right)^2} \times 10$$

$$= \sqrt{2.057 - 0.0784} \times 10 \quad = \sqrt{1.9786} \times 10$$

$$= 1.4057 \times 10 = 14.057$$

C.V. (Series B) $= \frac{\sigma}{X} \times 100 = \frac{14.06}{37.85} \times 100$ = 37.1%

Variance (Series B) $= \sigma^2$

Standard deviation of B = 14.057

Variance $= 14.057^2$ = 197.6

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 and variance of both the series is the same i.e., 14.057 and 197.6 respectively. From this fact, we should not conclude that two series have same variation. The difference in arithmetic mean has to be taken into account for correct interpretation.

14.8 SOME ILLUSTRATIONS

Illustration 17: 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 : **Computation**

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
		N = 100	**$\sum$fX= 5,910**	**$\sum fX^2$ = 3,80,100**

a) $\overline{X} = \frac{\sum \text{fX}}{n} = \frac{5,910}{100} = \text{Rs. } 59.10 \text{ Lakhs}$

b) Variance $= \frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2$

$= \frac{3,80,100}{100} - \left(\frac{5910}{100}\right)^2$

= 3801.00 – 3492.81

= Rs. 308.19 Lakhs

c) Standard Deviation $= \sqrt{Variance} = \sqrt{308.19}$

= 17.56 Lakhs

In the above illustration you may notice that by using sums of items and their squares to calculations involved are large. This method is a direct method in the sense that we have used the items directly and not calculated their deviation from any value. This method may be used only when size of items are small and their total number is also small.

Illustration 18: 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.

Calculation of Mean and Standard Deviation

Class Interval	f	Mid-point (X)	D = X-A (X-45)	$d'=\frac{d}{c}$ c = 10	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
	n = 58	**-**	**-**	**-**	$\sum fd' = 0$	$\sum fd'^2 = 148$

Mean $\overline{X} = A + \frac{\sum fx}{n} \times C$

$= 45 + \frac{0}{58} \times 10 = 45$

Standard Deviation $= C \times \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$

$= 10 \times \sqrt{\frac{148}{58} - \left(\frac{0}{50}\right)^2}$

$= 10 \times \sqrt{2.552} = 1.597 \times 10 = 15.97$

Illustration 19 : A state government decided to give old age pension to people over sixty years of age. The scale of pension were fixed as follows:

Age Group	Us. per month
60-65	250
65-70	300
70-75	350
75-80	400
80-85	450

The age 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
6	67	68								

Calculate the monthly average pension payable and standard deviation, variance and co-efficient of standard deviation.

Solution: Classification of Data

Age Group	Talley	Frequency
60-65	卌 II	7
65-70	卌	5
70-75	卌 I	6
75-80	IIII	4
80-85	III	3
		25

Calculation of Monthly Average Pension Payable and the Standard Deviation

Scale of Pension (Rs.)	f	$d'=\left(\frac{X-350}{50}\right)$	fd'	fd'^2
250	7	-2	-14	28
300	5	-1	-5	5
350	6	0	0	0
400	4	1	4	4
450	3	2	6	12
	25	-	-9	49

Here, A = 350, C = 50; $\sum f$ or n = 25; $\sum fd'$ = -9; and $\sum fd'^2$ = 49

$$\overline{X} = A + \frac{\sum fd'}{n} \times C$$

$$= 350 - \frac{9}{25} \times 50 = 332$$

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times c$$

$$= \sqrt{\frac{49}{25} - \left(\frac{-9}{25}\right)^2} \times 50 = 1.353 \times 50 = 67.65$$

$= \sigma^2 = 67.65^2 = 4576.52$

Coefficient of σ $= \frac{\sigma}{\overline{X}} \times 100$

$$= \frac{67.25}{332} \times 100 = 20.04\%$$

Thus, the monthly average pension is Rs. 332; standard deviation is Rs. 67.65 variance is 4776.52 and C.V. is 20.04%.

Illustration 20: 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 this data n_1= 50 and n_2= 40

$\overline{X}_1$= 72 and $\overline{X}_2$= 54

σ_1= 9 and σ_2= 6

Combined mean for group of 90 ($\overline{X}_{12}$) = $\frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1+n_2} = \frac{50\times72 + 40\times54}{90}$

$= \frac{3{,}600 + 2{,}160}{90} = 64$

Combined Standard Deviation for the group of 90

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2+d_1^2)+ n_2(\sigma_2^2+d_2^2)}{n_1+ n_2}}$$

Now, d_1= 64 – 72 = – 8 and d_2= 54 – 72 = –18

$$\sigma_{12} = \sqrt{\frac{50(80+64)+ 40(36+324)}{90}} = \sqrt{\frac{7{,}250+ 14{,}400}{90}}$$

$$= \sqrt{\frac{21{,}650}{90}} = \sqrt{240.54} = 15.51$$

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 greater of the given standard deviations. Combined mean will always be in between the range of the given mean, 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 mean, the combined standard deviation will be more away from the largest given standard deviation. 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 21: Calculate mean deviation about mean for data given previously in Illustration 18 and show that mean deviation is less than standard deviation.

Solution: **Calculation of Mean Deviation**

Class Interval	Frequency (f)	Mid-point (X)	$\lvert m - \overline{X}\rvert$ $\lvert d\rvert$	$f\lvert d\rvert$
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
	n = 58			$\sum f\|d\|$ **= 720**

From Illustration 18, we have $\overline{X}$= 45 and σ = 15.97

Mean Deviation about $\overline{X} = \frac{\Sigma f|d|}{n} = \frac{720}{58} = 12.41$

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 data.

Check Your Progress D

1) The following table gives weight in pounds of fat bullocks and fat sheep.

Fat Bullocks (Weight in lbs.)	Number	Fat Sheep (Weight in lbs.)	Number
850-900	2	150-175	8
900-950	24	175-200	30
950-1000	45	200-225	59
1000-1050	120	225-250	70
1050-1100	110	250-275	98
1100-1150	140	275-300	60
1150-1200	66	300-325	37
1200-1250	42	325-350	23
1250-1300	20	350-375	15
1300-1350	15	375-400	5

Determine if the Bullocks or the sheep are more variable in weight.

2) In a co-educational college boys and girls formed separate groups on the foundation day when every one had to put in physical labour. Compute standard deviation for boys and girls separately and for the combined group. Did the separation by sex make each work group more homogeneous.

Minutes of labour given by each individual	No. of Girls	No. of boys
60	20	120
55	60	100
50	100	200
45	450	355
40	450	350
35	300	500
30	250	350
25	100	20

14.9 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 fr.om 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 algebraic 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 some times 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.

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 Meant 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

(σ)= $\sqrt{\frac{\Sigma f d'^2}{n}-\left(\frac{fd'}{n}\right)^2} \times C$ Standard deviation is rigidly defined and based on all items.

14.10 KEY WORDS

Coefficient of Variation: Standard deviation divided by arithmetic mean expressed as a percentage.

Inter Quartile Range: A measure of dispersion which considers the spread in the middle 50%. It is $(Q_3 - Q_1)$ of the data.

Lorenz Curve: A double cumulative percentage graph used in determining the extent of inequalities of items.

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.

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.

14.11 ANSWERS TO CHECK YOUR PROGRESS

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

C) 3) 49.1; 4) 156.37;

D) 1) Bullocks: $\overline{X}$ = 1097.52; σ = 90.34; C.V. = 8.23%

Sheeps: $\overline{X}$ = 261.15; σ = 47.75; C.V. = 18.25%

2) Girls: $\overline{X}$ = 39.45; σ = 7.5; C.V. = 19.00%

Boys: $\overline{X}$ = 40.69; σ = 8.68; C.V. = 21.34%

$\overline{X}_{12}$= 40.11; σ_{12}= 8.18

14.12 TERMINAL QUESTIONS/EXERCISES

Questions:

1) What do you understand by dispersion? What purpose does it serve?.
2) What is the mean deviation? Review its advantages and disadvantages.
3) What is standard deviation? Explain its superiority over other measures of dispersion.
4) What is coefficient of variation? What is its role as a measure of variation? How does it differ from variance.
5) Define various measures of dispersion and explain their relative merits and limitations.

Exercises:

1). Calculate quartile deviation and mean deviation about 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 : R= 60 years, Q.D. =.10, M.D($\overline{X}$) = 9.52)

2). A frequency distribution for the duration of 20 long distance telephqne 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. ' 00)	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 = 8.9, Coefficient of M.D. about M = 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_d) = 9.08)

6). Calculate the mean deviation about Mode and its coefficient for the following data:

No. or Detects 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_d) = 102.13, Coefficient of M.D. (M_d) = 0.011)

8). Compute the mean deviation about mean and its co-efficient from the following data:

Marks	No. of Students	Marks	No. of Students
0-10	4	30-40	10
10-20	6	40-50	6
20-30	10	50-60	4

(Answer: M.D$(\overline{X})$ = 11.33 , Co-efficient of M.D. = 0.32)

9). 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 : X	B	C	D	E	F	G	H	I	J
Marks : 5	10	20	25	40	42	45	48	70	80

(Answer : 23.06)

10). Calculate standard deviation from the following data :

Mid-points	1	2	3	4	5	6	7	8	9
Frequency	2	60	101	152	205	155	79	40	1

(Answer : = 1.57)

11). Compute standard deviation for the following data which relate to the profits of 100 companies:

Profit (Rs. in lakhs)	8-10	10-12	12-14	14-16	16-18	18-20
No. of Companies	8	12	20	30	20	10

(Answer : σ = 2.77)

12). An analysis of production rejects resulted in the following figures. Calculate mean and standard deviation.

No. of Rejects per Operator	21-25	26-30	31-35	36-40	41-45	46-50	51-55
No. of Operators	8	15	28	42	15	12	3

(Answer : $\bar{X}$= 36.96; σ = 6.735)

13). 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 : σ_{12} = 14)

14). 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 66)	222
Less than 70	230

(Answer : σ = 13.9, C.V. = 37.3%)

15). You are given the data pertaining to kilowatt hours of electricity consumed by100 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) .

16). 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%)

17). 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	8	4
700-900	11	30
900-1100	26	12
1100-1300	10	9
1300-1500	8	16

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%)

18) In two factories A andB, 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) $\bar{X}_{12}$= Rs. 473.33, σ_{12} = 49.19

19) 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)

20) The mean of two samples of size 50 and 100 are 54.1 and 50.3 and the standard deviations are 8 and 7 respectively. Find the mean and standard deviation of the sample of size 150 obtained b combining the two samples.
(Answer : $\bar{X}_{12}$= 51.57, σ_{12} = 7.56)

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.

FURTHER READINGS

Arora, P.N. Sumeet Arora and Arora. A., 2007, *Comprehensive Statistical Methods.* S. Chand and Company Ltd., New Delhi.

Beri, G.C., 2005, *Business Statistics,* Tata Mc Graw-Hill Publishing Company, Ltd., New Delhi.

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.

Simpson, G, and.Kafka, F. Basic Statistics, Oxford & IBH Publishing 1 New Delhi.

UNIT 15 SIMPLE LINEAR CORRELATION

Structure

15.0 OBJECTIVES

After studying this unit, you should be able to:

- explain the concept of correlation,
- use scatter diagrams to visualize the relationship between two variables,
- compute the simple and rank correlation coefficients between two variables,

15.1 INTRODUCTION

In previous units, so far, we have discussed the statistical treatment of data relating to one variable only. In many other situations decision-makers need to consider the relationship between two or more variables. For example, the sales manager of a company may observe that the sales are not the same for each month. He/she also knows that the company's advertising expenditure varies from year to year. This manager would be interested in knowing whether a relationship exists between sales and advertising expenditure. If the manager could successfully define the relationship, he/she might use this result to do a better job of planning and to improve predictions of yearly sales with the help of the regression technique for his/her company. Similarly, a researcher may be interested in studying the effect of research and development expenditure on annual profits of a firm, the relationship that exists between price index and purchasing power etc. The variables are said to be closely related if a relationship exists between them. In this unit we discuss bi-variant analysis of Simple Linear Correlation and Simple Linear Regression will be covered in the next unit i.e. Unit-16.

The word 'bi-variate' is used to describe the situation in which two characteristics are measured on each variable or item, the characteristics being represented by the variables or item.

This unit, therefore, introduces the concept of correlation and statistical techniques of simple correlation.

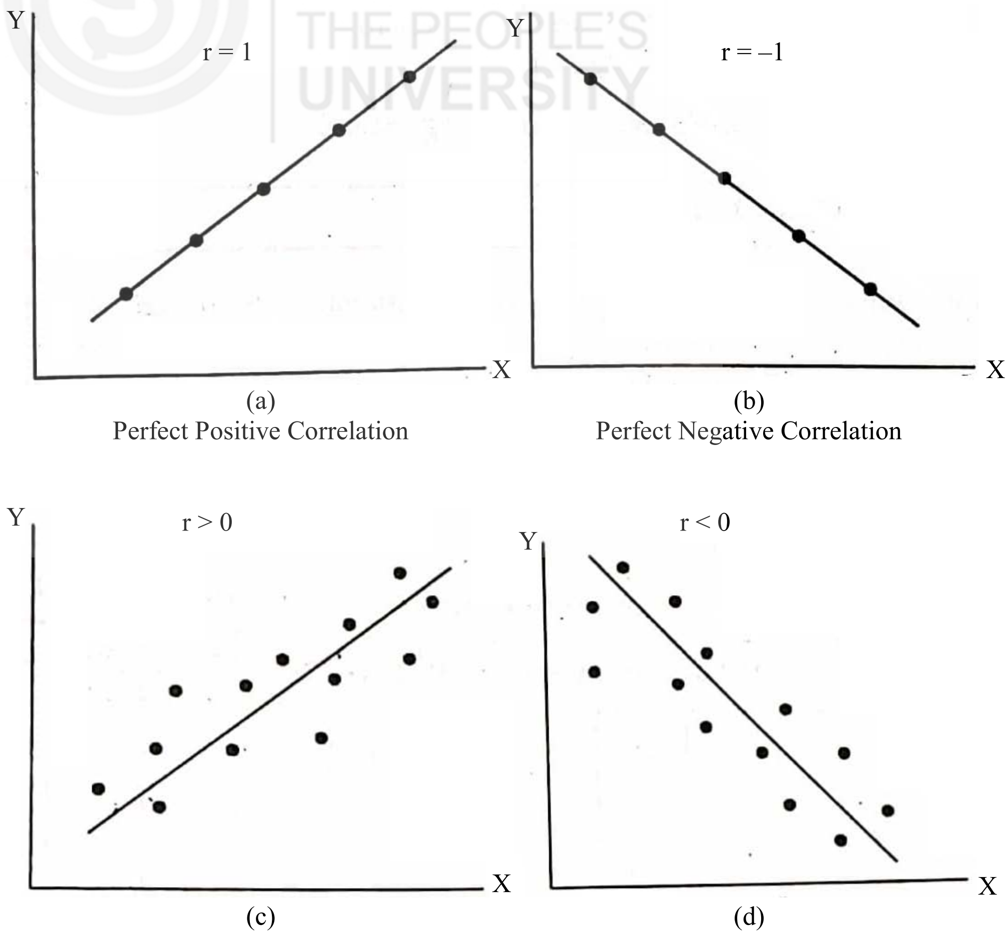
15.2 SIMPLE CORRELATION

15.2.1 Meaning

If two variables, say x and y vary or move together in the same or in the opposite directions they are said to be correlated or associated. Thus, correlation refers to the relationship between the variables. Generally, we find the relationship in certain types of variables. For example, a relationship exists between income and expenditure, absenteesim and production, advertisement expenses and sales etc. Existence of the type of relationship may be different from one set of variables to another set of variables. Let us discuss some of the relationships with the help of Scatter Diagrams.

15.2.2 Scatter Diagram

When different sets of data are plotted on a graph, we obtain **scatter diagrams**. A scatter diagram gives two very useful types of information. Firstly, we can observe patterns between variables that indicate whether the variables are related. Secondly, if the variables are related we can get an idea of the type of relationship that exists. The scatter diagram may exhibit different types of relationships. Some typical patterns indicating different correlations between two variables are shown in Figure 15.1.

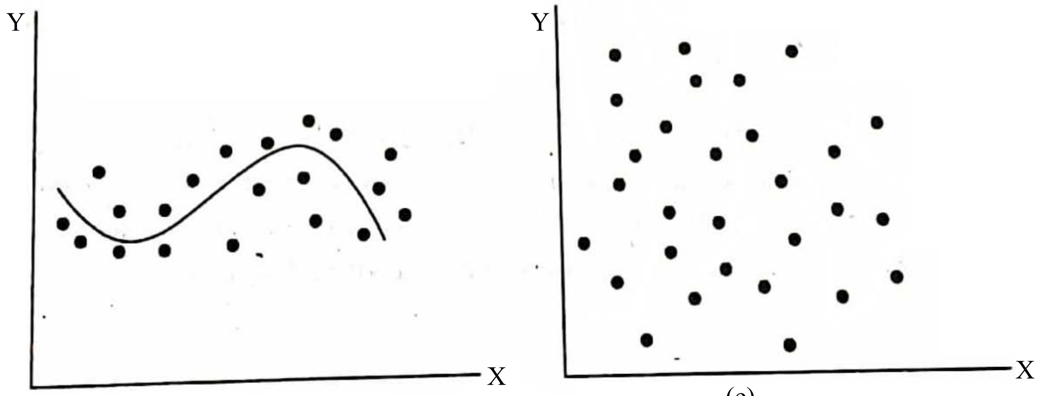


(a) Perfect Positive Correlation

(b) Perfect Negative Correlation

(c) Positive Correlation

(d) Negative Correlation



(e)
Non-linear Correlation

(e)
No Correlation

Figure 15.1 : Possible Relationships Between Two Variables, X and Y

If X and Y variables move in the same direction (i.e., either both of them increase or both decrease) the relationship between them is said to be **positive correlation** [Fig. 15.1 (a) and (c)]. On the other hand, if X and Y variables move in the opposite directions (i.e., if variable X increases and variable Y decreases or vice-versa) the relationship between them is said to be **negative correlation** [Fig. 15.1 (b) and (d)]. If Y is unaffected by any change in X variable, then the relationship between them is said to be **un-correlated** [Fig. 15.1 (f)]. If the amount of variations in variable X bears a constant ratio to the corresponding amount of variations in Y, then the relationship between them is said to be **linear-correlation** [Fig. 15.1 (a) to (d)], otherwise it is **non-linear or curvilinear correlation** [Fig. 15.1 (e)]. Since measuring non-linear correlation for data analysis is far more complicated, we therefore, generally make an assumption that the association between two variables is of the linear type.

If the relationship is confined to two variables only, it is called **simple correlation**. The concept of simple correlation can be best understood with the help of the following illustration which relates advertisement expenditure to sales of a company.

Illustration 1

Table 15.1 : A Company's Advertising Expenses and Sales Data (Rs. in crore)

Years:	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
Advertisement expenses (X)	6	5	5	4	3	2	2	1.5	1.0	0.5
Sales (Y)	60	55	50	40	35	30	20	15	11	10

The company's sales manager claims the sales variability occurs because the marketing department constantly changes its advertisement expenditure. He/she is quite certain that there is a relationship between sales and advertising, but does not know what the relationship is.

The different situations shown in Figure 15.1 are all possibilities for describing the relationships between sales and advertising expenditure for the company. To determine the appropriate relationship, we have to construct a scatter diagram shown in Figure 15.2, considering the values shown in Table 15.1.

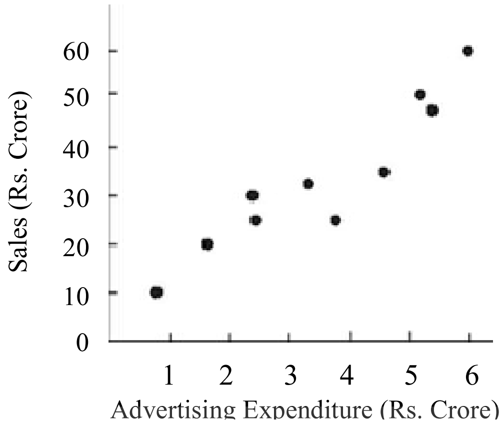


Figure 15.2 : Scatter Diagram of Sales and Advertising Expenditure for a Company.

Figure 15.2 indicates that advertising expenditure and sales seem to be linearly (positively) related. However, the strength of this relationship is not known, that is, how close do the points come to fall on a straight line is yet to be determined. The quantitative measure of strength of the linear relationship between two variables (here sales and advertising expenditure) is called the correlation coefficient. In the next section, therefore, we shall study the methods for determining the coefficient of correlation.

Let us understand through another example.

Illustration 2:

A teacher is interested in studying the relationship between the performance in Statistics and Economics of a class of 20 students. For this he compiles the scores on these subjects of the students in the last semester examination. Some data of this type are presented in Table 15.2.

Table 15.2: Scores of 20 Students in Statistics and Economics

Serial	Score in		Serial	Score in	
Number	Statistics	Economics	Number	Statistics	Economics
1	82	64	11	76	58
2	70	40	12	76	66
3	34	35	13	92	72
4	80	48	14	72	46
5	66	54	15	64	44
6	84	56	16	86	76
7	74	62	17	&I	52
8	84	66	18	60	40
9	60	52	19	82	60
10	86	82	20	90	60

A representation of data of this type on a graph is a useful device which will help us to understand the nature and form of the relationship between the two variables, whether there is a discernible relationship or not and if so whether it is linear or not. For this let us denote score in Economics by X and the score in Statistics by Y and plot the data of Table 15.1 on the x-y plane. It does not matter which is called X and which Y for this purpose. Such a plot is called Scatter Plot or Scatter Diagram. For data of Table 15.2 the scatter diagram is given in Fig. 15.3.

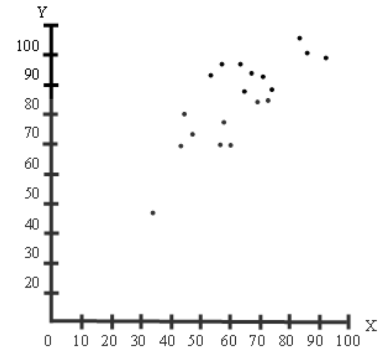


Fig. 15.3 Scatter Diagram of scores in Statistics and Economics

An inspection of Table 15.2 and Fig. 15.3 shows that there is a positive relationship between x and y. This means that larger values of x are associated with larger values of y and smaller values of x with smaller values of y. Further, the points seem to lie scattered around both sides of a straight line. Thus it appears that a linear relationship exists between x and y. However, this relationship is not perfect in the sense that there are deviations from such a relationship. It would indeed be useful to get a measure of the strength of this linear relationship.

Check Your Progress A

1) Suggest eight pairs of variables, four in each, which you expect to be positively correlated and negatively correlated

2) How does a scatter diagram approach help in studying the correlation between two variables?

15.3 CORRELATION COEFFICIENT

The coefficient of correlation helps in measuring the degree of relationship between two variables, X and Y. The methods which are used to measure the degree of relationship will be discussed below.

15.3.1 Karl Pearson's Correlation Coefficient

Karl Pearson's coefficient of correlation (r) is one of the mathematical methods of measuring the degree of correlation between any two variables X and Y is given as:

$$r = \frac{\sum dxdy}{\sqrt{\sum dx^2}\sqrt{\sum dy^2}} \quad (1)$$

Where $dx = X - \bar{X}$; $dy = Y - \bar{Y}$, $dx^2 = (X - \bar{X})^2$ and, $dy^2 = (Y - \bar{Y})^2$

This can also be written as:

$$r = \frac{\sum dxdy}{N\sigma x \ \times \sigma y}$$

Note: The above formula is used when $\bar{X}$ and $\bar{Y}$ are integers.

The following is the alternative formula, when $\bar{x}$ and $\bar{y}$ are not integers.

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}} \qquad (2)$$

Before we proceed to take up an illustration for measuring the degree of correlation, it is worthwhile to note some of the following important points.

i) 'r' is a dimensionless number whose numerical value lies between +1 to –1. The value +1 represents a perfect positive correlation, while the value –1 represents a perfect negative correlation. The value 0 (zero) represents lack of correlation. Figure 15.1 shows a number of scatter plots with corresponding values for correlation coefficient.

ii) The coefficient of correlation is a pure number and is independent of the units of measurement of the variables.

iii) The correlation coefficient is independent of any change in the origin and scale of X and Y values.

Remark: Care should be taken when interpreting the correlation results. Although a change in advertising may, in fact, cause sales to change, the fact that the two variables are correlated does not guarantee a cause and effect relationship. Two seemingly unconnected variables may often be highly correlated. For example, we may observe a high degree of correlation: (i) between the height and the income of individuals or (ii) between the size of the shoes and the marks secured by a group of persons, even though it is not possible to conceive them to be casually related. When correlation exists between such two seemingly unrelated variables, it is called spurious or non-sense correlation. Therefore we must avoid basing conclusions on spurious correlation.

Illustration 3

Taking as an illustration, the data of advertisement expenditure (X) and sales (Y) of a company for 10 years shown in Table 15.1, we proceed to determine the correlation coefficient between these variables.

Solution: Table 15.3: Calculation of Correlation Coefficient

Advertisement Expenditure Rs. (X)	Sales Rs. (Y)	XY	X^2	Y^2
6	60	360.0	35	3600
5	55	275.0	25	3025
5	50	250.0	25	2500
4	40	160.0	16	1600
3	35	105.0	9	1225
2	30	60.0	4	900
2	20	40.0	4	400
1.5	15	22.5	2.25	225
1.0	11	11.0	1	121
0.5	10	5.0	0.25	100
$\sum X = 30$	**$\sum Y = 326$**	**$\sum XY = 1288.5$**	**$\sum X^2 = 122.50$**	**$\sum Y^2 = 13696$**

We know that

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

$$= \frac{\frac{1288.5-(30)(326)}{10}}{\sqrt{122.5 - \frac{(30)^2}{10}}\sqrt{13696 - \frac{(326)^2}{10}}} = \frac{310.5}{315.7}$$

=0.9835

The calculated coefficient of correlation r = 0.9835 shows that there is a high degree of association between the sales and advertisement expenditure. For this particular problem, it indicates that an increase in advertisement expenditure is likely to yield higher sales. If the results of the calculation show a strong correlation for the data, either negative or positive, then the line of best fit to that data will be useful for forecasting (it is discussed in Unit-16 on 'Simple Linear Regression').

Illustration-4

Calculate correlation coefficient for the data given in illustration 2.

Solution:

Table 15.4: Calculation of Correlation Coefficient

Observation No.	X	Y	X^2	Y^2	$\overline{XY}$
1	82	64	6724	4096	5248
2	70	40	4900	1600	2800
3	34	35	1156	1225	1190
4	80	48	6400	2304	3840
5	66	54	4356	2916	3564
6	84	56	7056	3136	4704
7	74	62	5476	3844	4588
8	84	66	7056	4356	5544
9	60	52	3600	2704	3120
10	86	82	7396	6724	7052
11	76	58	5776	3364	4408
12	76	66	5776	4356	5016
13	92	72	8464	5184	6624
14	72	46	5184	2116	3312
15	64	44	4096	1936	2816
16	86	76	7396	5776	6536
17	84	52	7056	2704	4386
18	60	40	3600	1600	2400
19	82	60	6724	3600	4920
20	90	60	8100	3600	5400
Total	**1502**	**1133**	**116292**	**67141**	**87450**

From Table 15.4 we note that:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{1502}{20} = 75.1;$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{1133}{20} = 56.65;$$

$$\sigma_{\overline{X}} = \frac{1}{N}\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} = \frac{1}{20}\sqrt{116292 - \frac{(1502)^2}{20}} = \sqrt{174.59}; = 13.21;$$

$$\sigma_{\overline{y}} = \frac{1}{N}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}} = \frac{1}{20}\sqrt{67141 - \frac{1133^2}{20}} = \sqrt{147.83}; = 12.16;$$

$$\sigma xy = \frac{1}{\mathrm{N}}\left[\Sigma XY - \frac{(\Sigma \mathrm{X})(\Sigma \mathrm{Y})}{\mathrm{N}}\right] = \frac{1}{20}\left[87450 - \frac{1502 \times 1133}{20}\right] = 118.09$$

Thus using formula i.e.

$$r = \frac{\sigma xy}{\sigma x\ \sigma \mathrm{y}}$$

$$\mathrm{r} = \frac{118.09}{13.21\ \times 12.16} = 0.735$$

Now, let us use the another formula i.e.

$$r = \frac{\sum \mathrm{XY} - \frac{(\Sigma \mathrm{X})(\Sigma \mathrm{Y})}{\mathrm{N}}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

$$\mathrm{r} = \frac{87450 - 1502\ \times 1133}{\sqrt{\left(11292 - \frac{(1502)^2}{20}\right)}\sqrt{\left(67141 - \frac{(1133)^2}{20}\right)}} = 0.735$$

Thus, we see that both the formulae provide the same value of correlation coefficient r.

Now, you can check yourself that the same value of coefficient of correlation (r) is obtained by using the formula (1) as stated earlier. For this purpose you need the values to be computed in the table 15.4 as follow with five columns. (i) $(X - \bar{X}) = dx$; (ii) $Y - \bar{Y} = dy$; (iii) dx^2 ; (iv) dy^2 and, $dxdy$.

15.3.3 Spearman's Rank Correlation

The Karl Pearson's correlation coefficient, discussed above, is not applicable in cases where the direct quantitative measurement of a phenomenon under study is not possible. Sometimes we are required to examine the extent of association between two ordinally scaled variables such as two rank orderings. For example, we can study intelligence, efficiency, performance, competitive events, attitudinal surveys etc. In such cases, a measure to ascertain the degree of association between the ranks of two variables, X and Y, is called Rank Correlation. It was developed by **Edward Spearman**, its coefficient (R) is expressed by the following formula:

$R = 1 - \frac{6\sum D^2}{N^3 - N}$ where, N = Number of pairs of ranks, and $\sum D^2$= squares of difference between the ranks of two variables.

The following example illustrates the computation of rank correlation coefficient.

Illustration 5

Salesmen employed by a company were given one month training. At the end of the training, they conducted a test on 10 salesmen on a sample basis who were ranked on the basis of their performance in the test. They were then posted to their respective areas. After six months, they were rated in terms of their sales performance. Find the degree of association between them.

Salesmen:	1	2	3	4	5	6	7	8	9	10
Ranks in training (X):	7	1	10	5	6	8	9	2	3	4
Ranks on sales Peformance (Y):	6	3	9	4	8	10	7	2	1	5

Solution: Table 15.5: Calculation of Coefficient of Rank Correlation.

Sales men	Ranks Secured in Training X	Ranks Secured on Sales Y	Difference in Ranks D = (X–Y)	D2
1	7	6	1	1
2	1	3	–2	4
3	10	9	1	1
4	5	4	1	1
5	6	8	–2	4
6	8	10	–2	4
7	9	7	2	4
8	2	2	0	0
9	3	1	4	4
10	4	5	–1	1
				$\sum D^2 = 24$

Using the spearman's formula, we obtain

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6\sum 24}{10^3 - 10}$$

$$= 1 - \frac{144}{990} = 0.855$$

we can say that there is a high degree of positive correlation between the training and sales performance of the salesmen.

Illustration 6

Table 15.6: Rank of 10 candidates by two Examiners.

S.No.	Rank Given by		Difference	
	Examiner 1	**Examiner 2**	***D***	***D^2***
1	6.0	6.5	–0.5	0.25
2	2.0	3.0	–1.0	1.00
3	8.5	6.5	2.0	4.00

4	1.0	1.0	0.0	0.00
5	10.0	2.0	8.0	64.00
6	3.0	4.0	–1.0	1.00
7	8.5	9.5	–1.0	1.00
8	4.0	5.0	–1.0	1.00
9	5.0	8.0	–3.0	9.00
10	7.0	9.5	–2.5	6.25
			$\sum D = 0$	$\sum D^2 = 87.50$

Like Karl Pearson's coefficient of correlation the Spearman's rank correlation has a value + 1 for perfect matching of ranks, -1 for perfect mismatching of ranks and 0 for the lack of relation between the ranks.

Sometimes the data, relating to qualitative phenomenon, may not be available in ranks, but only in values. In such a situation it is necessary to assign the ranks to the values. Ranks may be assigned by taking either from largest to the smallest or vice versa. But the same method must be followed in case of both the variables.

Sometimes there is a tie between two or more ranks in the first and/or second series. **For example**, if the values of two items are same and presume that the rank of one item may be 4th rank, then instead of awarding 4th rank to the respective two observations, we award 4.5 [(4+5)/2] for each of the two observations. Now we will take up an illustration to understand how to award the ranks when the data is given in values and to calculate the rank, correlation. The illustration will also give clarity how to award the ranks when values of items in series are same.

Illustration 7

Calculate rank correlation from the following data related to a group of 10 students and percentage of marks secured.

Roll Nos. of the students	21	22	23	24	25	26	27	28	29	30
% of marks in statistics	45	66	55	45	80	75	50	55	60	45
% of marks in Accountancy	70	81	75	75	70	85	65	80	45	60

Solution:

The above data was given in percentage of marks not in the ranks. Therefore, for calculation of rank correlation, first, we have to assign the ranks to the given values. As we discussed earlier the ranks may be assigned either from the largest value to smallest value or visa-versa. Here, we assign the ranks from largest to smallest value which is normally in practice.

Calculation of rank correlation:

Roll Nos.	% of marks in statistics	% of marks in Accountancy	Ranks of % of marks in Statistics	Ranks of Marks in Accountancy	Difference in Ranks D	D^2
21	45	70	9	6.5	2.5	6.25
22	66	81	3	2	1	1.00
23	55	75	5.5	4.5	1	1.00
24	45	75	9	4.5	4.5	20.25
25	80	70	1	6.5	–5.5	30.25
26	75	85	2	1	1	1.00
27	50	65	7	8	–1	1.00
28	55	80	5.5	3	2.5	6.25
29	60	45	4	10	–6	36.00
30	45	60	9	9	0	0
						ΣD^2= 103.00

$$r = 1 - \frac{6(\Sigma D^2)}{N^3 - N} = 1 - \frac{103}{10^3 - 10} = 1 - \frac{103}{990} = 1 - 0.10 = 0.90$$

Explanation of assigning ranks:

For the values of percentage of marks in statistics for 80, 75, 66, 60 there are only single values. Therefore, ranks have been assigned 1,2,3,4. Whereas the next value 55 repeated two times in the data, therefore, 5 + 6 ranks divided by 2 = 5.5 rank has been allotted to the value of 55 two times.

Similarly, the value of 45 repeated three times in the data, therefore the ranks 8+9+10 divided by 3 equal to 9. Accordingly, the rank 9 has been allotted to value of 45 (in between) value of 55, 45. There is a value of 50, hence rank seven has been allotted to 50. In the same manner, you may try to observe the assigning of ranks to the values of percentage of marks in accountancy.

Check Your Progress B

1) Compute the degree of relationship between price of share (X) and price of debentures over a period of 8 years by using Karl Pearson's formula.

Years:	1996	1997	1998	1999	2000	2001	2002	2003
Price of Shares:	42	43	41	53	54	49	41	55
Price of debentures:	98	99	98	102	97	93	95	94

2) Consider the above exercise and assign the ranks to price of shares and price of debentures. Find the degree of association by applying Spearman's formula.

15.4 LET US SUM UP

In this unit, fundamental concepts, meaning and techniques of correlation (or association) have been discussed. Scatter diagrams, which exhibit some typical pattern indicating different kinds of relationships have been illustrated. A scatter plot of the variables may suggest that the two variables are related but the value of the Karl Pearson's correlation coefficient (r) quantifies the degree of this association. The closer the relation coefficient is to ± 1.0, the stronger the linear relationship between the two variables. Spearman's rank correlation for data with rank is outlined. Finally, we discussed the procedure of assigning the ranks to the variables, if the data is in the values for computation of Rank correlation.

15.5 KEY WORDS

Correlation Analysis: Refer to a measure of association between two random variables. If two random variables have been such that when one gets changed the other will do so in a related manner, they are regarded to be correlated. Variables which are independent are not correlated. The correlation coefficient is a number between –1 and +1. It could be calculated from a number of pairs of observations which are normally referred to a points (x, y) a coefficient of 1 implies perfect positive correlation, -1 perfect negative correlation and 0 no correlation.

Rank Correlation Coefficient: There happen to be many occasions when it may not be convenient, economic or even possible to give value to variables. However, various items can be ranked. In such cases, a rank correlation coefficient may be used.

Scatter Diagram: A diagram showing the joint variation of two variables X and Y. Each member is represented by a point whose coordinates, on ordinary rectangular axes, are the values of the variables. A set of n observations thus provides n points on the diagram and the scatter or clustering of the points exhibits the relationship between X and Y.

15.6 ANSWERS TO SELF ASSESSMENT EXERCISES

B) 1. $r_k = -0.071$

2. R = –0.185

15.7 TERMINAL QUESTIONS/EXERCISES

1) What do you understand by the term correlation? Distinguish between different types of correlation with the help of scatter diagrams?

2) Explain the difference between Karl Pearson's correlation co-effiecient and spearsman's rank correlations co-efficient. Under what situations, in the latter preferred to the former?

3) With the help of an example, explain the procedure you would follow in assigning the ranks when the data as given in values and same values of the observations are common.

4) Calculate the co-efficient of correlation for the ages of husband and wife:

Age of husband	23	27	28	29	30	31	33	35	36	39
Age of wife	18	22	23	24	25	26	28	29	30	32

5) Determine the correlation coefficient between x and y

x	5	7	9	11	13	15
y	1.7	2.4	2.8	3.4	3.7	4.4

6) Ten students obtained the following marks in the mathematics and statistics. Calculate the rank correlation coefficient:

Student	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in statistics	84	51	91	60	68	62	86	58	53	47

7) Ten competitors in a musical contest were ranked by 3 judges, A, B and C in the following order:

Competitors:	1	2	3	4	5	6	7	8	9	10
Rank by A	1	6	5	10	3	2	4	9	7	8
Rank by B	3	5	8	4	7	10	2	1	6	9
Rank by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common liking in music.

15.8 FURTHER READINGS

A number of good text books are available for the topics dealt with in this unit. The following books may be used for more indepth study.

Richard I. Levin and David S. Rubin, 1996, Statistics for Management. Prentice Hall of India Pvt. Ltd., New Delhi.

Peters, W.S. and G.W. Summers, 1968, Statistical Analysis for Business Decisions, Prentice Hall, Englewood-cliffs.

Hooda, R.P., 2000, Statistics for Business and Economics, MacMillan India Ltd., New Delhi.

Gupta, S.P. 1989, Elementary Statistical Methods, Sultan Chand & Sons: New Delhi.

Chandan, J.S. - Statistics for Business and Economics, Vikas Publishing House Pvt. Ltd., New Delhi.

UNIT 16 SIMPLE LINEAR REGRESSION

Structure

16.0 OBJECTIVES

After going through this unit, you shall be able to:

- explain the concept of regression
- estimate the linear regression
- explain the method of least squares
- apply linear regression methods to given data
- use regression equations for predictions; and
- identify the relationship and difference between correlation and regression coefficient

16.1 INTRODUCTION

In the previous unit we have learnt about simple linear correlation and understood that correlation tells whether exists a relationship between two variable or not but it does not reflect cause and effect relationship between two variables. Therefore, we cannot predict the value of one variable for a given value for other variable. This limitation is removed by regression analysis. In regression analysis, the relationship between variable are expressed in the form of a mathematical equation. It is assumed that one variable is cause and the other is the effect. Please note that regression is a statistical tool which helps understand the relationship between variables and predicts the unknown values of the dependent variable from known values of the independent variable.

16.2 THE CONCEPT OF REGRESSION

In regression analysis we have two types of variables: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

In the simplest case of regression analysis there is one dependent variable and one independent variable. Let us assume that consumption expenditure of a household is related to the household income. For example, it can be postulated that as household income increases, expenditure also increases. Here, consumption expenditure is the dependent variable and household income is the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected n pairs of observations in X and Y. The next step is to find out the nature of relationship between X and Y

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the **linear equation**. This means that the relationship between X and Y is in the form of a straight line and is termed linear regression. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Now the question arises, 'How do we identify the equation form?' There is no hard and fast rule as such. The form of the equation depends upon the reasoning and assumptions made by us. However, we may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the type of equation to be fitted. If the points are more or less in a straight line, then linear equation is assumed. On the other hand, if the points are not in a straight line and are in the form of a curve, a suitable non-linear equation (which resembles the scatter) is assumed.

We have to take another decision, that is, the identification of dependent and independent variables. This again depends on the logic put forth and purpose of analysis: whether 'Y depends on X' or 'X depends on Y'. Thus there can be two regression equations from the same set of data. These are: **i) Y is assumed to be dependent on X (this is termed 'Y on X' line), and ii) X is assumed to be dependent on Y (this is termed 'X on Y' line)**.

You may by now be wondering why the term 'regression', which means 'reduce'. This name is associated with a phenomenon that was observed in a study on the relationship between the stature of father (x) and son (y). It was observed that the average stature of sons of the tallest fathers has a tendency to be less than the average stature of these fathers. On the other hand, the average stature of sons of the shortest fathers has a tendency to be more than the average stature of these fathers. This phenomenon was called regression towards the mean. Although this appeared somewhat strange at that time, it was found later that this is due to natural variation within subgroups of a group and the same phenomenon occurred in most problems and data sets. The explanation is that many tall men come from families with average stature due to vagaries of natural variation and they produce sons who are shorter than them on the whole. A similar phenomenon takes place at the lower end of the scale. Let us discuss simple linear regression.

16.3 SIMPLE LINEAR REGRESSION

When we identify the fact that the correlation exists between two variables, we shall develop an estimating equation, known as regression equation or estimating line, i.e., a methodological formula, which helps us to estimate or predict the unknown value of one variable from known value of another variable. In the words of Ya-Lun-Chou, "regression analysis attempts to establish the nature of the relationship between variables, that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting." For example, if we confirmed that advertisement expenditure (independent variable), and sales (dependent variable) are correlated, we can predict the required amount of advertising expenses for a given amount of sales or vice-versa. Thus, the statistical method which is used for prediction is called regression analysis. And, when the relationship between the variables is linear, the technique is called **simple linear regression**.

Hence, the technique of regression goes one step further from correlation and is about relationships that have been true in the past as a guide to what may happen in the future. To do this, we need the regression equation and the correlation coefficient. The latter is used to determine that the variables are really moving together.

The objective of simple linear regression is to represent the relationship between two variables with a model of the form shown below:

$Y = \beta_0 + \beta_1 X + e_i$

wherein

Y = value of the dependent variable,

β_0 = Y-intercept,

β_1 = slope of the regression line,

X = value of the independent variable,

e_i = error term (i.e., the difference between the actual Y value and the value of Y predicted by the model.

i = represents the observation number, ranges from 1 to n. Thus Y_3 is the third observation of the dependent variable and X_6 is the sixth observation of the independent variable.

16.3.1 Estimating The Linear Regression: Two Variable Case

If we consider the two variables (X variable and Y variable), as discussed earlier, we shall have two regression lines. They are:

i) Regression of Y on X

ii) Regression of X on Y.

The first regression line (Y on X) estimates value of Y for given value of X. The second regression line (X on Y) estimates the value of X for given value of Y. These two regression lines will coincide, if correlation between the variable is either perfect positive or perfect negative.

Illustration 1

The amount of rainfall and agricultural production for ten years are given in Table 16.1

Table 16.1 Rainfall and Agricultural Production

Rainfall (in mm)	Agricultural Production (in tonnes)
60	33
62	37
65	38
71	42
73	42
75	45
81	49
85	52
88	55
90	57

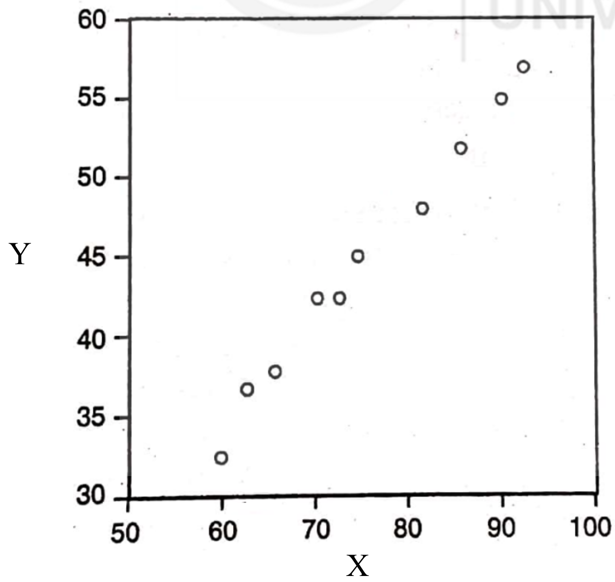


Figure 16.1 Scatter Diagram

We plot the data on a graph paper. The scatter diagram looks something like Figure 16.1 we observe from figure 16.1 that the prints do not like strictly on a straight line. But they show an upward rising tendency where a straight line on the fitted. Let us draw the regression line along with the scatter plot.

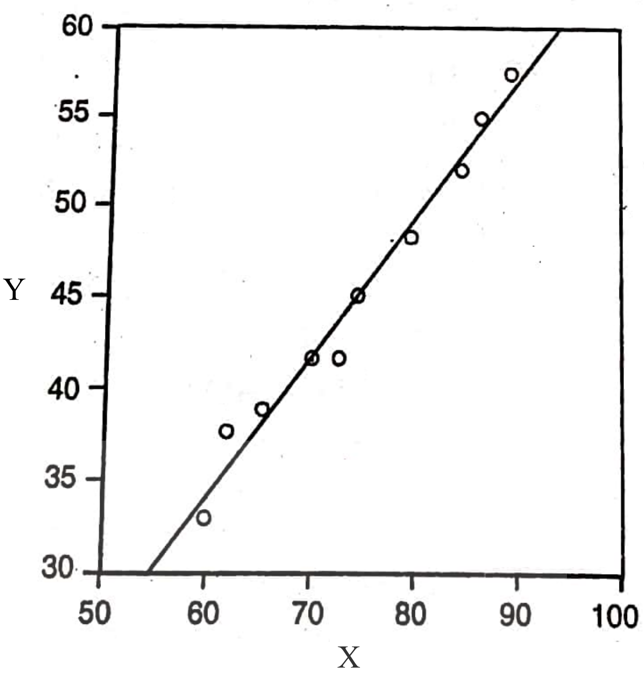


Figure 16.2

When we draw the regression lines with the help of a scatter diagram as shown earlier in Fig. 16.1, we may get an infinite number of possible regression lines for a set of data points. We must, therefore, establish a criterion for selecting the best line. The criterion used is the **Least Squares Method**. According to the least squares criterion, the best regression line is the one that minimizes the sum of squared vertical distances between the observed (X, Y) points and the regression line, i.e., $\Sigma(Y - \hat{Y})^2$ is the least value and the sum of $\Sigma(Y - \hat{Y}) = 0$. It is important to note that the distance between (X, Y) points and the regression line is called the 'error'.

16.3.2 Simple Linear Regression Equations

As we discussed above, there are two regression equations, also called estimating equations, for the two regression lines (Y on X, and X on Y). These equations are, algebraic expressions of the regression lines, expressed as follows:

Regression Equation of Y on X

$\hat{Y} = a + bx$

where, $\hat{Y}$ is the computed values of Y (dependent variable) from the relationship for a given X, 'a' and 'b' are constants (fixed values), 'a' determines the level of the fitted line at Y-axis (Y-intercept), 'b' determines the slope of the regression line, X represents a given value of independent variable.

The alternative simplified expression for the above equation is:

$\hat{Y} - \bar{Y} = byx\ (X - \bar{X})$

$$byx = r\frac{\sigma_y}{\sigma_x} = \frac{(\Sigma XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}$$

Regression equation of X on Y

$\hat{X} = a + by$

Alternative simplified expression is:

$$bxy = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}$$

It is worthwhile to note that the estimated simple regression line always passes through $\bar{X}$ and $\bar{Y}$. The following illustration shows how the estimated regression equations are obtained, and hence how they are used to estimate the value of Y for given X value.

Illustration 2

From the following 12 months sample data of a company, estimate the regression lines.

(Rs. in lakh)

Advertisement Expenditure:	0.8	1.0	1.6	2.0	2.2	2.6	3.0	3.0	4.0	4.0	4.0	4.6
Sales:	22	28	22	26	34	18	30	38	30	40	50	46

Solution:

Table 16.2: Calculations for Least Square Estimates of a Company.

(Rs. in lakh)

Advertising (X)	**Sales**			
	(Y)	X^2	Y^2	XY
0.8	22	0.64	484	17.6
1.0	28	1.00	784	28.0
1.6	22	2.56	484	35.2
2.0	26	4.00	676	52.0
2.2	34	4.84	1156	74.8
2.6	18	6.76	324	46.8
3.0	30	9.00	900	90.0
3.0	38	9.00	1,444	114.0
4.0	30	16.00	900	120.0
4.0	40	16.00	1600	160.0
4.0	50	16.00	2,500	200.0
4.6	46	21.16	2,116	211.6
ΣX=32.8	**ΣY=384**	$\Sigma X^2 = 106.96$	ΣY^2=13368	ΣXY-1150.0

Now we establish the best regression line (estimated by the least square method).

i) We know the regression equation of Y on X is:

$$\hat{Y} - \bar{Y} = byx\ (X - \bar{X})$$

$$\bar{Y} = \frac{384}{12} = 32;\ \bar{X} = \frac{32.8}{12} = 2.733$$

$$byx = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}$$

$$= \frac{1{,}150 - \frac{(32.8)(384)}{12}}{106.96 - \frac{(32.8)^2}{12}} = 100.4/17.31 = 5.8$$

Now Y on X equation is $\hat{Y} - \bar{Y} = byx\ (\hat{X} - \bar{X})$

$\hat{Y} - 32 = 5.8\ (X - 2.733)$

$\hat{Y} = 5.8\ X - 15.85 + 32 = 5.8\ X + 16.15$

Or $\hat{Y} = 16.15 + 5.8X$

which is shown in Figure 16.2. Note that, as said earlier, this line passes through $\bar{X}$ (2.733) and $\bar{Y}$ (32).

ii) We know the regression equation of X on Y is

$$\hat{X} - \bar{X} = bxy\ (Y - \bar{Y})$$

$$bxy = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}} = \frac{1{,}150 - \frac{(32.8)(384)}{12}}{13368 - \frac{(328)^2}{12}} = \frac{100.4}{1{,}080} = 0.093$$

Now X on Y equation is :

$\hat{X} - 2.733 = 0.093\ (Y - 32)$

$\hat{X} - 2.733 = 0.093Y - 2.976$

$\hat{X} = 2.733 - 2.976 - 0.093Y$

$\hat{X} = -0.243 + 0.093Y$

We have the values of $\bar{X} = 2.733$ and $\bar{Y} = 32$

Now we calculate the bxy value:

$$bxy = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}$$

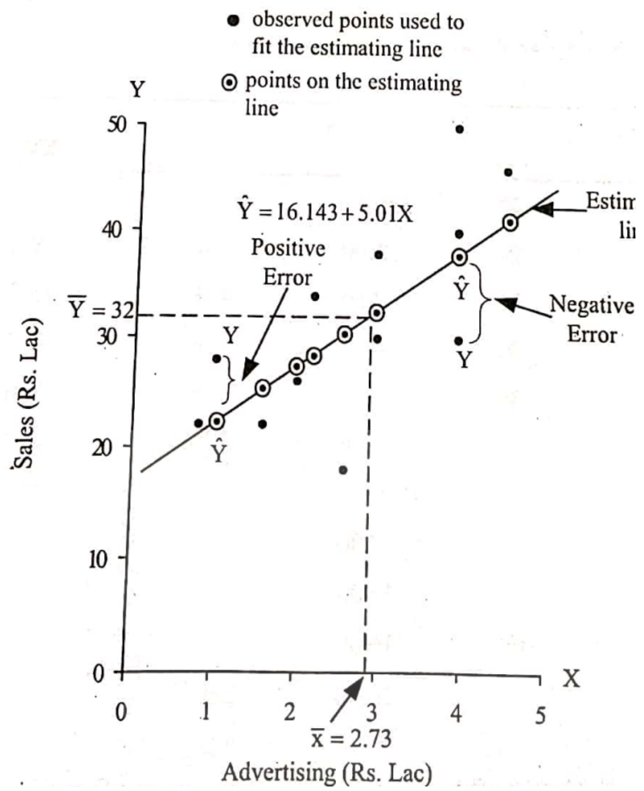


Figure 16.2: Least Squares Regression Line of a Company's Advertising Expenditure and Sales.

It is worthwhile to note that the relationship displayed by the scatter diagram may not be the same if the estimating equation is extended beyond the data points (values) considered in computing the regression equation.

16.3.3 Using Regression for Prediction

Regression, a statistical technique, is used for predictive purposes in applications ranging from predicting demand sales to predicting production and output levels. **In the above illustration 2**, we obtained the regression models of the company. With these models estimate: i) the value of sales when the company decided to spend Rs. 2,50,000 on advertising, and ii) the cost of advertisement when the company desires to reach the target of Rs. 50 Lakhs during the next quarter.

Solution:

i) To find $\hat{Y}$, the estimate of expected sales, we substitute the specified advertising level into the regression model. For example, if we know that the company's marketing department has decided to spend Rs. 2,50,000/- (X = 2.5) on advertisement during the next quarter, the most likely estimate of sales ($\hat{Y}$) is :

$$\hat{Y} = 16.15 + 5.8\,(2.5) = 30.65$$

= Rs. 30,65,000

Thus, an advertising expenditure of Rs. 2.5 lakh is estimated to generate sales for the company to the tune of Rs. 65,000.

ii) To find $\hat{X}$, the estimate cost of advertisement, when company desires to get the target of Rs. 50 lakhs sales during next quarter, the most likely estimation of advertisement cost ($\hat{X}$) is:

$$\hat{X} = -0.25 + 0.093\ (50)$$

$$= -0.25 + 4.65 = 4.4$$

= Rs. 4,40,000.

Thus, the target sales of Rs. 50,00,000/- may be achieved with the estimated cost of Rs. 4,40,000 on advertisement.

Check Your Progress A

You are given the following data relating to age of Autos and their maintenance costs. Obtain the two regression equations by the method of least squares and estimate the likely maintenance cost when the age of Auto is 5 years.

Age of Auto (years)	:	2	4	6	8
Maintainance Cost (Rs.00)	:	10	20	25	30

16.3.4 Method of Least Squares

As discussed earlier that in the method of scattered diagram, we may get infinite numbers of possible regression lines for a set of data points. Therefore, it is necessary to establish a criterion for selecting the next line. The criterion used in the Least Square Method under this method $\sum(Y-\hat{Y})^2$ is the least value and $\sum(Y-\hat{Y})$ is zero.

As we know the basic equation of least square method that y on x equation is: $\hat{Y} = a + bx$ and x on y equation is $\hat{X} = a + by$.

We can obtain the values of the coefficient a and b of the least square regression line through the following equations:

$\sum Y = Na + b\sum x$……… (i)

$\sum XY = a\sum X + b\sum X^2$……… (ii)

Let us take the following illustration for formulation of best regression lines i.e. least square regression lines.

Illustration – 3:

Assume that quantity of agricultural production depends on the amount of rainfall and fit a linear regression to the data given.

Rainfall (in mm)	:	60	62	65	71	73	75	81	85	88	90
Agricultural Production (in tones)	:	33	37	38	42	42	45	49	52	55	57

In this case dependent variable (Y) is quantity of agricultural production and independent variable (X) is amount of rainfall. The regression equation to be fitted is

$Y_i = a + bX_i$

For the above equation we find out the normal equations by the method of least squares. Next we construct a table as follows:

Table 16.3: Computation of Regression Line

X	Y	X^2	XY	$\hat{Y}$	$Y - \hat{Y}$
60	33	3600	1980	33.85	-0.85
62	37	3844	2294	35.34	1.66
65	38	4225	2470	37.57	0.43
71	42	5041	2982	42.03	-0.03
73	42	5329	3066	43.51	-1.51
75	45	5625	3375	45.00	0.00
81	49	6561	3669	49.46	-0.46
85	52	7225	4420	52.43	-0.43
88	55	7744	4840	54.66	0.34
90	57	8100	5130	56.15	0.85
$\sum X = \mathbf{750}$	$\sum Y = \mathbf{450}$	$\sum X^2 = \mathbf{57294}$	$\sum XY = \mathbf{34526}$	$\sum \hat{Y} = \mathbf{450}$	$\sum e_i = \mathbf{0}$

Now we will solve the following equation

$\sum Y = Na + b\sum x$……… (i)

$\sum XY = a\sum X + b\sum X^2$……… (ii)

By substituting values from the above table (16.3) in the above normal equation (i) and (ii), we will get following

$450 = 10a + 750b$ ………… (iii)

$34,526 = 750a + 57{,}294b$ ……. (iv)

Before substituting the above values of the two equations (iii & iv) we have to adjust the value connected with either a or b coefficient as equal by the value of suitable multiplier.

Here, if we multiply the equation (iii) with the value 75 we may equalize the value connected with coefficient a, we will get:

$450 = 10a + 750b \times 75$

$33{,}750 = \cancel{750a} + 56{,}250b$ (adjusted of iii)

(-) $34{,}526 = \cancel{750a} + 57{,}294b$ $(as\ (iv)$

$-$ 776 = . $-$ 1,044 b

Now, $b = \frac{-776}{-1{,}044} = 0.743$

We will find the value of coefficient a by considering the equation (iii) above i.e.

$$450 = 10a + 750\ (0.743)$$

$$450 = 10a + 557.25$$

$$-10a = 557.25 - 450$$

$$a = \frac{107.25}{-10} = -10.73$$

So the regression line is $\hat{Y} = -10.73 + 0.743X$.

Notice that the sum of errors $\sum e_i$ for the estimated regression equation is zero (see the last column of Table 16.3).

The computation given in Table 16.3 often involves large numbers and poses difficulty. Hence we have a short-cut method for calculating the values of a and b from the normal equations.

Under this shortcut method:

$$a = \bar{Y} - b\bar{X}$$

$$b = \sum XY / \sum X^2$$

Here, the denotion $x = X - \bar{X}$ means deviation of X (independent variable) from the value of $\bar{X}$

The denotion $y = Y - \bar{Y}$ means the deviation of Y (dependent variable) from the $\bar{Y}$.

Hence $xy = (X - \bar{X})(Y - \bar{Y})$

Since these formulae are derived from the normal equations we get the same values for a and b in this method also. For the data given we compute the values of a and b by this method. For this purpose we construct Table 16.4.

Table 16.4 Computation of Regression Line (short-cut method)

	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$x-X^2$	xy
	60	33	–15	–12	225	180
	62	37	–13	–8	169	104
	65	38	–10	–7	100	70
	71	42	–4	–3	16	12
	73	42	–2	–3	4	6
	75	45	0	0	0	0
	81	49	6	4	36	24
	85	52	10	7	100	70
	88	55	13	10	136	130
	90	57	15	12	225	180
Total	**750**	**450**	**0**	**0**	**1044**	**776**

$$\bar{X} = \frac{750}{10} = 75 \text{ and } \bar{Y} = \frac{750}{10} = 45$$

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} = \frac{776}{1044} = 0.743$$

$$a = \bar{Y} - b\bar{X} = 45 - 0.743 \times 75 = -10.73$$

Thus the regression line in this method also $\hat{Y} = -70.73 + 0.743X$

Coefficient b is called the regression coefficient. This coefficient reflects the amount of increase in Y when there is a unit increase in X. In regression equation the coefficient b = 0.743 implies that if rainfall increase by 1 mm. agricultural production will increase 0.743 thousand tonne.

You might have observed that, it may be noted, this short cut method is the easiest for calculation only when the arithmetic mean of both X and Y series are having absolute value (10, 25,32 etc.) not in fraction value (i.e. 10.62, 53.12, 83.95 etc.).

Check you Progress B

From the following data, obtain the two regression equation by Least squares method and estimate the sales if the purchases are 95 lakhs. The data is Rs. In lakhs.

Sales	:	91	97	108	121	67	124	51	73	111	57
Purchases	:	71	75	69	97	70	91	39	61	80	47

16.4 RELATIONSHIP BETWEEN CORRELATION AND REGRESSION COEFFICIENTS

The following points about the regression should be noted:

1) The geometric mean of the two regression coefficients (byx and bxy) gives coefficient of correlation.

   That is, $r = \pm\sqrt{(bxy)\ (byx)}$

   Consider the values of regression coefficients from the previous illustration to know the degree of correlation between advertising expenditure and sales.

   $r = \pm\sqrt{0.093 \times 5.801} = 0.734$

2) Both the regression coefficients will always have the same sign (+ or –).

3) Coefficient of correlation will have the same sign as that of regression coefficients. If both are positive, then r is positive. In case both are negative, r is also negative. For example, bxy = –1.3 and byx = –0.65, then r is:

   $\pm\sqrt{-1.3x - 0.65} = -0.919$ but not $+0.919$

4) Regression coefficients are independent of change of origin, but not of scale.

16.5 DIFFERENCE BETWEEN CORRELATION AND REGRESSION

After having an understanding about the concept and application of simple correlation (discussed in unit 15) and simple regression, we can draw the difference between them. They are:

1) Correlation coefficient 'r' between two variables (X and Y) is a measure of the direction and degree of the linear relationship between them, which is mutual. It is symmetric (i.e., $r_{xy} = r_{yx}$) and it is inconsiderable which, of X and Y, is dependent variable and which is independent variable. Whereas regression analysis aims at establishing the functional relationship between the two variables under study, and then using this relationship to predict the value of the dependent variable for any given value of the independent variable. It also reflects upon the nature of the variables (i.e., which is the dependent variable and which is independent variable). Regression coefficients, therefore, are not symmetric in X and Y (i.e., $r_{xy} \neq r_{yx}$).

2) Correlation need not imply cause and effect relationship between the variables under study. But regression analysis clearly indicates the cause and effect relationship between the variables. The variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.

3) Correlation coefficient 'r' is a relative measure of the linear relationship between X and Y variables and is independent of the units of measurement. It is a number lying between ±1. Whereas the regression coefficient byx (or bxy) is an absolute measure representing the change in the value of the variable Y (or X) for a unit change in the value of the variable X (or Y). Once the functional form of the regression curve is known, by susbstituting the value of the dependent variable we can obtain the value of the independent variable which will be in the unit of measurement of the variable.

4) There may be spurious (non-sense) correlation between two variables which is due to pure chance and has no practical relevance. For example, the correlation between the size of shoe and the income of a group of individuals. There is no such thing as spurious regression.

5) Correlation analysis is confined only to study of linear relationship between the variables and, therefore, has limited applications. Whereas regression analysis has much wider applications as it studies linear as well as non-linear relationships between the variables.

16.6 LET US SUM UP

In this unit, fundamental concepts and techniques of simple linear regression have been discussed i.e.in case of two variables only.

Once it is identified that correlation exists between the variables, an estimating equation known as regression equation could be developed by the least squares method for prediction. Relationship between correlation and regression coefficient and the conceptual differences between correlation and regression have been highlighted. The techniques of regression analysis are widely used in business decision making and data analysis.

16.7 KEY WORDS

Linear Relationship: The relationship between two variables described by a straight line.

Least Squares Criterion: The criterion for determining a regression line that minimizes the sum of squared errors.

Simple Regression Analysis: A regression model that uses one independent variable to explain the variation in the dependent variable.

16.8 ANSWERS TO CHECK YOUR PROGRESS

A) $Y\ on\ X : \hat{Y} = 5 + 3.25x$

$X\ on\ Y : \hat{X} = -3 + 0.297y$

B) Y = 14.81 + 0.613x

X = -5.2 + 1.36y

Estimation = Rs. 124 lakhs

16.9 TERMINAL QUESTIONS

1) What do you understand by the term regression? Explain its significance.

2) Distinguish between correlation and regression.

3) Discuss about least square method.

4) A personal manager of a firm is interested in studying as to how the number of worker absent on a given day is related to the average temperature on that day. A random sample of 12 days was used for the study. The data is given below:

No. of Workers absent	6	4	8	9	3	8	5	2	4	10	7	6
Average temperature (^{0}C)	12	30	15	18	40	30	45	35	23	15	25	35

a). State the independent variable and dependent variable.

b). Draw a scatter diagram.

c). Determine the regression lines (i) X on Y and (ii) Y on X

5) The following table gives the demand and price for a commodity for 6 days.

Price (Rs.):	4	3	6	9	12	10
Demand (mds):	46	65	50	30	15	25

a) Develop the estimating regression equations.

b) Predict demand for price (Rs.) = 5,8 and 11.

6) A sales manager of a soft drink company is studying the effect of its latest advertising campaign. People chosen at random were called and asked how many bottles they had bought in the past week and how many advertisements of this product they had seen in the past week.

No. of ads (X)	4	0	2	7	3	4	2	6
Bottles Purchased (Y)	6	5	4	16	10	9	6	14

a). Develop the regression equations that best fits the data through the method of least squares.

b). Predict Y value when X = 78.

c). Predict X value when Y = 20.

7) Obtain the lines of regression from the following data.

X	25	22	28	26	35	20	22	40	20	18
Y	18	15	20	17	22	14	16	21	15	14

i) Estimate the value of Y if the value of X is 25, and

ii) Estimate the value of X if the value of Y is 45.

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 for assessment. These are for your practice only.

16.10 FURTHER READINGS

A number of good text books are available for the topics dealt with in this unit. The following books may be used for more indepth study.

Richard I. Levin and David S. Rubin, 1996, Statistics for Management. Prentice Hall of India Pvt. Ltd., New Delhi.

Peters, W.S. and G.W. Summers, 1968, Statistical Analysis for Business Decisions, Prentice Hall, Englewood-cliffs.

Hooda, R.P., 2000, Statistics for Business and Economics, MacMillan India Ltd., New Delhi.

Gupta, S.P. 1989, Elementary Statistical Methods, Sultan Chand & Sons: New Delhi.

Chandan, J.S. - Statistics for Business and Economics, Vikas Publishing House Pvt. Ltd., New Delhi.

UNIT 17 INDEX NUMBERS

Structure

17.0 OBJECTIVES

After studying this unit, you should be able to:

- define and explain the meaning of Index numbers,
- discuss the characterstics and uses of Index numbers
- identify and avoid various issues faced while developing index numbers for some special purposes,
- discuss the classification of index numbers
- construct and calculate index numbers applying different methods, and
- describe the limitations of index numbers to avoid errors in interpretations.

17.1 INTRODUCTION

In the previous block-5 we have learnt how to calculate the statistical data relating bi-variant in nature by applying the statistical devices. They are **simple linear correlation and simple linear regression** which provide to establish the relationship between the two variables. In this unit we shall discuss the methods of constructing various types of index numbers for different purposes. This device is an extension of the time series analysis because an index number combines two or more time series variables related to non-comparable units. You would have read in newspapers or heard on the television/the radio that the cost of living index has increased by so many points, hence for government employees another slab of Dearness Allowance has been declared. Probably you might have wondered what is this cost of living index?

Many of you must also be aware of the Stock Exchange Share Price Index – commonly referred to as BSE SENSEX or, more recently, NSE SENSEX. In fact, these various types of index series have come to be used in many activities such as industrial production, export, prices, etc. In this Unit, you will study and understand the meaning and uses of index numbers, various problems resulting from the incorrect use of index numbers, methods for construction of various index numbers, and their limitations.

17.2 MEANING AND CONCEPT OF INDEX NUMBERS

When we talk that the general level of industrial production has registered an increase of 4 per cent, it is obvious that we are referring to the production of all those items that are produced by the industrial sector. However, production of some of these items may be increasing while that of others may be decreasing or may remain constant. The rate of increase or decrease and the units in which these items are expressed may differ. For instance, cement may be quoted per kg, cloth may be per meters, cars may be per unit etc. In such a situation, when the purpose is to measure the changes in the average level of prices or production of industrial products for comparing over a time or with respect to geographic location, it is not appropriate to apply the technique of measure of central tendency because it is not useful when series are expressed in different units or/and in different items.

It is in these situations, that we need a specialised average, known as index numbers. These are often termed as 'economic barometers'.

An index number may be defined as a special average which helps in comparison of the level of magnitude of a group of related variables under two or more situations.

Index numbers are a series of numbers devised to measure changes over a specified time period (the time period may be daily, weekly, monthly, yearly, or any other regular time interval), or compare with reference to one variable or a group of related variables. Thus, each number in a series of specified index number is:

a) A pure number i.e., it does not have any unit.

b) Calculated according to a pre-determined formula.

c) Generated at regular time intervals, sometimes during the same time interval at different places.

d) The regular generation of numbers form a chronological series.

e) With reference to some specified period and number known as base period and base number, the latter is always 100. For example, if the consumer price index, with base year 1996 is calculated to be 180 for the year 2003, it means that consumer prices have increased by 80 per cent in 2003 as compared to the prices prevalent in 1996.

17.2.1 Characteristics of Index Number

Main Characteristics of the measurement of Index number is as follows:

1) Relative measurement
2) Specialized average
3) Measurement of changes not capable of direct measurement
4) Measurement of common characteristics of a group of items
5) Comparison on the basis of time or place
6) Expressed in percentage
7) Universal use

1) **Relative measurement**: Index number are used for comparing relative change in a variable or group of variables at different point of time or place.

2) **Specialized average**: Index number is a special type of average that provides a measurement of relative changes in a variable or group of variables.

3) **Measurement of changes not capable of direct measurement**: with the help of index number, we can measure the changes in magnitude which are not capable in direct measurement due to their complex nature.

4) **Measurement of common characteristics of a group of items**: Index express the common characteristics of a group of items change in index does not always mean that there is a change in all the variables for example, an increase in price index does not mean that price of all the commodities are increased.

5) **Comparison on the basis of time or place**: Index number is used to measure the relative changes either on the basis of time or on the basis of place for example production of wheat in Uttar Pradesh for two different period or production of wheat in Uttar Pradesh and Haryana in the same period.

6) **Expressed in percentage**: Index numbers are expressed in percentages to show the relative change through the sign of percentage (%) is never used.

7) **Universal use**: The technique of Index number is being used extensively in all the fields now a days be it changes in production, trade, etc.

17.3 USES OF INDEX NUMBERS

Though originally the index number was developed for measuring the effect of change in prices, today they have become indispensable for analyzing the data related to business and economic activity. This statistical tool can be used in several ways as follows:

1) Decision makers use index numbers as part of intermediate computations to understand other information better. Nominal income can be transformed into real income. Similarly, nominal sales into real sales & so on …, through an appropriate index number. Consumer price index, also known as cost of living index, is arrived at for a specified group of consumers in respect of prices of specific commodities and services which they usually purchase. This index serves as an indicator of 'real' wages (or income) of the consumers. For example, an individual earns Rs. 100/- in the year 1970 and his earnings increase to Rs. 300/- in the year 1980. If during this period, consumer price index increases from 100 to 400 then the consumer is not able to purchase the same quantity of different commodities with Rs. 300, which he was able to purchase in the year 1970 with his income of Rs. 100/-. This means the real income has declined. Thus real income can be calculated by dividing the actual income by dividing the consumer price index:

$$Real\ Income\ in\ 1980 = \frac{Actual\ income\ in\ 1980}{Consumer\ price\ index\ of\ 1980} \times 100$$

$$= \frac{300}{400} \times 100 = Rs. 75/- \text{ with respect to 1970 as base year}$$

Therefore, the consumer's real income in the year 1980 is Rs. 75/- as compared to his income of Rs. 100/- in the year 1970. We can also say that because of price increase, even though his income has increased, his purchasing power has decreased.

2) Different types of price indices are used for wage and salary negotiations, for compensating in price rise in the form of DA (Dearness Allowance).

3) Various indices are useful to the Government in framing policies. Some of these include taxation policies, wage and salary policies, economic policies, custom and tariffs policies etc.

4) Index numbers can also be used to compare cost of living across different cities or regions for the purpose of making adjustments in house rent allowance, city compensatory allowance, or some other special allowance.

5) Indices of Industrial Production, Agricultural Production, Business Activity, Exports and Imports are useful for comparison across different places and are also useful in framing industrial policies, import/export policies etc.

6) BSE SENSEX is an index of share prices for shares traded in the Bombay Stock Exchange. This helps the authorities in regulating the stock market. This index is also an indicator of general business activity and is used in framing various government policies. For example, if the share prices of most of the companies comprising any particular industry are continuously falling, the government may think of changes in its policies specific to that industry with a view to helping it.

7) Sometimes, it is useful to correlate index related to one industry to the index of another industry or activity so as to understand and predict changes in the first industry. For example, the cement industry can keep track of the index of construction activity. If the index of construction activity is rising, the cement industry can expect a rise in demand for cement.

8) If you are informed that the price of one kilogram sunflower oil was Rs.0.50 per kg. in the year 1940 and in the year 1980 it was Rs. 30 and in the year 2004 it is reported to be Rs. 70, per kg in the year 2018 the price was Rs. 160 per kg, and if you are asked this question: shall sunflower oil be sold again in the future for either Rs.0.50 or Rs. 30 or Rs. 70 per kg? Surely, you answer would be 'No'.

17.4 ISSUES IN CONSTRUCTION OF INDEX NUMBERS

There are three major issues which may be faced in the construction of index numbers. They are: 1) Collection of Data; 2) Selection of Base Year and 3) Selection of Appropriate Index. Let us discuss them in detail:

1) Collection of Data: Data collection through a sample method is one of the issues in the construction of index numbers. The data has to be as reliable, adequate, accurate, comparable, and representative, as possible. Here a large number of questions need to be answered. The answers ultimately depend on the purpose and individual judgement. For example, one needs to decide the following:

i) **Identification of Commodities to be Included**: How many and which category of commodities to include? A large number of items may be present. It is not possible to include all of them, only those items deserve to be included in the construction of an index number as would make it more representative. For example, if we are required to construct indices for shares on the Bombay Stock Exchange, there are several shares listed and traded, it is not possible to include all of them. Therefore, it has to be decided which sample number of shares (may be 30 or 40) should represent the general movement of share prices of the Bombay Stock Exchnage. Therefore, it is worthwhile to note that the selection of items must be deliberate and in keeping with the relevance and significance of each individual item to the purpose for which the index is constructed.

ii) **Sources of Data**: From where to collect data? It is an important and difficult issue. The source depends on the information requirement. For example, one may need to collect prices and quantities consumed related to certain commodities for a consumer price index. However, there may be a large number of retailers and wholesalers, selling the commodities, and quoting different prices. To get the details, only a few representative shops (which represent the typical purchasing points of the people under question) need to be selected. Thus, based on a representative sample survey, sources should be from where accurate, adequate, and timely data can be available.

iii) **Timings of Data Collection**: It is also equally important to collect the data at an appropriate time. Referring to the example of consumer price index, prices are likely to vary on different days of the month. For certain commodities prices may vary at different times of the same day. Take an example, vegetable prices are usually high in the morning when fresh vegetables arrive and are low in the late evening when sellers are closing for the day and wish to clear the perishable stock. For each commodity, individual judgment needs to be exercised to represent reality and to serve the purpose for which an index is to be used.

2) Selection of Base Year: A base period is the reference period for comparing and analysing the changes in prices or quantities in a given period. For many index number series, value of a particular time period, usually a year, is taken as reference period against which all subsequent index numbers in the series are calculated and compared.

In some other cases, especially when cost of living needs to be compared across the cities, the value of cost of living prevailing in a selected city is taken as a base against which cost of living in other cities is compared.

In yet other cases, we may be required to compare one index number series against another series. In such a context, a 'base' common to all series is more appropriate.

In the light of the above considerations, therefore, the period/year selected as base period/year must be a 'normal' period. Normal period is a period with price or quantity figures neither too low, nor too high. It should not have been affected by abnormal occurrences, such as floods, (if interested in agricultural production), wars, sudden recession etc. What is normal should also be decided keeping in view the purpose of constructing an index number, and the specific situation.

3) Selection of an Appropriate Index: Different methods of indices give different results, when applied to the same data. Utmost care must be taken in selection of a formula which is the most suitable for the purpose. Whether to use an unweighted or weighted index is a difficult question to answer. It depends on the purpose for which the index number is required to be used. For example, if we are interested in an index for the purpose of negotiating wages or compensating for price rise, only a weighted index would be worthwhile to use.

Which weights to be used? Whether base year quantities or current year quantities or some other weights are to be used is an important question to answer. Weights which realistically reflect the relative importance of items included in the construction of an index is perhaps the only answer. The purpose for which an index is needed will of course remain a vital factor to reckon with.

17.5 CLASSIFICATION OF INDEX NUMBERS

There are three principal types of indices: price indices, quantity indices, and value indices.

Price Indices: This type of indices is the most frequently used. Price indices consider prices of a commodity or a group of commodities and compare changes of prices from one period to another period and also compare the difference in price from one place to another. For example, the familiar Consumer Price Index measuring overall price changes of consumer commodities and services is used to define the cost of living.

Quantity Indices: The major focus of consideration and comparison in these indices are the quantities either of a single commodity or a group of commodities. For example, the focus may be to understand the changes in the quantity of paddy production in India over different time periods. For this purpose, a single commodity's quantity index will have to be constructed. Alternatively, the focus may be to understand the changes in food grain production in India, in this case all commodities which are categorized under food grains will be considered while constructing the quantity index.

Value Indices: Value indices actually measure the combined effects of price and quantity changes. For many situations either a price index or quantity index may not be enough for the purpose of a comparison. For example, an index may be needed to compare cost of living for a specific group of persons in a city or a region. Here comparison of expenditure of a typical family of the group is more relevant. Since this involves comparing expenditure, it is the value index which will have to be constructed. These indices are useful in production decisions, because it avoids the effects of inflation.

The formula, therefore is:

$$Value\ Indices\ Iv = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$$

Check Your Progress A

1) State with reasons, on your agreement or disagreement on the following statements.

   a) Index numbers are specialised averages.

   b) The index number for a base year is always zero.

   c) A value index measures either price or quantity changes.

d) In times of inflation, a quantity index provides a better measure of actual output than a corresponding value index.

e) Through appropriate indices, normal increase can be transformed into real income.

f) Probability sampling is the most appropriate method for selecting commodities while constructing indices.

g) A base period may be described as a "normal" period if it is the most recent period for which we have data.

2) In magazines and newspapers you might have come across many index numbers. Name four such index numbers and briefly state what does each one of them indicate?

3) List out the problems that arise in connection with the construction of an index number.

4) Try to cite one example each where (a) price index, (b) quantity index, and (c) value index is not appropriate.

17.6 METHODS OF CONSTRUCTING INDEX NUMBERS

In the previous section, we have discussed different types of indices, i.e., price indices, quantity indices, and value indices. We shall now focus on the construction of price and quantity indices and their limitations.

Different formulae have been introduced by statisticians for constructing composite index numbers. They may be categorized into two broad groups as given below:

I) Unweighted Indices; and

II) Weighted Indices

The formula and its use in constructing each category of indices, listed above, are discussed in the following sections. Let us first acquaint ourselves with the symbols used in construction of index numbers. They are as follows:

P_0 denotes price per unit of a commodity in the base period.

P_1 denotes price per unit of the same commodity in the current period (current period is one in which the index number is calculated with reference to the base period).

Similar measurements are assigned to Q_0, Q_1 and V_0, V_1.

Capital letters P, Q, and V are used for denoting price index, quantity index, and value index numbers, respectively.

Thus, P_{01} refers to price index for period 1. (P_1) with respect to base period (P_0). Similar meanings are assigned to quantity (Q_{01}) and value (V_{01}) indices. It may be noted that indices are expressed in per cent.

17.6.1 Unweighted Index Numbers:

This type of indices are also referred to as simple index numbers. In this method of constructing indices, weights are not expressly assigned. These are further classified under two categories:

1) Simple Aggregative Index

2) Simple Average of Relatives Index

Let us study the construction of indices under these two methods:

1) **Simple Aggregative Index**: This is the simplest and least satisfactory method of constructing indices. In the case of price indices, through this method, the total of unit cost of each commodity in the current year is divided by the total of unit cost of the same commodity in the base year and the quotient is multiplied by 100. Symbolically,

$$P_{01} = \left(\frac{\sum P_1}{\sum P_0}\right) \times 100$$

Similarly, the quantity index may be expressed as:

$$Q_{01} = \left(\frac{\sum q_1}{\sum q_0}\right) \times 100$$

Illustration 1: By considering the hypothetical data for the year 1990 and 2000 the following computation was done for construction of price index and quantity index.

Table 17.1 Computation of Index by Simple Aggregative Method

Item	Year 1990		Year 2000	
	Price (Rs.)	Quantity	Price (Rs.)	Quantity
Wheat	700	4 qts	950	3.5 qts
Clothing	200	30 mts	300	35 mts
Gas	150	4 cylinder	220	6 cylinders
Electricity	0.80	800 units	1.10	1,000units
House Rent	400	1 dwelling	800	1 dwelling
	1450.80 $\sum P_0$	839 $\sum q_0$	2271.1 $\sum p_1$	1045.5 $\sum q_1$

The price index for the year 2000 with reference to base year 1990 the simple aggregative method is

$$P_{01} = \left(\frac{\sum P_1}{\sum P_0}\right) \times 100 = \frac{2271.1}{1450.8} \times 100 = 156.54$$

Thus, the prices in respect of commodities considered in the index have shown an increase of 56.54 per cent in 2000 as compared to 1990.

This method suffers from the following two limitations:

1) The unit size affects the index number. For instance, in the above illustration if the price of wheat was quoted in terms of per kg. Rs. 7/- in 1990 and Rs. 9.5 in 2000) the index might be very different.

2) Relative importance of different commodities is not reflected in the index. For example, in the above illustration a total of Rs. 2,800/- is spent on wheat, which is the most important item of expenditure. This is not reflected in this method.

Analogously, the Quantity Index by the simple aggregate method is:

$$Q_{01} = \left(\frac{\sum q_1}{\sum q_0}\right) \times 100$$

Consider the illustration 1 for quantity index

$$Q_{01} = \frac{1045.5}{839} \times 100 = 124.61$$

Here, you should note that the 'P' in the formulae of price index will be replaced by 'q' in constructing index. This expression is applicable to the formulae of different methods.

Limitation: The units of quantities being different cannot be added and the quantities do not represent appropriate variables for the purpose of comparing expenditure.

2) Simple Average of Relatives Index

In this method of constructing price index, first of all price relatives have to be computed for the different items included in the index then the average of these is calculated simbolically,

$$P_{01} = \frac{\sum\left(\frac{P_1}{P_0} \times 100\right)}{N} \text{ or } \frac{\underline{\textit{Sum of the Price Relatives}}}{\textit{No.of Items}}$$

Using the same data by considering only prices given in the illustration-1, the computation of price index as simple average of price relatives is as follows:

Illustration 2

Table 17.2: Computation of Index by Simple Average of Relatives Method

Item	Units	Year 1990 Price (Rs.)	Year 2000 Price (Rs.)	Price relatives $\frac{P_1}{P_0} \times 100$
Wheat	Qts	700	950	(950 / 700) × 100 = 135.7
Clothing	Mts	200	300	(300 / 200) × 100 = 150.0
Gas	Cylinder	150	220	(220 / 150) × 100 = 140.7
Electricity	Units	0.80	1.10	(1.10 / 0.80) × 100 = 137.5
House Rent	dwelling	400	800	(800 / 400) × 100 = 200
	N = 5			$\sum\left(\frac{P_1}{P_0} \times 100\right) = 763.9$

$$P_{01} = \frac{\sum\left(\frac{P_1}{P_0} \times 100\right)}{N} = \frac{763.9}{5} = 152.78$$

Thus, the index of simple average of price relatives shows 52.78 per cent increase in price.

For construction of Quantity Index, quantity relatives should be obtained and averaged. The formula for quantity index in this method is:

$$Q_{01} = \frac{\sum\left(\frac{q_1}{q_0} \times 100\right)}{N}$$

Which you may compute on your own by using the data given in Illustration-1.

This method also has its limitations. First, each price/quantity relative is given equal importance, which is not realistic. Secondly, the arithematic mean is not the right type of average for ratios, and percentages.

Check Your Progress B

a) Calculate:

i) The price index number by simple aggregative and average of relatives methods from the following data (price per kg).

Commodities	**Price in 2015 (Rs.)**	**Price in 2018 (Rs.)**
Apple	35	60
Mango	30	45
Watermelon	5	10

ii) What are the limitations of both the methods?

b) For the give data find:

Simple aggregative Index for the year 2017 over the year 2016.

(i) Simple Aggregative Index for the year 2018 over the year 2017.

Commodity	**2016 (Rs.)**	**2017 (Rs.)**	**2018 (Rs.)**
A (100 gm)	12	15	15.60
B (per piece)	3	3.60	3.30
C (per kg)	5	6	5.70
Aggregate	20	24.60	54.60

17.6.2 Weighted Index Numbers

In the earlier two methods each item received equal weight/importance in the construction of an index, whereas in the weighted index methods, weights are expressly assigned to each item which is included in an index construction.

This weighting allows us to consider more information than just the change in price/ quantity over time. The problem only is to decide how much weight (importance) to consider for each of the items included in the sample. This is further divided into two methods.

1) **Weighted Aggregative Index, and**

2) **Weighted Average of Relatives Index.**

Let us discuss these two methods one after another.

1) Weighted Aggregative Index: In this group, we shall study three specific methods commonly used in business research. They are: (a) Laspeyre's index, (b) Paasche's index, and (c) Fisher's ideal index. After understanding the concepts of the three indices we will take up an illustration for construction of these indices.

a) **Laspeyre's Index:** In this method, weights assigned to each commodity are the quantities consumed in the base year for price indices. For quantity index weights used are the prices of commodities in the base year. Thus, according to Laspeyre:

Price Index $(P_{01}{}^{La}) = \left(\frac{\sum P_1 q_0}{\sum P_0 q_0}\right) \times 100,$ and

Quantity Index = $(Q_{01}{}^{La}) = \left(\frac{\sum q_1 P_0}{\sum q_0 P_0}\right) \times 100,$

It is to be noted that this method is most popular for constructing "Consumer Price Index". It is, therefore, considered as aggregate expenditure method which is one of the methods for constructing Consumer Price Index.

Since each index number depends upon price and quantity of the same base year, the researcher can compare the index of one period directly with the index of another period. For instance, assume that the cement price index is 115 in 1995 and 143 in 2001, taking 1991 as base year. The firm concludes that the price level of cement has increased by 15 per cent from 1991 to 1995 and has increased 43% from 1991 to 2000.

b) **Paasche's Index**: In this method, quantities consumed in the current year are used as weights in construction of price indices, where as in construction of quantity index, weights used are the prices of items in the current year. Thus according to Paasche:

Price Index $(P_{01}{}^{Pa}) = \left(\frac{\sum P_1 q_1}{\sum P_0 q_1}\right) \times 100,$ and

Quantity Index = $(Q_{01}{}^{Pa}) = \left(\frac{\sum q_1 P_1}{\sum q_0 P_1}\right) \times 100,$

Comparison of Laspeyre's and Paasche's Indices

From the practical point of view, Laspeyre's index is usually preferred over Paasche's index. This is because as long as base period is fixed, the weights assigned will remain unchanged. Therefore, calculations and comparisons are easier. On the other hand, weights in Paasche's formula continue to change with the change in the current year so that the price index for every year has to be computed using fresh/different weights.

Another interesting property of Laspeyre's index is that it tends to overestimate the value of indices. It is argued that when prices increase, the consumers reduce the consumption of commodities (which are price elastic) for which price rise has been highest. Thus the use of base year quantities increases the value of the numerator, thus increasing the value of index number. The same is true when prices are falling. The Paasche's index, on the other hand, has a tendency to underestimate. This is because when prices are rising, reduced current quantities are used as weights which reduces the value of the index. When price changes are not very rapid, there is not much difference between the index values given by the two methods.

c) **Fisher's Ideal Index**: Irving Fisher used geometric mean of the Laspeyre's and Paache's indices to overcome the shortcomings of the both. Thus,

Price Index $(P_{01}{}^{F}) = \sqrt{\left(\frac{\sum P_1 q_0}{\sum P_0 q_0}\right)\left(\frac{\sum P_1 q_1}{\sum P_0 q_1}\right)} \times 100$

Analogously, Fisher's quantity index is

Quantity Index = $(Q^{F}{}_{01}) = \sqrt{\left(\frac{\sum q_1 P_0}{\sum q_0 P_0}\right)\left(\frac{\sum q_1 P_1}{\sum q_0 P_1}\right)} \times 100$

Thus fisher's ideal index of price/quantity =

$$\sqrt{\text{Laspeyre's Index} \times \text{Paasche's Index}}$$

Fisher's index is superior because it uses geometric mean (which is best applicable for average of ratios and percentages) of Laspeyre's and Paache's indices. Also, because it is comparatively free from bias of over estimation and under estimation. Fisher's index satisfies the requirement of time reversal test and factor reversal test. This index is, therefore, called ideal index. So far we have discussed the three different indices of weighted aggregates method.

For illustration, let us observe the following data of 2013 and 2018, and also required computation for construction of (i) Laspeyre's, (ii) Paasche's, and (iii) Fisher's indices made in the table.

Illustration 3: Table 17.3 Computation of Weighted Aggregated Index

Commodity	Year 2013 (Base Year)		Year 2018 (Current Year)		$P_0 q_0$	$P_1 q_0$	$P_0 q_1$	$P_1 q_1$
	Prices (P_0)	Qty. (q_0)	Prices (P_1)	Qty. (q_1)				
A	800	6	950	8	4800	5700	6400	7600
B	600	3	800	4	1800	2400	2400	3200
C	400	5	425	4	2000	2125	1600	1700
D	250	2	300	2	500	600	500	600
					$\sum P_0 q_0$ = 9100	$\sum P_1 q_0$ = 10824	$\sum P_0 q_1$ = 10900	$\sum P_1 q_1$ = 13100

i) **Laspeyre's Price Index** or $P_{01}^{La} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$

$$= \frac{10824}{9100} \times 100 = 118.94$$

This shows that prices for the group (sample commodities) have increased by 18.94 per cent in 2018 as compared to those prevailing in 2013.

The quantity index according to Laspeyre's formula is computed as shown below:

$$(Q_{01}) = \frac{\sum q_1 P_0}{\sum q_0 P_0} \times 100$$

The sum of $q_1\ P_0$ and $q_0\ P_0$ may be taken from the Table 17.3 as $\sum P_0 q_1 = \sum q_1 p_0$, and $\sum P_0 q_0 = \sum q_0 p_0$.

$$Q_{01}^{Pa} = \frac{10900}{9100} \times 100 = 119.78$$

This shows a 19.78 percent increase in aggregate quantity consumption for this group in 2018 as compared to 2013.

ii) **Paache's Price Index or** $(P_{01}^{Pa}) = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$

$$= \frac{13100}{10900} \times 100 = 120.18$$

Thus, according to the Paache's Index the price index reveals an increase of 20.18 per cent in prices in 2018 as against 2013.

Analogously, Paasche's quantity index is

$$(Q_{01}^{Pa}) = \left(\frac{\sum q_1 P_1}{\sum q_0 P_1}\right) \times 100$$

The values of $\sum q_1 P_1$ and $\sum q_0 P_1$ in the Table 17.3, as they are equivalen to $\sum P_1 q_1$ and $\sum P_1 q_0$ respectively.

Thus, $(Q_{01}^{Pa}) = \frac{13100}{10824} \times 100 = 121.03$

It shows a 21.03 per cent increase in quantity consumption for this group in 2018 as compared to 2013.

iii) **Fisher's Index or** $(\boldsymbol{P_{01}^F}) = \sqrt{\left(\frac{\sum P_1 q_0}{\sum P_0 q_0}\right)\left(\frac{\sum P_1 q_1}{\sum P_0 q_1}\right)} \times 100$

$$(P_{01}^F) = \sqrt{\left(\frac{10824}{9100}\right)\left(\frac{13100}{10900}\right)} \times 100 \quad = \sqrt{1.43} \times 100 = 119.55$$

Therefore, Fisher index value is comparatively free from bias of underestimation and overestimation as in Laspeyre's and Paachre's indices. However, it is more complicated to construct.

Fisher's Quantity Index or $(\boldsymbol{Q_{01}^F}) = \sqrt{\left(\frac{\sum q_1 P_0}{\sum q_0 P_0}\right)\left(\frac{\sum q_1 P_1}{\sum q_0 P_1}\right)} \times 100$

which you may compute and interpret on your own using the data in the Table 17.3.

Illustration-4: Construct Index Number of prices of items in the year 2018 from the following data by:

1). Laspeyres method; 2). Paasche's method; 3) Fisher's method

Items	Price (2011)	Quantity (2011)	Price (2018)	Quantity (2018)
A	10	10	5	25
B	35	4	35	10
C	30	3	15	15
D	10	25	20	20
E	40	3	40	5

Solution: Table 17.4 (Computation of Index Numbers)

Items	P_0	q_0	P_1	q_1	P_0q_0	P_0q_1	P_1q_0	P_1q_1
A	10	10	5	25	100	250	50	125
B	35	4	35	10	140	350	140	350
C	30	3	15	15	90	450	45	225
D	10	25	20	20	250	200	100	80
E	40	3	40	5	120	200	120	200
					Σ=700	Σ=1450	Σ=455	Σ=980

1) Laspeyres method : $(P_{01}{}^{La}) = \frac{\sum P_1q_0}{\sum P_0q_0} \times 100$ = (455/700) × 100 = 65

2) Paasche's method: $(P_{01}{}^{Pa}) = \frac{\sum P_1q_1}{\sum P_0q_1} \times 100$ = (980/1450) × 100 = 67.58

3) Fisher's method: $(P_{01}{}^{F}) = \sqrt{\left(\frac{\sum P_1q_0}{\sum P_0q_0}\right)\left(\frac{\sum P_1q_1}{\sum P_0q_1}\right)} \times 100$

$$= \sqrt{0.43927} \times 100 = 66.27$$

2) Weighted Average of Relatives Index: In this method, the construction of the index number is similar to the simple average of relatives method, in respect of computation of price relatives, as discussed in Section 17.6.1. However, to overcome the limitation of simple average of relatives method, the weights used are the values of consumption for each commodity either in the base period, or in the current period.

This method is also called as Family Budget method, which is considered as one of the methods to construct consumer price index. It can be defined symbolically as:

$(\boldsymbol{P_{01}}) = \frac{\sum\left[\left(\frac{P_1}{P_0}\times 100\right) P_0q_0\right]}{\sum P_0q_0}$, in simple $\frac{\sum \text{pv}}{\sum v}$

As an illustration let us consider the data given in Table 17.5 which also contains required computations for constructing index number through weighted average of relatives method.

Illustration-5: Table 17.5: Computation of Index Number through Weighted Average of Relatives Method

Items	Year 2005 (base Year)		Year 2015 (Current Year)		V P_0q_0	P $\left(\frac{P_1}{P_0} \times 100\right)$ Price Relatives	PV
	Prices P_0	Qty. q_0	Prices P_1	Qty. q_1			
A	7	25	12	21	175	171.43	30000.25
B	2	12	2.5	12	24	125.00	3000.00
C	3	4	5	3	12	166.67	2000.04
					∑V= 211	∑PV = 35000.29	

Then, the price index $(P_{01}) = \frac{\sum PV}{\sum V} = \frac{35000.29}{211} = 165.88$

This means that according to this method, the rise in prices in 2015 as compared to the base year 2005 is 65.88 per cent. In this method, the index of quantity relatives is expressed as:

$$(Q_{01}) = \frac{\sum\left[\left(\frac{q_1}{q_0} \times 100\right) q_0 P_0\right]}{\sum q_0 P_0} = \frac{\sum qV}{\sum V}$$

which you may compute and interpret on your own by using the data in Table 17.5.

Check Your Progress C

Compute price index number by Weighted Aggregates method (Laspeyre's, Paache's and Fisher's) and weighted Average of Relatives method, from the following data (Price quoted in Rs. per kg. and production in qtls).

Items	1990		2000	
	Price	**Production**	**Price**	**Production**
Wheat	8	700	12	900
Rice	7	900	16	1,400
Sugar	12	300	16	500

17.7 TESTS FOR INDEX NUMBERS

A perfect index number, which measures the change in the level of a phenomenon from a specific period to another period, should satisfy certain tests. In this section, we discuss the two types of tests of index numbers. They are: **(i) Time reversal test, and (2) factor reversal test**.

17.7.1 The Time Reversal Test

If we observe the construction of index numbers, we found that there are two aspects. They are period and/ or quantity. Therefore, if we reverse the time subscripts, such as base period (0) and current period (1), of a price or/and quantity index, the result should be the reciprocal of the original index number.

Algebraically, it is expressed as: $P_{0.1} \times P_{1.0} = 1$

Where, $P_{0.1}$= Index number for current period (P_1) with the base period (P_0)

$P_{1.0}$= Index number for base period (P_0) with the current period (P_1)

As we discussed the three method of construction of indices under Weighted Aggregative Index in Section 17.6.2, Fisher's Ideal Index Satisfies this test. Hence this method is considered as ideal index.

Now, let us discuss this as below:

Fisher's Ideal Index $P_{0.1} = \sqrt{\left(\frac{\sum P_1 q_0}{\sum P_0 q_0}\right)\left(\frac{\sum P_1 q_1}{\sum P_0 q_1}\right)}$

and if time subscripts are reversed i.e.,

$$P_{1.0} = \sqrt{\left(\frac{\sum P_0 q_1}{\sum P_1 q_1}\right)\left(\frac{\sum P_0 q_0}{\sum P_1 q_0}\right)}$$

With the above, now, we verify the result of time reversal test i.e.

$$P_{0.1} \times P_{1.0} = 1$$

Hence,

$$P_{0.1} \times P_{1.0} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}}$$

17.7.2 The Factor Reversal Test

Irving Fisher suggested one more test i.e. Factor Reversal Test to be applied to weighted index numbers to verify the validity. According to him "Just as our formula should permit the interchange of the two times without giving inconsistent results so it ought to permit interchanging the prices (P) and quantities (q) without giving inconsistent results, i.e., the two results multiplied together should give the true ratio".

Thus, with the usual notations a 'value index' ($P_{0.1} \times q_{1.0}$) formula is given by:

$$P_{0.1} \times q_{0.1} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

Where, $P_{0.1}$ = The price change for the current period over the base period.

$q_{0.1}$ = Quantity change for the current period over the base period.

$\sum P_1 q_1$ = The total value in the current period.

$\sum P_0 q_0$ = The total value in the base period.

The Fisher's ideal index only satisfied this test, as shown below:

$$P_{0.1}^{F} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

and, if factors (part q) are reversed i.e.

$$q_{0.1} = \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_1}}$$

Hence,

$$P_{0.1} \times q_{0.1} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum q_1 P_0}{\sum q_0 P_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_1}}$$

$$= \sqrt{\frac{\sum P_1 q_1}{\sum P_0 q_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_0}} = \frac{\sum P_1 q_1}{\sum P_0 q_0} = P_{0.1} \times q_{0.1}$$

Illustration-6: We show the following data that Fisher's ideal index satisfies the Time Reversal Test and Factor Reversal Test:

Commodity	**Price**		**No. of Units**		$P_0 q_0$	$P_1 q_0$	$P_0 q_1$	$P_1 q_1$
	2005 (P_0)	2018 (P_1)	2005(q_0)	2018 (q_1)				
I	6	10	50	56	300	500	336	560
II	2	2	100	100	200	200	240	240
III	4	6	60	60	240	360	240	360
IV	10	12	30	30	300	360	240	288
V	8	12	40	40	320	480	288	432
Total					1360	1900	1344	1880

i) Time Reverssal Test:

$$P_{0.1}^{F} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}}$$

$$P_{1.0} = \sqrt{\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}} = \sqrt{\frac{1344}{1880} \times \frac{1360}{1900}} = 1$$

$$P_{0.1} \times P_{1.0} = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344} \times \frac{1344}{1880} \times \frac{1360}{1900}} = 1$$

ii) Factor Reversal Test:

$$\text{Price ratio: } P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}}$$

$$Quantity\ ratio: q_{01} = \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_1}} \times \sqrt{\frac{1344}{1360} \times \frac{1880}{1900}}$$

$$P_{01} \times q_{01} ratio \quad \sqrt{\frac{1900}{1360} \times \frac{1880}{1344} \times \frac{1344}{1360} \times \frac{1880}{1900}} = \frac{1880}{1360}$$

$$New\ Value\ ratio\ P_{0.1} \times q_{0.1} = \frac{\sum P_1 q_1}{\sum P_0 q_0}\ is\ equal\ to\ \frac{1880}{1360}$$

17.8 CONSUMER PRICE INDEX NUMBER (CPI)

This method is also known as Cost of Living Index Number (CPI). This index is to serve as a measure of change in the prices of goods and services commonly consumed by a homogeneous section of people, such as the classes – lower middle, middle, upper middle, industrial workers, urban and rural areas etc. These indices are helpful in deciding dearness allowances, wages/ salaries, negotiations, framing price policy, taxation policy, other economic and welfare policies.

The common method for selecting from the consumption basket is to conduct a family living style survey among the population group (section) for which the consumer price index is to be constructed. Prices of selected commonly consumed items are also collected from various retail markets used by such consumers and also the quantity of consumption [normally expressed in terms of weights (w)]. When the price of one commodity varies, a simple average is applied. For example, if index number is constructed for each of five groups using weighted average of the price group, the weights used are proportional to the expenditure on the consumed items by an average family. The overall index (CPI) is computed as an weighted average of group indices and the weights being again the proportional expenditure on different groups (e.g. 30 per cent on food).

As stated in the explanation at Laspeyre's Method and accordingly using the formula of Laspeyre's

$$\text{CPI: } I = \frac{\sum W\left(\frac{P_1}{P_0} \times 100\right)}{\sum W}$$

Where, w = $\frac{P_0 q_0}{\sum P_0 q_0}$, is the weight of a group index.

Illustration-7: Let us observe how to construct the Consumer Price Index for food with the help of the following data pertains to current price, base price and weights of seven items:

Construction of an Index for food

Items	Price		$P\left(\frac{P_1}{P_0} \times 100\right)$	Weights (w)	Pw
	P_1	P_0			
Wheat	50	40	125.0	30	3750.0
Pulses	45	30	150.0	20	3000.0
Rice	60	40	150.0	10	1500.0
Sugar	40	50	200.0	5	1000.0
Oil	75	60	125.0	15	1875.0
Potato	60	50	120.0	15	1800.0
Meat	200	150	133.3	5	666.5
Total				100	13591.5

CPI (Food) = $\frac{\sum W\left(\frac{P_1}{P_0} \times 100\right)}{\sum W} = \frac{13591.5}{100} = 135.92$

Check Your Progress D

Construct Consumer Price Index number from the data given belowL

Item	:	A	B	C	D	E
Price of Base Year (Rs.)	:	85	15	45	55	17
Price of Current Year (Rs.)	:	115	20	61	100	23
Weights	:	35	15	10	25	15

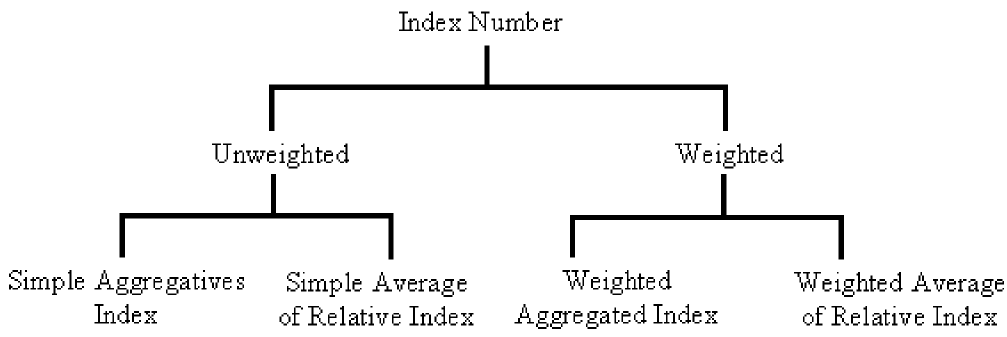
17.9 LET US SUM UP

An index number is a specialised average which helps in comparison of the level of magnitude of a group of related variables with respect to time, geographical location or other characteristics such as production, income, employment, etc. It combines two or more time series variables related to non-comparable units.

Index numbers can be used in several ways, such as study trends and tendencies of business activities, provide guidelines in framing suitable policies, measure real purchasing power of money, help in transforming nominal wage into real wage and so on. The researcher may face various problems in the construction of different types of indices. They may be selection of the base period, collection of data, selection of commodities, choice of averages and weights, selection of an appropriate index. These issues must be clarified before constructing indices.

There are three principal types of indices (i) price indices, (ii) quality indices, and (iii) value indices. Among these three, price indices is the most common in analysing the data.

There are different methods of constructing index numbers which is illustrated through the following chart:



Choice of an appropriate method depends upon the purpose of constructing indices.

You have been shown how to use the Laspeyre's, Paache's and Fisher's formulae for calculation of price as well as quantity indices. Only Fisher's Ideal Index number satisfies the Time Reversal Test and Factor Reversal Test. We have also discussed how to measure change in consumer price or cost of living.

17.10 KEY WORDS

Base period: It is the reference period against which comparisons are made.

Cost of Living Index: Numbers represent the average change in the prices paid by the consumer on specified goods and services over a period of time, popularly known as "Consumer Price Index Number".

Index Number: A ratio for measuring differences in the magnitude of a group of related variables over time.

Price Index: A measure of how much the price variables change over a period of time.

Price Relative: In the construction of an index number, price relative for a commodity in the ratio of the current year price to base year price of that commodity.

Quality Index: A measure which studies the quantity of a variable changes from one period to another period.

Value Index: A measure for changes in total monetary worth over a time.

17.11 ANSWERS TO CHECK YOUR PROGRESS

A 1) a) Agree b) Disagree c) Disagree d) Agree

e) Agree f) Disagree g) Disagree.

B a) (i) Simple aggregative $P_{01} = 164.3$

(ii) Average of Relatives $P_{01} = 173.8$

b) (i) Simple Aggregative Index for the year 2017 over the year 2018 = 123

(ii) Simple Aggrivative Index for the year 2018 over the year 2016 = 273

C Weighted Aggregates index number:

${P_{01}}^{La} = 183.9$; ${P_{01}}^{Pa} = 180.4$; ${P_{01}}^{F} = 181.9$;

Weighted Average of Relatives Index $(P_{01}) = 183.9$

D CPI = 146.65

17.12 TERMINAL QUESTIONS/EXERCISES

1) What do you mean by an index number? Explain the uses of index numbers for analysing the data.

2) Discuss various issues that arise in connection with the construction of an index number.

3) Briefly explain different methods for construction of indices and their limitations.

4) Why do we consider Fisher's index as an ideal index?

5) Write short notes on:

   a) Price Index

   b) Quantity Index

   c) Value Indices

   d) Consumer Price Index Numbers

6) A drug processing plant utilized four different materials in the manufacturing of a medicine. The following data indicates the final inventory levels (in tons) and prices (per kg) for these materials for the years 2010 and 2015.

Items	2010		2015	
	Inventory	Price (Rs.)	Inventory	Price (Rs.)
A	96	45	108	41
B	495	26	523	32
C	1,425	5	1,608	8
D	208	12	196	9

Find the price indices and quantity indices by using the methods of unweighted index numbers and comment on the results.

7) A department of Statistics has collected the following data describing the prices and quantities of harvested crops for the years 1990, 2000 and 2004 (Price in Qtls. and Production in tons).

Items	1990		2000		2004	
	Price	Production	Prince	Production	Prince	Production
Paddy	200	1.050	500	1,300	600	1,450
Wheat	250	940	550	1,220	700	1,450
Groundnut	350	400	800	500	1,000	480

Construct the price and quantity indices of Laspeyre's Index, Paache's Index and Fisher's Index in 2000 and 2004, using 1990 as the base period and verify whose index number satisfies the Time Reversal Test and Factor Reversal Test. Give your comments on the results.

8) From the given data in Problem No. 7, find out the following:

i) Weighted average of Relative Prices Index number for 2004 using 1990 and 2000 as base.

ii) Weighted average of Relative Quantity Index for 2004 using 2000 as the base.

iii) Give your comments on the price indices.

Given below is the annual income of an Engineer and the general index number of prices during 2010-2017. Construct the index number to show the change in the real income of the Engineer.

Year	2010	2011	2012	2013	2014	2015	2016	2017
Income: (in 000' Rs.)	255	265	286	312	336	380	405	420
Price Index No.	100	108	116	153	140	192	248	235

9) A survey of the budget of working class families in an industrial area gave the following information.

Expression %	:	Food 30%	Rent 15%	Clothing 20%	Fuel 10%	Others 25%
Price in 2015 (Rs.)	:	100	20	70	20	40
Price in 2016 (Rs.)	:	90	20	60	15	55

What is the change in the cost of living in 2016, as compared with 2015?

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.

17.13 FURTHER READINGS

A number of good text books are available for the topics dealt with in this unit.

The following books may be used for more indepth study.

Hooda, R.P, 2001. Statistics for Business and Economics, Macmillan India Ltd.

Richard I. Levin and David S. Rubin, 1996, Statistics for Management, Prentice Hall of India Pvt. Ltd.

Gupta, S.P., Statistical Methods, 2000, Sultan Chand and Sons.

Gupta, C.B. and Vijay Gupta, 2001. An Introduction to Statistical Methods, Vikas Publishing House Pvt. Ltd., New Delhi.

UNIT 18 TIME SERIES ANALYSIS

Structure

18.0 OBJECTIVES

After studying this unit, you should be able to:

- define the concept of time series,
- appreciate the role of time series in short-term forecasting,
- explain the components of time series, and
- estimate the trend values by different methods

18.1 INTRODUCTION

In the previous unit, you have learnt types of the index numbers and various methods in constructing index numbers. The nature of data varied from case to case. You have come across quantitative data for a group of respondents collected with a view to understanding one or more parameters of that group, such as investment, profit, consumption, weight etc. But when a nation, state, an institution or a business unit etc., intend to study the behavior of some element, such as price of a product, exports of a product, investment, sales, profit etc., as they have behaved over a period of time, the information shall have to be collected from a fairly long period, usually at equal time intervals. Thus, a set of any quantitative data collected and arrangement on the basis of time is called 'Time Series'. The unit of time may be a decade, a year, a month, or a week etc.

Usually, the quantitative data of the variable under study are denoted by y_1, y_2,….y_n and the corresponding time units are denoted by t_1, t_2,….t_n. The variable 'y' shall have variations, as you will see ups and downs in the values. These changes account for the behavior of that variable.

Instantly it comes to our mind that 'time' is responsible for these changes, but this is not true. Because, the time (t) is not the cause and changes in the variable (y) are not the effect. The only fact, therefore, which we must understand is that there are a number of causes which affect the variable and have operated on it during a given time period. Hence, time becomes only the forecasting any event helps in the process of decision making. Forecasting is possible if we are able to understand the past behavior of that particular activity. For understanding the past behavior, a researcher needs not only the past data but also a detailed analysis of the same. Thus, in this unit we will discuss the need for analysis of time series, fluctuations of time series which account for changes in the series over a period of time, and measurement of trend for forecasting.

18.2 DEFINITION AND UTILITY OF TIME SERIES ANALYSIS

Based on the above discussion we can understand the definition given by a few statisticians. They are:

"A time series consists of statistical data which are collected, recorded over successive increments".

"When quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series".

The analysis of time series is of great utility not only to research workers but also to economists, businessmen and scientists etc., for the following reasons:

1) It helps in understanding past behavior of the variables under study.

2) It facilitates in forecasting the future behavior with the help of the changes that have taken place in the past.

3) It helps in planning future course of action.

4) It helps in knowing current accomplishment.

5) It is helpful to make comparisons between different time series and significant conclusions draw therefrom.

Thus, we can say that the need for time series analysis arises because:

- we want to understand the behavior of the variables under study,
- we want to know the expected quantitative changes in the variable under study, and
- we want to estimate the effect of various causes in quantitative terms

In a nutshell, the time series analysis is not only useful for researchers, business research institutions, but also for Governments for devising appropriate future growth strategies.

18.3 COMPONENTS OF TIME SERIES

If you are informed that the price of one kilogram sunflower oil was Rs.0.50 in the year 1940 and in the year 1980 it was Rs. 30 and in the year 2004 it is reported to be Rs. 70, and if you are asked this question: shall sunflower oil be sold again in the future for either Rs.0.50 and Rs. 30 per kg? Surely, you answer would be 'No'.

Another Question: Shall sunflower oil be sold again in future for Rs.60 per kg? No doubt, you answer would be 'Yes'. Have you ever thought about how you answered the above two questions? Probably you have not! The analysis of these answers shall lead us to arrive at the following observations:

- There are several causes which affect the variable gradually and permanently. Therefore we are prompted to answer 'No" for the first question.
- There are several causes which affect the variable for the time being only. For this reason we are prompted to answer 'Yes' for the second question.

The causes which affect the variable gradually and permanently are terms as "Long-Term Causes". The examples of such causes are: increase in the rate of capital formation, technological innovations, the introduction of automation, changes in productivity, improved marketing etc. The effect of long term causes is reflected in the tendency of a behavior, to move in an upward or downward direction, termed as 'Trend' or 'Secular Trend'. It reveals as to how the time series has behaved over the period under study.

The causes which affect the variables for the time being only are labelled as 'Short-Term Causes". The short term causes are further divided into two parts, they are 'Regular' and 'Irregular'. Regular causes are further divided into two parts, namely 'cyclical causes' and seasonal causes'. The cyclical variations are also termed as business cycle fluctuations, as they influence the variable. A business cycle is composed of prosperity, recession, depression and recovery. The periodic movements from prosperity of recovery and back again to prosperity vary both in time and intensity. The seasonal causes, like weather conditions, business climate and even local customs and ceremonies together play an important role in giving rise to seasonal movements to almost all the business activities. For instance the yearly weather conditions directly affect agricultural production and marketing.

It is worthwhile to say that the seasonal variations analysis will be possible only if the season-wise data are available. This fact must be checked first. For analysing the seasonal effect various methods are available. Among them seasonal index by 'Ratio to Moving Average Method' is the most widely used. However, if collected data provides only yearly values, there is no possibility of obtaining seasonal variations. Therefore, the residual amount after eliminating trend will be the effect or irregular or random causes.

Irregular causes are also termed as 'Erratic' or 'Random' causes. Random variations are caused by infrequent occurrences such as wars, strikes, earthquakes, floods etc. These reasons either go very deep downwards or very high upwards.

The foregoing paragraphs have, in a way, led us to enumerate the components of the time series. The components from the basis for 'Time Series Analysis'.

Long-term causes	:	Secular Trend or Trend (T)
Short-term causes	:	
Regular	:	Cyclical (C)
	:	Seasonal (S)
Irregular or Random	:	Erratic (I)

18.4 DECOMPOSITION OF TIME SERIES

Decomposition and analysis of a time series are one and the same thing. The original data or observed data 'O' is the result of the effects generated by the long-term and short-term causes, namely (1) Trend = T, (2) cyclical – C, (3) Seasonal = S, and (4) Irregular = I. Finding out the values for each of the components is called decomposition of a time series. Decomposition is done either by the additive model or the multiplicative model of analysis. Which of these two models is to be used in analysis of time series depends on the assumption that we might make about the nature and relationship among the four components.

Additive Model: It is based on the assumption that the four components are independent on one another. Under this assumption, the pattern of occurrence and the magnitude of movements in any particular component are not affect4ed by the other components. In this model the values of the four components are expressed in the original units of measurement. Thus, the original data or observed data 'Y' is the total of the four component values, that is,

$$Y = T + S + C + I$$

where, T, S, C and I represents the trend variations, seasonal variations cyclical variations, and erratic variations, respectively.

Multiplicative Model: It is based on the assumption that the causes giving rise to the four components are interdependent. Thus, the original data or observed data 'Y' is the product of four component values, that is:

$$Y = T \times S \times C \times I$$

In this model the values of all the components, except trend values, are expressed as percentages.

In business research, normally the multiplicative model is more suited and used more frequently for the purposes of analysis to time series. Because, the data related to business and economic time series is the result of interaction of a number of factors which individually cannot be held responsible for generating any specific type of variations.

Let us consider an example for construction of time series according to the Multiplicative Model. Table 18.1 represents trend, seasonal and cyclical-erratic components of a hypothetical series.

Table 18.1: Hypothetical time series and its components (quarterly)

			Components		
Year	Quarter	Series (O)	Trend (T)	Seasonal (100 S)	Cyclical – Irregular (100 CI)
1	1	79	80	120	82
	2	58	85	80	85
	3	84	90	92	102
	4	107	95	108	105
2	1	130	100	120	108
	2	93	105	80	132
	3	121	110	92	120
	4	161	115	108	130
3	1	216	120	120	150
	2	132	125	80	132
	3	150	130	93	125
	4	163	135	108	112
4	1	176	140	120	105
	2	112	145	80	97
	3	128	150	93	93
	4	142	155	108	85

According to multiplicative model

Y = T × S × C × I

Thus, 79 (1 year and 1 quarter) = $80 \times \frac{120}{100} \times \frac{82}{100}$

130 (2 year and 1 quarter) = $100 \times \frac{120}{100} \times \frac{108}{100}$

Thus, each quarterly figure (Y) is the product of the T, S and CI. Such as synthetic composition looks like an actual time series and has encouraged use of the model as the basis of the analysis of time series data.

18.5 PRELIMINARY ADJUSTMENTS

Before we proceed with the task of analysing a time series data, it is necessary to do relevant adjustments in the raw data. They are:

1) **Calendar Variations**: As we are aware, all the calendar months do not have the same number of days. For instance, all production in the month of February may be less than other months because of fewer days and if we take the holidays into account the variation is greater. Therefore, adjustments for calendar variations have to be made.

2) **Price Changes**: As price level changes are inevitable, it is necessary to convert monetary values into real values after taking into consideration the price indices. In fact this is process of deflating which will be discussed in Unit 17 (Index Numbers) of this course.

3) **Population changes**: Population grown constantly. This also calls for adjustment in the data for the population changes. In such cases, if necessary, per capita values may be computed (dividing original figures by the total population).

Check Your Progress A

1) Do you agree or disagree on the following statement. Give reasons of your opinion.

   a) Time is cause for the ups and downs in the values of the variable under study.

   b) The variable under study in time series analysis is denoted by 'y'.

   c) 'Trend' values are major component of the time series.

   d) Analysis of time series helps in knowing current accomplishment

   e) Weather conditions, customs, habits etc., are causes for cuyclical variations.

   f) The analysis of time series is done to know the expected quantity changes in the variable under study.

2) Why do we analyse a time series?

3) List out the components of a time series.

18.6 METHODS OF MEASUREMENT OF TREND

The effect of long-term causes is seen in the trend values we compute. A trend is also known as 'secular trend' or 'long-term trend' as well. There are several methods of isolating the trend of which we shall discuss only two methods which are most frequently use in the business and economic time series data analysis. They are: Moving Average Method, and Method of Least Square.

18.6.1 Moving Average Method

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 cycles i.e. periodic movements. Even if the periodic move with the period of the cycle i.e., periodic movements. Even if the periodic move merit is absent in the time series, the irregularities of data can be reduced to a large extent by moving average process. If we choose this method, we should select a period for calculation. The period may be 3 years or 5 years or 6 years or 12 years etc., which is to be decided by considering the duration of the cycle.

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 18.2.

Table 18.2: 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). As said above, 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 point of relevant time interval. Study Table 18.3 carefully to understand the procedure.

Table 18.3: Computation of Odd Period Moving Average

Years	Sales (*000 tonnes)	3 Yearly Moving Totals	5 Yearly Moving Totals
2001	15	--	--
2002	25	72	24
2003	32	81	27

2004	24	75	25
2005	19	60	20
2006	17	--	--

You should note that the moving average for the first three years (2001, 2002 and 2003) i.e., 72 is associated with the middle year 2002. Having dropped the first year, the moving average of the next three years i.e. 2002, 2003 and 2004 is placed against 2003; 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-1** carefully and understand the procedure involved in centering.

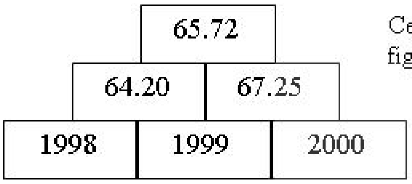
Illustration-1: Compute 4 yearly moving averages for the following data:

Years	:	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Sales (Rs. In '000)	:	75	60	54	69	86	65	63	80	90	72

Solution: Computation of 4 yearly moving average.

Year	**Sales (Rs. In 000s)**	**4 Yearly Moving Total**	**4 Yearly Moving Average**	**4 Yearly Moving Average Centered**
1997	75	--	--	--
1998	60	--	--	--
1999	54	258	64.20	--
2000	69	269	67.25	67.88
2001	86	274	68.50	69.62
2002	65	283	70.75	72.12
2003	63	294	73.50	73.50
2004	80	298	74.50	75.37
2005	90	305	76.25	--
2006	72	--	--	--

The total 258 of the first four figures (years 1997 to 2000) and their average 64.20 is written against the middle of this time period i.e., middle of the years 1998 and 1999. This middle time period is a specially designed year taking last six months from 1998 and the first six months from 1999. Similarly, the total 269 corresponding to year 1998 to 2001 and their average 67.25 is written against the specially designed year i.e., the mid-year of 1999 and 2000. This process continues till the last average 76.25 and the total 305 is noted against the mid-year of 2004 and 2005. To find out the first centred moving average 65.72 (i.e., a figure of moving average which will coincide with the year 1999), we have to find the mid-value 64.20 and 67.25, the first two figures in Column 4. This can be easily seen with the help of diagram given below:



Centered Moving Average figures in Column 4

The diagram shows that the figure which coincides with the year 1999 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 centered moving average and is entered in the last column. The various entered moving averages are, thus, calculated by taking successively mean of the two consecutive figures from Column 4.

18.6.2 Least Square Method

This is also known as straight line method. This method is most commonly used in research to estimate the trend of time series data, as it is mathematically designed to satisfy two conditions. They are:

1) Sum of $(Y + Y_c)$ = 0, and

2) Sum of $(Y + Y_c)^2$= least

The straight line method gives a line of best fit on the given data. The straight line which can satisfy the above conditions and make use of regression equation, is given by:

Y_c = a + bx

Where, 'Y_c represents the trend value of the time series variable y, 'a' and 'b' are constant values of which 'a' is the trend value at the point of origin and 'b' is the amount by which the trend value changes per unit of time, and 'x' is the unit of time (value of the independent variable).

The values of constant, 'a' and 'b' are determined by the following two normal equations.

$\sum y = na + b\sum x \ldots\ldots\ldots..(i)$

$\sum xy = a\sum x + b\sum x^2 \ldots..(ii)$

The process of finding values of constants a and b can be made simple by using a shortcut method, that is, by taking the origin year in such a way that it gives the toal of 'x' (∑x) equal to 'zero'. This becomes possible if we take the median year as origin period. Thus, the negative values in the first half of the series balance out the positive values in the second half. Thus, the earlier normal equation shall be changed as follows, with reference to $\sum x = 0$.

$\sum y = a$ (as $\sum bx$ becomes zero)

$\sum xy = b\sum x^2$ (as $a\sum x$ becomes zero)

Therefore, the values of two constants are obtained by the following formulae:

$a = \frac{\sum y}{N}$, and $b = \frac{\sum xy}{\sum x^2}$

It is to be noted that when the number of time units involved is even, the point of origin will have to be chosen between the two middle time units.

Let us consider an illustration to understand the procedure for estimation of the trend by using the method of least squares.

Illustration 2: The decision making body of a fertilizer firm producing fertilizer wants to predict future sales trend for the year 2006 and 2008 based on the analyses of its past sales pattern. The sales of the firm for the last 7 years, for this purpose are given below:

Year	1998	1999	2000	2001	2002	2003	2004
Sales (in ’000 tonnes)	70	75	90	98	85	91	100

Solution: to find the straight line equation (Y_c = a + bx) for the given time series data, we have to substitute the values of already arrived expression, that is:

$a = \frac{\sum y}{N}$, and $b = \frac{\sum xy}{\sum x^2}$

In order to make the total of x = ‘zero’, we must take median year (i.e. 2001) as origin. Study the following table carefully to understand the procedure for fitting the straight line.

Table 18.4: Computation of Trend

Year	Sales (in ’000 tonnes)	x	x^2	Xy	Trend (Y_c = a + bx)
1998	70	–3	9	–210	74.5
1999	75	–2	4	–150	78.6
2000	90	–1	1	–90	82.8
2001	98	0	0	0	87.2
2002	85	1	1	85	91.2
2003	91	2	4	182	95.4
2004	100	3	9	300	99.5
N = 7	$\sum y = 609$	$\sum x = 0$	$\sum x^2 = 28$	$\sum xy$ - 117	609.0

$a = \frac{\sum y}{N} = \frac{609}{7} = 87$, and $b = \frac{\sum xy}{\sum x^2} = \frac{117}{28} = 4.18$

Thus, the straight line trend equation is: Y_c= 87 + 4.18x

From the above equation, we can also find the monthly increase in sales as follows:

$\frac{4.180}{12} = 348.33$tons

The reason for this is that the trend values increased by a constant amount 'b' every year. Hence the annual increase in sales in 4:18 thousand tons.

Trend values are to be obtained as follow:

Y_{1998}= 87+4.18 (–3) = 74.5

Y_{1999}= 87+4.18(–2) = 78.6 and so on …..

Predicting with decomposed components of the time series: The management wants to estimate fertilizer sales for the years 2006 and 2008.

Estimation of sales for 2006, 'x' would be 5 (because for 2004 'x' was 3)

Y_{2006}= 87 + 4.18 (5) = 1.7.9 thousand tonnes

Estimation of sales for 2008, 'x' would be 7.

Y_{2008}= 87 + 4.18 (7) = 116.3 thousand tonnes

Illustration-3: Fit a straight line trend by the method of least square from the following data and find the trend values.

Year	1958	1959	1960	1961	1962
Sales (in lakhs of units)	65	95	80	115	105

Solution: We have n = 5

$\therefore n \text{ is odd}$

Taking middle year i.e. 1960 as the origin, we gets

Table 18.5: Computation

Year	Sales	X	X^2	XY
1958	65	–2	4	–130
1959	95	–1	1	–95
1960	80	0	0	0
1961	115	1	1	115
1962	105	2	4	210
Total	ΣY=460	ΣX=0	ΣX^2=10	ΣXY=100

$\therefore n = 5, \Sigma X = 0, \Sigma X^2 = 10, \Sigma Y = 460 \text{ and } \Sigma XY = 100$

$a = \Sigma Yn\ 460 \Rightarrow a = 460/5 = 92$

$b = \Sigma XY \Sigma X^2 = \dfrac{100}{10} = 10$

$\therefore \textit{the equation of straight line trend is } Yc = a + bx \Rightarrow yc = 92 + 10X$

$\textit{for the year } 1958, x = \ -2$

$\Rightarrow$ Yc $(1958) = 92 + 10(-2) = 92 - 20 = 72$

$\textit{for the year } 1959, X = \ -1$

$\Rightarrow$ Yc $(1959) = 92 + 10(-1) = 92 - 10 = 82$

$\textit{for the year } 1960, X = \ 0$

$\Rightarrow$ Yc $(1960) = 92 + 10(0) = 92 - 0 = 92$

$\textit{for the year } 1961, X = \ 1$

$\Rightarrow$ Yc $(1961) = 92 + 10(1) = 92 + 10 = 102$

$\textit{for the year } 1962, X = \ 2$

$\Rightarrow$ Yc $(1962) = 92 + 10(2) = 92 + \ 20 = 112$

We have,

Year	Trend Value	And the straight line trend is Yc = 92 + 10X
1958	72	
1959	82	
1960	92	
1961	102	
1962	112	

Check Your Progress B

1) Do you agree or disagree the following statements. Justify your opinion.

   a) The multiplicative model is based on the assumption that the causes giving rise to the four components are dependent.

   b) The total of the difference between original data and trend values (obtained by straight line method) will never be zero.

   c) In the least square trend equation Y_c = a + bx, if b is positive it indicates a rising trend.

   d) The additive model of time series analysis is expressed as: Y = T + S + C + I.

2) Enumerate the methods of isolating trend.

3) What do you mean by moving average? Explain the procedure for calculation of moving average when the data is given in odd and even periods.

4) Foodgrain production (in lakh tones) is given below (figures are imaginary). Find the Trend by using a) 3 yearly and 4 yearly moving average method b) Straight Line Method. Tabulate the trend values. C) Predict the production for the year 2022.

Years	Production
2008	40
2009	60
2010	45
2011	83
2012	130
2013	135
2014	150
2015	120
2016	200

18.7 LET US SUM UP

This unit has introduced you to the concept of time series and its analysis with a view to making more accurate and reliable forecasts for the future.

A set of quantitative data arranged on the basis of TIME are referred to as 'Time Series'. The analysis of time series is done to understand the dynamic conditions for achieving the short-term and long-term goals of institution(s). With the help of the techniques of time series analysis the future pattern can be predicted on the basis of past trends.

The quantitative values of the variable under study are denoted by denoted by y_1, y_2, y_3….. and the corresponding time units are denoted by x_1, x_2, x_3…… . The variable 'y' shall have variations, you will see ups and downs in the values. There are a number of causes during the given time period which affect the variable. Therefore, time becomes the basis of analysis. Time is not the cause and the changes in the values of the variable are not effect.

The causes which affect the variable gradually and permanently are termed as Long-term causes. The causes which affect the variable only for the time being are termed as Short-term causes. The time series are usually the result of the effects of one or more of the four components. These are trend variations (T) seasonal variations (S), cyclical variations (C) and irregular variations(I)

When we try to analyse the time series, we try to isolate and measure the effects fo various kinds of these components one a series.

We have two models for analysing time series:

1) Addictive model, which considers the sum of various components resulting in the given values of overall time series data and symbolically it would be expressed as Y = T + C + S + I.

2) The multiplicative model assumes that the various components interact in a multiplicative manner to produce the given values of the overall time series data dn symbolically it would be expressed as : y = T × C × S × I.

The trend analysis brings out the effect of long-term causes. There are different methods of isolating trends, among these we have discussed only two methods i.e., Moving Average Method and Least Square Method.

Long-term predictions can be made on the basis of trends, and only the least square method of trend computation offers this possibility.

18.8 KEY WORDS

Cyclical Variations: A type of variation in time series, in which the values of variables vary up and down around the secular trend line.

Irregular Variations: A type of element of a time series, refers to such variations in business activity which do not repeat according to a definite pattern and the values of variables are completely unpredictable.

Seasonal Variation: Pattern of change in a time series within a year and the same changes tend to be repeated from year to year.

Secular Trend: A type of variation in a time series, the long-term tendency of a time series to grow or decline over a period of time.

Time Series: is a data on any variable accumulated at regular time intervals.

18.9 ANSWERS TO SELF ASSESSMENT EXERISES

A) 1) a) Disagree b) Agree c) Agree d) Agree

e) Disagree f) Agree

3) Secular trend, Seasonal variation, Cyclical variation, and Irregular Variation

B) 1) a) Disagree b) Agree c) Disagree d) Agree

4) a) 3 years M.A. = 48.33, 62.67, 86, 116, 138.33, 101.67, 156.67

4 years M.A. = 273, 353, 433, 51135, 540, 570

b) $Y_1 = 107 + 18.03$ x

c) Estimated production for 2022 is 287.3 lakh tonnes.

18.10 TERMINAL QUESTIONS

1) What is time series? Why do we analyse a time series?

2) Explain briefly the components of time series.

3) Explain briefly the additive and multiplicative models of time series. Which of these model is more commonly used and why.

4) From the following data, compute trend values, using 3 yearly and 4 yearly moving average.

Years	2010	2011	2012	2013	2014	2015	2016	2017
Yields (in tones)	24	28	38	33	49	50	66	68

5) The production (in thousand tons) in a sugar factory during 2010 to 2017 has been as follows:

Years	2010	2011	2012	2013	2014	2015	2016	2017
Production	35	38	49	41	56	58	76	75

i) Find the trend values by applying the method of least square.

ii) What is the monthly increase in production?

iii) Estimate the production of sugar for the year 2020.

6) The following data relates to a survey of use car sales in a city for the period 2006-2014. Predict sales for 2022 by using the linear trend equation.

Years	2006	2007	2008	2009	2010	2011	2012	2013	2014
Sales	214	320	305	298	360	450	340	500	520

7) Calculate 4 yearly and 5 yearly moving average for the following time series data:

Quarter	2014	2015	2016	2017
I	62	68	75	80
II	58	62	68	75
III	72	74	81	85
IV	65	77	80	85

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.

18.11 FURTHER READINGS

A number of good text books are available for the topics dealt with in this unit. The following books may be used for more indepth study.

Mentgomery, D.,C. and L.A. Johnson, 1996, '*Forecasting and Time Series Analysis*' McGraw Hill: New York.

Chandan, J.S., 2001, *Statistics for Business and Economics*, Vikas Publishing House Pvt. Ltd., New Delhi

Gupta, S.P. and H.P. Gupta, 2001, *Business Statistics*, S. Chand, New Delhi.

C.B. Gupta & Vijay Gupta, Vikash Publishing Honk Pvt. Ltd., New Delhi.

LOGARITHMS

79

	0	1	2	3	4	5	6	7	8	9	Mean Differences. 1	2	3	4	5	6	7
10	00000	00432	00860	01284	01703						42	85	127	170	212	254	297
						02119	02531	02938	03342	03743	40	81	121	162	202	242	283
11	04139	04532	04922	05308	05690						37	77	116	191	193	232	270
						06070	06446	06819	07188	07555	37	74	111	148	185	222	259
12	07918	08279	08636	08991	09342						36	71	106	142	177	213	248
						09691	10037	10380	10721	11059	34	68	102	136	170	204	238
13	11394	11727	12057	12385	12710						33	66	98	131	164	197	229
						13033	13354	13672	13988	14301	32	63	95	126	158	190	221
14	14613	14922	15229	15534	15836						30	61	91	122	152	183	213
						16137	16435	16732	17026	17319	29	59	88	118	147	177	206
15	17609	17898	18184	18469	18752						28	57	85	114	142	171	199
						19033	19312	19590	19866	20140	28	55	83	110	138	165	193
16	20412	20683	20952	21219	21484						27	53	80	107	134	160	187
						21748	22011	22272	22531	22789	27	52	78	104	130	156	182
17	23045	23300	23553	23805	24055						26	50	76	101	126	151	176
						24304	24551	24797	25042	25285	25	49	73	98	122	147	174
18	25527	25768	26007	26245	26482						24	48	71	95	119	143	167
						26717	26951	27184	27476	27646	23	46	69	93	116	139	162
19	27875	28103	28330	28556	28780						23	45	68	90	113	135	158
						29003	29226	29447	29667	29885	22	44	66	88	110	132	154
20	30103	30320	30535	30750	30963	31175	31387	31597	31806	32015	21	43	64	85	106	127	148
21	32222	32428	32634	32838	33041	33244	33445	33646	33846	34044	20	41	61	81	101	121	141
22	34242	34439	34635	34830	35025	35218	35411	35603	35793	35984	20	39	58	77	97	116	135
23	36173	36361	36549	36736	36922	37107	37291	37475	37658	37840	19	37	56	74	93	111	130
24	38021	38202	38382	38561	38739	38917	39094	39270	39445	39620	18	35	53	71	89	106	124
25	39794	39967	40140	40312	40483	40654	40824	40993	41162	41330	17	34	51	68	85	102	119
26	41497	41664	41830	41996	42160	42325	42488	42651	42813	42975	16	33	49	66	82	98	115
27	43136	43297	43457	43616	43775	43933	44091	44248	44404	44560	16	32	47	63	79	95	111
28	44716	44871	45025	45179	45332	45484	45637	45788	45939	46090	15	30	46	61	76	91	107
29	46240	46389	46538	46687	46835	46982	47129	47276	47422	47567	15	29	44	59	74	88	103
30	47712	47857	48001	48144	48287	48430	48572	48714	48855	48996	14	29	43	57	72	86	100
31	49136	49276	49415	49554	49693	49831	49969	50106	50243	50379	14	28	41	55	69	83	97
32	50515	50650	50786	50920	51054	51188	51322	51455	51587	51720	13	27	40	54	67	80	94
33	51851	51983	52114	52244	52375	52504	52634	52763	52892	53020	13	26	39	52	65	78	91
34	53148	53275	53403	53529	53656	53782	53908	54033	54158	54283	13	25	38	50	63	76	88
35	54407	54531	54654	54777	54900	55023	55145	55267	55388	55509	12	24	37	49	61	73	85
36	55630	55751	55871	55991	56110	56229	56348	56467	56585	56703	12	24	36	48	60	71	83
37	56820	56937	57054	57171	57287	57403	57519	57634	57749	57864	12	23	35	46	58	70	81
38	57978	58092	58206	58320	58433	58546	58659	58771	58883	58995	11	23	34	45	57	68	79
39	59106	59218	59329	59439	59550	59660	59770	59879	59988	60097	11	22	33	44	55	66	77
40	60206	60314	60423	60531	60638	60746	60853	60959	61066	61172	11	21	32	43	54	64	75
41	61278	61384	61490	61595	61700	61805	61909	62014	62118	62221	10	21	31	42	53	63	74
42	62325	62428	62531	62634	62737	62839	62941	63043	63144	63246	10	20	31	41	51	61	71
43	63347	63448	63548	63649	63749	63849	63949	64048	64147	64246	10	20	30	40	50	60	70
44	64345	64444	64542	64640	64738	64836	64933	65031	65128	65225	10	20	29	39	49	59	68
45	65321	65418	65514	65610	65706	65801	65896	65992	66087	66181	10	19	29	38	48	57	67
46	66276	66370	66464	66558	66652	66745	66839	66932	67025	67117	9	19	28	37	47	56	65
47	67210	67302	67394	67486	67578	67669	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

LOGARITHMS 80

	0	1	2	3	4	5	6	7	8	9	Mean Diffe				
											1	2	3	4	5
50	69897	69984	70070	70157	70243	70329	70415	70501	70586	70672	9	17	26	34	43
51	70757	70842	70927	71012	71096	71181	71265	71349	71433	71517	8	17	25	34	42
52	71600	71684	71767	71850	71933	72016	72099	72181	72263	72346	8	17	25	33	42
53	72428	72509	72591	72673	72754	72835	72916	72997	73078	73159	8	16	24	32	41
54	73239	73320	73400	73480	73560	73640	73719	73799	73878	73957	8	16	24	32	40
55	74036	74115	74194	74273	74351	74429	74507	74586	74663	74741	8	16	23	31	39
56	74819	74896	74974	75051	75128	75205	75282	75358	75435	75511	8	15	23	31	39
57	75587	75664	75740	75815	75891	75967	76042	76118	76193	76268	8	15	23	30	38
58	76343	76418	76492	76567	76641	76716	76790	76864	76938	77012	7	15	22	30	37
59	77085	77159	77232	77305	77379	77452	77525	77597	77670	77743	7	15	22	29	37
60	77815	77887	77960	78032	78104	78176	78247	78319	78390	78462	7	14	22	29	36
61	78533	78604	78675	78746	78817	78888	78958	79029	79099	79169	7	14	21	28	36
62	79239	79309	79379	79449	79518	79588	79657	79727	79796	79865	7	14	21	28	35
63	79934	80003	80072	80140	80209	80277	80346	80414	80482	80550	7	14	20	27	34
64	80618	80686	80754	80821	80889	80956	81023	81090	81158	81224	7	13	20	27	34
65	81291	81358	81425	81491	81558	81624	81690	81757	81823	81889	7	13	20	26	33
66	81954	82020	82086	82151	82217	82282	82347	82413	82478	82543	7	13	20	26	33
67	82607	82672	82737	82802	82866	82930	82995	83059	83123	83187	6	13	19	26	32
68	83251	83315	83378	83442	83506	83569	83632	83696	83759	83822	6	13	19	25	32
69	83885	83948	84011	84073	84136	84198	84261	84323	84386	84448	6	12	19	25	31
70	84510	84572	84634	84696	84757	84819	84880	84942	85003	85065	6	12	19	25	31
71	85126	85187	85248	85309	85370	85431	85491	85552	85612	85673	6	12	18	24	31
72	85733	85794	85854	85914	85974	86034	86094	86153	86213	86273	6	12	18	24	30
73	86332	86392	86451	86510	86570	86629	86688	86747	86806	86864	6	12	18	24	30
74	86923	86982	87040	87099	87157	87216	87274	87332	87390	87448	6	12	17	23	29
75	87506	87564	87622	87679	87737	87795	87852	87910	87967	88024	6	12	17	23	29
76	88081	88138	88195	88252	88309	88366	88423	88480	88536	88593	6	11	17	23	29
77	88649	88705	88762	88818	88874	88930	88986	89042	89098	89154	6	11	17	22	28
78	89209	89265	89321	89376	89432	89487	89542	89597	89653	89708	6	11	17	22	28
79	89763	89818	89873	89927	89982	90037	90091	90146	90200	90255	6	11	17	22	28
80	90309	90363	90417	90472	90526	90580	90634	90687	90741	90795	5	11	16	22	27
81	90848	90902	90956	91009	91062	91116	91169	91222	91275	91328	5	11	16	21	27
82	91381	91434	91487	91540	91593	91645	91698	91751	91803	91855	5	11	16	21	27
83	91908	91960	92012	92064	92117	92169	92221	92273	92324	92376	5	10	16	21	26
84	92428	92480	92531	92583	92634	92686	92737	92788	92840	92891	5	10	15	20	26
85	92942	92993	93044	93095	93146	93197	93247	93298	93349	93399	5	10	15	20	26
86	93450	93500	93551	93601	93651	93702	93752	93802	93852	93902	5	10	15	20	25
87	93952	94002	94052	94101	94151	94201	94250	94300	94349	94399	5	10	15	20	25
88	94448	94498	94547	94596	94645	94694	94743	94792	94841	94890	5	10	15	20	25
89	94939	94988	95036	95085	95134	95182	95231	95279	95328	95376	5	10	15	19	24
90	95424	95472	95521	95569	95617	95665	95713	95761	95809	95856	5	10	14	19	24
91	95904	95952	95999	96047	96095	96142	96190	96237	96284	96332	5	9	14	19	24
92	96379	96426	96473	96520	96567	96614	96661	96708	96755	96802	5	9	14	19	24
93	96848	96895	96942	96988	97035	97081	97128	97174	97220	97267	5	9	14	18	23
94	97313	97359	97405	97451	97497	97543	97589	97635	97681	97727	5	9	14	18	23
95	97772	97818	97864	97909	97955	98000	98046	98091	98137	98182	5	9	14	18	23

ANTILOGARITHMS

81

	0	1	2	3	4	5	6	7	8	9	Mean Differences. 1	2	3	4	5	6	7
.00	10000	10023	10046	10069	10093	10116	10139	10162	10186	10209	2	5	7	9	12	14	16
.01	10233	10257	10280	10304	10328	10351	10375	10399	10423	10447	2	5	7	10	12	14	17
.02	10471	10495	10520	10544	10568	10593	10617	10641	10666	10691	2	5	7	10	12	15	17
.03	10715	10740	10765	10789	10814	10839	10864	10889	10914	10940	3	5	8	10	13	15	18
.04	10965	10990	11015	11041	11066	11092	11117	11143	11169	11194	3	5	8	10	13	15	18
.05	11220	11246	11272	11298	11324	11350	11376	11402	11429	11455	3	5	8	11	13	16	18
.06	11482	11508	11535	11561	11588	11614	11641	11668	11695	11722	3	5	8	11	13	16	19
.07	11749	11776	11803	11830	11858	11885	11912	11940	11967	11995	3	5	8	11	14	16	19
.08	12023	12050	12078	12106	12134	12162	12190	12218	12246	12274	3	6	8	11	14	17	20
.09	12303	12331	12359	12388	12417	12445	12474	12503	12531	12560	3	6	9	11	14	17	20
.10	12589	12618	12647	12677	12706	12735	12764	12794	12823	12853	3	6	9	12	15	18	21
.11	12882	12912	12942	12972	13002	13032	13062	13092	13122	13152	3	6	9	12	15	18	21
.12	13183	13213	13243	13274	13305	13335	13366	13397	13428	13459	3	6	9	12	15	18	21
.13	13490	13521	13552	13583	13614	13646	13677	13709	13740	13772	3	6	9	13	16	19	22
.14	13804	13836	13868	13900	13932	13964	13996	14028	14060	14093	3	6	10	13	16	19	22
.15	14125	14158	14191	14223	14256	14289	14322	14355	14388	14421	3	7	10	13	16	20	23
.16	14454	14488	14521	14555	14588	14622	14655	14689	14723	14757	3	7	10	13	17	20	24
.17	14791	14825	14859	14894	14928	14962	14997	15031	15066	15101	3	7	10	14	17	21	24
.18	15136	15171	15205	15241	15276	15311	15346	15382	15417	15453	4	7	11	14	18	21	25
.19	15488	15524	15560	15596	15631	15668	15704	15740	15776	15812	4	7	11	14	18	22	25
.20	15849	15885	15922	15959	15996	16032	16069	16106	16144	16181	4	7	11	15	18	22	26
.21	16218	16255	16293	16331	16368	16406	16444	16482	16520	16558	4	8	11	15	19	23	26
.22	16596	16634	16672	16711	16749	16788	16827	16866	16904	16943	4	8	12	15	19	23	27
.23	16982	17022	17061	17100	17140	17179	17219	17258	17298	17338	4	8	12	16	20	24	28
.24	17378	17418	17458	17498	17539	17579	17620	17660	17701	17742	4	8	12	16	20	24	28
.25	17783	17824	17865	17906	17947	17989	18030	18072	18113	18155	4	8	12	17	21	25	29
.26	18197	18239	18281	18323	18365	18408	18450	18493	18535	18578	4	8	13	17	21	25	30
.27	18621	18664	18707	18750	18793	18836	18880	18923	18967	19011	4	9	13	17	22	26	30
.28	19055	19099	19143	19187	19231	19275	19320	19364	19409	19454	4	9	13	18	22	26	31
.29	19498	19543	19588	19634	19679	19724	19770	19815	19861	19907	5	9	14	18	23	27	32
.30	19953	19999	20045	20091	20137	20184	20230	20277	20324	20370	5	9	14	19	23	28	32
.31	20417	20464	20512	20559	20606	20654	20701	20749	20797	20845	5	10	14	19	24	29	33
.32	20893	20941	20989	21038	21086	21135	21184	21232	21281	21330	5	10	15	19	24	29	34
.33	21380	21429	21478	21528	21577	21627	21677	21727	21777	21827	5	10	15	20	25	30	35
.34	21878	21928	21979	22029	22080	22131	22182	22233	22284	22336	5	10	15	20	25	31	36
.35	22387	22439	22491	22542	22594	22646	22699	22751	22803	22856	5	10	16	21	26	31	37
.36	22909	22961	23014	23067	23121	23174	23227	23281	23336	23388	5	11	16	21	27	32	37
.37	23442	23496	23550	23605	23659	23714	23768	23823	23878	23933	5	11	16	22	27	33	38
.38	23988	24044	24099	24155	24210	24266	24322	24378	24434	24491	6	11	17	22	28	34	39
.39	24547	24604	24660	24717	24774	24831	24889	24946	25003	25061	6	11	17	23	29	34	40
.40	25119	25177	25236	25293	25351	25410	25468	25527	25586	25645	6	12	18	23	29	35	41
.41	25704	25763	25823	25882	25942	26002	26062	26122	26182	26242	6	12	18	24	30	36	42
.42	26303	26363	26424	26485	26546	26607	26669	26730	26792	26853	6	12	18	24	31	37	43
.43	26915	26977	27040	27102	27164	27227	27290	27353	27416	27479	6	13	19	25	31	38	44
.44	27542	27606	27669	27733	27797	27861	27925	27990	28054	28119	6	13	19	26	32	39	45
.45	28184	28249	28314	28379	28445	28510	28576	28642	28708	28774	7	13	20	26	33	39	46
.46	28840	28997	28973	29040	29107	29174	29242	29309	29376	29444	7	13	20	27	34	40	47
.47	29512	29580	29648	29717	29785	29854	29923	29992	30061	30130	7	14	21	28	34	41	48

ANTILOGARITHMS

82

	0	1	2	3	4	5	6	7	8	9	Mean Di				
											1	2	3	4	
.50	31623	31696	31769	31842	31916	31989	32063	32137	32211	32285	7	15	22	29	3
.51	32359	32434	32509	32584	32659	32735	32809	32885	32961	33037	8	15	23	30	3
.52	33113	33189	33266	33343	33420	33497	33574	33651	33729	33806	8	15	23	31	3
.53	33884	33963	34041	34119	34198	34277	34356	34435	34514	34594	8	16	24	32	4
.54	34674	34754	34834	34914	34995	35075	35156	35237	35318	35400	8	16	24	32	4
.55	35481	35563	35645	35727	35810	35892	35975	36058	36141	36224	8	16	25	33	4
.56	36208	36392	36475	36559	36644	36728	36813	36898	36983	37068	8	17	25	34	4
.57	37154	37239	37325	37411	37497	37584	37670	37757	37844	37931	9	17	26	35	4
.58	38019	38107	38194	38282	38371	38459	38548	38637	38726	38815	9	18	27	35	4
.59	38905	38994	39084	39174	39264	39355	39446	39537	39628	39719	9	18	27	36	4
.60	39811	39902	39994	40087	40179	40272	40365	40458	40551	40644	9	19	28	37	4
.61	40738	40832	40926	41020	41115	41210	41305	41400	41495	41591	9	19	28	38	4
.62	41687	41783	41879	41976	42073	42170	42267	42364	42462	42560	10	19	29	39	4
.63	42658	42756	42855	42954	43053	43152	43251	43351	43451	43551	10	20	30	40	5
.64	43652	43752	43853	43954	44055	44157	44259	44361	44463	44566	10	20	30	41	5
.65	44668	44771	44875	44978	45082	45186	45290	45394	45499	45604	10	21	31	42	5
.66	45709	45814	45920	46026	46132	46238	46345	46452	46559	46666	11	21	32	43	5
.67	46774	46881	46989	47098	47206	47315	47424	47534	47643	47753	11	22	33	44	5
.68	47863	47973	48084	48195	48306	48417	48529	48641	48753	48865	11	22	33	45	5
.69	48978	49091	49204	49317	49431	49545	49659	49774	49888	50003	11	23	34	46	5
.70	50119	50234	50350	50466	50582	50699	50816	50933	51050	51168	12	23	35	47	5
.71	51286	51404	51523	51642	51761	51880	52000	52119	52240	52360	12	24	36	48	6
.72	52481	52602	52723	52845	52966	53088	53211	53333	53456	53580	12	24	37	49	6
.73	53703	53827	53951	54075	54200	54325	54450	54576	54702	54828	13	25	38	50	6
.74	54954	55081	55208	55336	55463	55590	55719	55847	55976	56105	13	26	38	51	6
.75	56234	56364	56494	56624	56754	56885	57016	57148	57280	57412	13	26	39	52	6
.76	57544	57677	57810	57943	58076	58210	58345	58479	58614	58749	13	27	40	54	6
.77	58884	59020	59156	59293	59429	59566	59704	59841	59979	60117	14	27	41	55	6
.78	60256	60395	60534	60674	60814	60954	61094	61235	61376	61518	14	28	42	56	7
.79	61659	61802	61944	62087	62230	62373	62517	62661	62806	62951	14	29	43	58	7
.80	63096	63241	63387	63533	63680	63826	63973	64121	64269	64417	15	29	44	59	7
.81	64565	64714	64863	65013	65163	65313	65464	65615	65766	65917	15	30	45	60	7
.82	66069	66222	66374	66527	66681	66834	66988	67143	67298	67453	15	31	46	62	7
.83	67608	67764	67920	68077	68234	68391	68549	68707	68865	69024	16	32	47	63	7
.84	69183	69343	69503	69663	69823	69984	70146	70307	70469	70632	16	32	48	64	81
.85	70795	70958	71121	71285	71450	71614	71779	71945	72111	72277	17	33	50	66	8
.86	72444	72611	72778	72946	73114	73282	73451	73621	73790	73961	17	34	51	68	8
.87	74131	74302	74473	74645	74817	74989	75162	75336	75509	75683	17	35	52	69	8
.88	75858	76033	76208	76384	76560	76736	76913	77090	77268	77446	18	35	53	71	8
.89	77625	77804	77983	78163	78343	78524	78705	78886	79068	79250	18	36	54	72	91
.90	79433	79616	79799	79983	80168	80353	80538	80724	80910	81096	19	37	56	74	9
.91	81283	81470	81658	81846	82035	82224	82414	82604	82794	82985	19	38	57	76	9
.92	83176	83368	83560	83753	83946	84140	84333	84528	84723	84918	19	39	58	78	9
.93	85114	85310	85507	85704	85901	86099	86298	86497	86696	86896	20	40	60	79	9
.94	87096	87297	87498	87700	87902	88105	88308	88512	88716	88920	20	41	61	81	102
.95	89125	89331	89536	89743	89950	90157	90365	90573	90782	90991	21	42	62	83	104
.96	91201	91411	91622	91833	92045	92257	92470	92683	92897	93111	21	42	64	85	106

RECIPROCALS OF NUMBERS. FROM 1 TO 10

[*Numbers in difference columns to be subtracted, not added.*] 83

	0	1	2	3	4	5	6	7	8	9	Mean Differences 1	2	3	4	5	6	7	8
1·0	1·000	9901	9804	9709	9615	9524	9434	9346	9259	9174								
1·1	·9091	9009	8929	8850	8772	8696	8621	8547	8475	8403								
1·2	·8333	8264	8197	8130	8065	8000	7937	7874	7813	7752								
1·3	·7692	7634	7576	7519	7463	7407	7353	7299	7246	7194								
1·4	·7143	7092	7042	6993	6944	6897	6849	6803	6757	6711	5	10	14	19	24	29	33	38
1·5	·6667	6623	6579	6536	6494	6452	6410	6369	6329	6289	4	8	13	17	21	25	29	33
1·6	·6250	6211	6173	6135	6098	6061	6024	5988	5952	5917	4	7	11	15	18	22	26	29
1·7	·5882	5848	5814	5780	5747	5714	5682	5650	5618	5587	3	6	10	13	16	20	23	26
1·8	·5556	5525	5495	5464	5435	5405	5376	5348	5319	5291	3	6	9	12	15	17	20	23
1·9	·5263	5236	5208	5181	5155	5128	5102	5076	5051	5025	3	5	8	11	13	16	18	21
2·0	·5000	4975	4950	4926	4902	4878	4854	4831	4808	4785	2	5	7	10	12	14	17	19
2·1	·4762	4739	4717	4695	4673	4651	4630	4608	4587	4566	2	4	7	9	11	13	15	17
2·2	·4545	4525	4505	4484	4464	4444	4425	4405	4386	4367	2	4	6	8	10	12	14	16
2·3	·4348	4329	4310	4292	4274	4255	4237	4219	4202	4184	2	4	5	7	9	11	13	14
2·4	·4167	4149	4132	4115	4098	4082	4065	4049	4032	4016	2	3	5	7	8	10	12	13
2·5	·4000	3984	3968	3953	3937	3922	3906	3891	3876	3861	2	3	5	6	8	9	11	12
2·6	·3846	3831	3817	3802	3788	3774	3759	3745	3731	3717	1	3	4	6	7	8	10	11
2·7	·3704	3690	3676	3663	3650	3636	3623	3610	3597	3584	1	3	4	5	7	8	9	11
2·8	·3571	3559	3546	3534	3521	3509	3497	3484	3472	3460	1	2	4	5	6	7	9	10
2·9	·3448	3436	3425	3413	3401	3390	3378	3367	3356	3344	1	2	3	5	6	7	8	9
3·0	·3333	3322	3311	3300	3289	3279	3268	3257	3247	3236	1	2	3	4	5	6	7	9
3·1	·3226	3215	3205	3195	3185	3175	3165	3155	3145	3135	1	2	3	4	5	6	7	8
3·2	·3125	3115	3106	3096	3086	3077	3067	3058	3049	3040	1	2	3	4	5	6	7	8
3·3	·3030	3021	3012	3003	2994	2985	2976	2967	2959	2950	1	2	3	4	4	5	6	7
3·4	·2941	2933	2924	2915	2907	2899	2890	2882	2874	2865	1	2	3	3	4	5	6	7
3·5	·2857	2849	2841	2833	2825	2817	2809	2801	2793	2786	1	2	2	3	4	5	6	6
3·6	·2778	2770	2762	2755	2747	2740	2732	2725	2717	2710	1	2	2	3	4	5	5	6
3·7	·2703	2695	2688	2681	2674	2667	2660	2653	2646	2639	1	1	2	3	4	4	5	6
3·8	·2632	2625	2618	2611	2604	2597	2591	2584	2577	2571	1	1	2	3	3	4	5	5
3·9	·2564	2558	2551	2545	2538	2532	2525	2519	2513	2506	1	1	2	3	3	4	4	5
4·0	·2500	2494	2488	2481	2475	2469	2463	2457	2451	2445	1	1	2	2	3	4	4	5
4·1	·2439	2433	2427	2421	2415	2410	2404	2398	2392	2387	1	1	2	2	3	3	4	5
4·2	·2381	2375	2370	2364	2358	2353	2347	2342	2336	2331	1	1	2	2	3	3	4	5
4·3	·2326	2320	2315	2309	2304	2299	2294	2288	2283	2278	1	1	2	2	3	3	4	4
4·4	·2273	2268	2262	2257	2252	2247	2242	2237	2232	2227	1	1	2	2	3	3	4	4
4·5	·2222	2217	2212	2208	2203	2198	2193	2188	2183	2179	0	1	1	2	2	3	3	4
4·6	·2174	2169	2165	2160	2155	2151	2146	2141	2137	2132	0	1	1	2	2	3	3	4
4·7	·2128	2123	2119	2114	2110	2105	2101	2096	2092	2088	0	1	1	2	2	3	3	4
4·8	·2083	2079	2075	2070	2066	2062	2058	2053	2049	2045	0	1	1	2	2	3	3	4
4·9	·2041	2037	2033	2028	2024	2020	2016	2012	2008	2004	0	1	1	2	2	2	3	3
5·0	·2000	1996	1992	1988	1984	1980	1976	197[illegible]	1969	1965	0	1	1	2	2	2	3	3
5·1	·1961	1957	1953	1949	1946	1942	1938	1934	1931	1927	0	1	1	2	2	2	3	3
5·2	·1923	1919	1916	1912	1908	1905	1901	1898	1894	1890	0	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

RECIPROCALS OF NUMBERS. FROM 1 TO 10

(Numbers in difference columns to be subtracted, not added.)

	0	1	2	3	4	5	6	7	8	9	Mean Differences 123
5·5	·1818	1815	1812	1808	1805	1802	1799	1795	1792	1789	011
5·6	·1786	1783	1779	1776	1773	1770	1767	1764	1761	1757	011
5·7	1754	1751	1748	1745	1742	1739	1736	1733	1730	1727	011
5·8	1724	1721	1718	1715	1712	1709	1706	1704	1701	1698	011
5·9	·1695	1692	1689	1686	1684	1681	1678	1675	1672	1669	011
6·0	·1667	1664	1661	1658	1656	1653	1650	1647	1645	1642	011
6·1	·1639	1637	1634	1631	1629	1626	1623	1621	1618	1616	011
6·2	·1613	1610	1608	1605	1603	1600	1597	1595	1592	1590	011
6·3	·1587	1585	1582	1580	1577	1575	1572	1570	1567	1565	001
6·4	·1562	1560	1558	1555	1553	1550	1548	1546	1543	1541	001
6·5	·1538	1536	1534	1531	1529	1527	1524	1522	1520	1517	001
6·6	·1515	1513	1511	1508	1506	1504	1502	1499	1497	1495	001
6·7	·1493	1490	1488	1486	1484	1481	1479	1477	1475	1473	001
6·8	·1471	1468	1466	1464	1462	1460	1458	1456	1453	1451	001
6·9	·1449	1447	1445	1443	1441	1439	1437	1435	1433	1431	001
7·0	·1429	1427	1425	1422	1420	1418	1416	1414	1412	1410	001
7·1	·1408	1406	1404	1403	1401	1399	1397	1395	1393	1391	001
7·2	·1389	1387	1385	1383	1381	1379	1377	1376	1374	1372	001
7·3	·1370	1368	1366	1364	1362	1361	1359	1357	1355	1353	001
7·4	·1351	1350	1348	1346	1344	1342	1340	1339	1337	1335	001
7·5	·1333	1332	1330	1328	1326	1325	1323	1321	1319	1318	001
7·6	·1316	1314	1312	1311	1309	1307	1305	1304	1302	1300	001
7·7	·1299	1297	1295	1294	1292	1290	1289	1287	1285	1284	000
7·8	·1282	1280	1279	1277	1276	1274	1272	1271	1269	1267	000
7·9	·1266	1264	1263	1261	1259	1258	1256	1255	1253	1252	000
8·0	·1250	1248	1247	1245	1244	1242	1241	1239	1238	1236	000
8·1	·1235	1233	1232	1230	1229	1227	1225	1224	1222	1221	000
8·2	·1220	1218	1217	1215	1214	1212	1211	1209	1208	1206	000
8·3	·1205	1203	1202	1200	1199	1198	1196	1195	1193	1192	000
8·4	·1190	1189	1188	1186	1185	1183	1182	1181	1179	1178	000
8·5	·1176	1175	1174	1172	1171	1170	1168	1167	1166	1164	000
8·6	·1163	1161	1160	1159	1157	1156	1155	1153	1152	1151	000
8·7	·1149	1148	1147	1145	1144	1143	1142	1140	1139	1138	000
8·8	·1136	1135	1134	1133	1131	1130	1129	1127	1126	1125	000
8·9	·1124	1122	1121	1120	1119	1117	1116	1115	1114	1112	000
9·0	1111	1110	1109	1107	1106	1105	1104	1103	1101	1100	000
9·1	·1099	1098	1096	1095	1094	1093	1092	1090	1089	1088	000
9·2	·1087	1086	1085	1083	1082	1081	1080	1079	1078	1076	000
9·3	1075	1074	1073	1072	1071	1070	1068	1067	1066	1065	000
9·4	·1064	1063	1062	1060	1059	1058	1057	1056	1055	1054	000
9·5	·1053	1052	1050	1049	1048	1047	1046	1045	1044	1043	000
9·6	·1042	1041	1039	1038	1037	1036	1035	1034	1033	1032	000
9·7	·1031	1030	1029	1028	1027	1026	1025	1024	1022	1021	000